Accessing Spoken Interaction Through Dialogue Processing

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
von der Fakultät für Informatik
der Universität Karlsruhe genehmigte Dissertation von

Doctoral thesis accepted by the Faculty of Computer Science at Karlsruhe University by

Klaus Ries
aus
Wiesbaden / Hessen / Deutschland
born in Wiesbaden, Hesse, Germany

Tag der mündlichen Prüfung: *Defended on*	14.12.2001
Gutachter: *Thesis Reader*	Prof. Dr. Alexander Waibel
Zweitgutachter: *Second Thesis Reader*	Prof. Dr. Eduard Hovy *University of Southern California at Los Angeles*

Zusammenfassung

Unser Leben, unsere Leistungen und unsere Umgebung, alles wird derzeit durch Schriftsprache dokumentiert. Die rasante Fortentwicklung der technischen Möglichkeiten Audio, Bilder und Video aufzunehmen, abzuspeichern und wiederzugeben kann genutzt werden um die schriftliche Dokumentation von menschlicher Kommunikation, zum Beispiel Meetings, zu unterstützen, zu ergänzen oder gar zu ersetzen. Diese neuen Technologien können uns in die Lage versetzen Information aufzunehmen, die anderweitig verloren gehen, die Kosten der Dokumentation zu senken und hochwertige Dokumente mit audiovisuellem Material anzureichern. Die Indizierung solcher Aufnahmen stellt die Kerntechnologie dar um dieses Potential auszuschöpfen. Diese Arbeit stellt effektive Alternativen zu schlüsselwortbasierten Indizes vor, die Suchraumeinschränkungen bewirken und teilweise mit einfachen Mitteln zu berechnen sind.

Die Indizierung von Sprachdokumenten kann auf verschiedenen Ebenen erfolgen: Ein Dokument gehört stilistisch einer bestimmten Datenbasis an, welche durch sehr einfache Merkmale bei hoher Genauigkeit automatisch bestimmt werden kann. Durch diese Art von Klassifikation kann eine Reduktion des Suchraumes um einen Faktor der Größenordnung 4-10 erfolgen. Die Anwendung von thematischen Merkmalen zur Textklassifikation bei einer Nachrichtendatenbank resultiert in einer Reduktion um einen Faktor 18.

Da Sprachdokumente sehr lang sein können müssen sie in thematische Segmente unterteilt werden. Ein neuer probabilistischer Ansatz sowie neue Merkmale (Sprecherinitiative und Stil) liefern vergleichbare oder bessere Resultate als traditionelle schlüsselwortbasierte Ansätze. Diese thematische Segmente können durch die vorherrschende Aktivität charakterisiert werden (erzählen, diskutieren, planen, . . .), die durch ein neuronales Netz detektiert werden kann. Die Detektionsraten sind allerdings begrenzt da auch Menschen diese Aktivitäten nur ungenau bestimmen. Eine maximale Reduktion des Suchraumes um den Faktor 6 ist bei den verwendeten Daten theoretisch möglich. Eine thematische Klassifikation dieser Segmente wurde ebenfalls auf einer Datenbasis durchgeführt, die Detektionsraten für diesen Index sind jedoch gering.

Auf der Ebene der einzelnen Äußerungen können Dialogakte wie Aussagen, Fragen, Rückmeldungen (aha, ach ja, echt?, . . .) usw. mit einem diskriminativ trainierten Hidden Markov Model erkannt werden. Dieses Verfahren kann um die Erkennung von kurzen Folgen wie Frage/Antwort-Spielen erweitert werden (Dialogspiele). Dialogakte und -spiele können eingesetzt werden um Klassifikatoren für globale Sprechstile zu bauen. Ebenso könnte ein Benutzer sich an eine bestimmte Dialogaktsequenz erinnern und versuchen, diese in einer grafischen Repräsentation wiederzufinden.

In einer Studie mit sehr pessimistischen Annahmen konnten Benutzer eines aus vier ähnlichen und gleichwahrscheinlichen Gesprächen mit einer Genauigkeit von $\approx 43\%$ durch eine graphische Repräsentation von Aktivität bestimmt. Dialogakte könnte in diesem Szenario ebenso nützlich sein, die Benutzerstudie konnte aufgrund der geringen Datenmenge darüber keinen endgültigen Aufschluß geben. Die Studie konnte allerdings für detailierte Basismerkmale wie Formalität und Sprecheridentität keinen Effekt zeigen.

Abstract

Written language is one of our primary means for documenting our lives, achievements, and environment. Our capabilities to record, store and retrieve audio, still pictures, and video are undergoing a revolution and may support, supplement or even replace written documentation. This technology enables us to record information that would otherwise be lost, lower the cost of documentation and enhance high-quality documents with original audiovisual material. The indexing of the audio material is the key technology to realize those benefits. This work presents effective alternatives to keyword based indices which restrict the search space and may in part be calculated with very limited resources.

Indexing speech documents can be done at a various levels: Stylistically a document belongs to a certain database which can be determined automatically with high accuracy using very simple features. The resulting factor in search space reduction is in the order of 4-10 while topic classification yielded a factor of 18 in a news domain.

Since documents can be very long they need to be segmented into topical regions. A new probabilistic segmentation framework as well as new features (speaker initiative and style) prove to be very effective compared to traditional keyword based methods. At the topical segment level activities (storytelling, discussing, planning, ...) can be detected using a machine learning approach with limited accuracy; however even human annotators do not annotate them very reliably. A maximum search space reduction factor of 6 is theoretically possible on the databases used. A topical classification of these regions has been attempted on one database, the detection accuracy for that index, however, was very low.

At the utterance level dialogue acts such as statements, questions, backchannels (aha, yeah, ...), etc. are being recognized using a novel discriminatively trained HMM procedure. The procedure can be extended to recognize short sequences such as question/answer pairs, so called dialogue games. Dialog acts and games are useful for building classifiers for speaking style. Similarily a user may remember a certain dialog act sequence and may search for it in a graphical representation.

In a study with very pessimistic assumptions users are able to pick one out of four similar and equiprobable meetings correctly with an accuracy $\approx 43\%$ using graphical activity information. Dialogue acts may be useful in this situation as well but the sample size did not allow to draw final conclusions. However the user study fails to show any effect for detailed basic features such as formality or speaker identity.

Acknowledgments

Whenever acknowledgments are given, there is a danger of leaving someone out or forgetting something they have done. Most certainly this work would have not been possible without the discussions, inspiration and support by many people.

Prof. Alex Waibel, my thesis advisor, has provided support and a highly dynamic scientific environment over the years which I cannot acknowledge enough. He offered the opportunity to conduct most of this research at Carnegie Mellon University (CMU) in Pittsburgh. My second thesis reader, Prof. Ed Hovy from ISI/UCLA provided extremely helpful comments which most certainly improved this work a lot – it was a joy to interact with him and I was delighted to have him over in Karlsruhe for my defense.

Carnegie Mellon University and the members of my group would all deserve mentions. Most important was the DARPA funded project Clarity for the completition of this thesis. Klaus Zechner and Thomas Polzin were collaborators in this project, Alon Lavie and Lori Levin being the lead scientists. Donna Gates provided help with Spanish linguistics and Michael Bett provided support for the system integration. The John Hopkins Workshop in 1997 provided the initial drive to go into the project, a big thanks to all of you, of course foremost to Dan, Andreas and Liz – it was a joy to work in that group.

Pittsburgh grew on me and I had a wonderful time there. There were so many people of all nationalities which made life like a dream. Just to name some of them, Clark and Joanie, Matthias and Ju, Ralph and Sasha, Willy and Clara, Andy, Heather, Laurie, Dani, Vasiliki, Marcus, Dominic, ... Of course there are too many missing in this list but still I want to mention our house on Forbes Terrace. It was a home and focal point over the years and was something special, despite all of its flaws.

Going back to the other side of the Atlantic my parents Maria and Ludwig and my brother Thorsten were there all the time and supported me and my spirits. I am glad to have you, you brought me up to that point and made it possible for me to go all the way. And finally, back in Germany for defending my thesis, I met Christiane and she made sense of it all – from here we go for life after the defense.

Contents

Chapter 1

Introduction

1.1 Indexing Speech Data

“Wer schreibt der bleibt” (those who write are being remembered) is a German proverb and there is certainly truth to it: Written material can be stored easily and read later on, it can be indexed and retrieved electronically, it can be made available worldwide over the Internet and it can be skimmed visually. Spoken and visual records are now being captured thanks to the ubiquous availability of digital sensors such as cameras and microphones on stationary and mobile computing platforms, cellphones and digital cameras. Adequate network and storage capacity for speech data are available or are becoming available to store any or all of our human encounters. The open question is how to make use of that data, especially how to find information effectively.

This thesis shows that audio records of meetings and other types of spoken communications may be indexed by non-keyword based indices which may be more practical, complementary or better than keyword based methods alone (Tab. 1.1). Non-keyword indices such as the type and style of interaction are often easily remembered, may be easy to annotate manually and automatically and are recognized properties of language. Spoken interactions are different from most written text and from TV and audio broadcasts as they are usually a joint achievement within a defined group. Consequently the language is less formal (Biber, 1988), contains a low number of unique keywords (Tab. 1.2) which are likely ideosynchratic and the speech recognition problem is very difficult (Waibel et al. (2001a), Sec. 2.4.5.2). Additionally the language is less constrained by formal genre constraints which may lead to a greater variation in the style of the interaction [1]. It is therefore plausible that – in contrast to the experience in written lan-

[1] Examples of genre constraints in TV news-shows are the length, number and types of segments of a show. A segment type may be charecterized by introduction or closing phrases (*Chris-*

Feature	Memorization	Example manifestations	Extraction / Detection
Time	Often only contextual.	Clock time, relation to other events.	Wallclock time, time express-sions, visual cues.
Location	Often very well.	Physical coordinates, function of room.	Global Positioning System, image analysis, fixed location of recording system.
People	Often very well.	Names, functions of the participants.	Face and speaker identification or enrollment.
Style	Often very well.	Kind of interaction, type of dialog.	Via dialog acts and low level features.
Topic	Semantic remembered, not keywords.	Keywords, topic changes.	LVCSR, manual transcription.
Emotion	Own emotion, emoting events may be important.	Often only internal, emotions are not always displayed.	Prosody, facial and other gestures, skin conductivity.
Details	"flashbulb-effect": Very few situations but highly accurate.	Combing in hair, gaze, pen falls from table, etc.	Unclear, many sensors needed.

Table 1.1: **Indices for Spoken Interactions:** Features are detectable properties of spoken communication, indices are features that are known and may be used by the information seeker for information access. Retrieving spoken interactions often makes sense for participants of the conversation who may memorize specific elements of the interaction (Sec. 2.4.4). An outsider may never even understand the discourse nor find good indices (Fig. 2.2).

guage – spoken interactions can be indexed effectively using non-keyword based methods. When creating an indexing system for speech data the location, time and usually also the participants of the conversation can be determined easily and automatically at recording time. Style – besides emotion – is therefore the most important realistic index that is missing for a speech retrieval system (Tab. 1.1). The style of a conversation may not only be remembered well but may also be used by non-participants as a priori indicator for the relevance of a conversation. Emotion, on the other hand, is neither always displayed by discourse participants nor is displayed emotion easy to detect reliably with an unintrusive apparatus (Sec. 3.8, Polzin (1999)).

Details such as (head) movements, smiles, jokes etc., which we might remember in rare occasions, are not considered in this thesis. These features are often remembered as a bundle (flashbulb memory) and may help to find landmarks from

tiane Amanpour for CNN from Kabul, Afghanistan) and by specific interactions (interviews, "live on the scene" reports, anchor-speech, expert discussion and testimonial).

which to navigate through speech data. They may therefore be useful for a participant of a conversation but they are likely unavailable for an outsider. The usage of these “details” is however questionable if other indices are already strong enough to find these landmarks. Other open questions related to “details” are how much can realistically be gained given our memorization (Sec. 2.4.4), whether realistic retrieval interfaces can be build and whether these features can be made available at reasonable cost. Given these uncertainties descriptions of situations which are rich in detail were not considered in this thesis.

The focus was therefore on the use of stylistic information for information access to speech data and is demonstrated at a number of levels (Fig. 1.1). At each of these levels stylistical classifications are described and detected automatically. Identifying speech documents at a very high level such TV-broadcasts vs. personal communications turns out to be easy for machines and provides a reasonable reduction is search space. A speech document can be broken down into topical regions using features such as speaker initiative and speaking style and the resulting algorithms are competitive or better than keyword based methods in terms of accuracy and may require no speech recognition. It is shown that these topical segments can be annotated manually and automatically with activities such such as storytelling and discussion. Finally, at the lowest level, statements, questions, backchannels (aha, yeah) (dialogue acts) as well as short sequences of dialogue acts such as question/answer pairs (dialogue games), are detected automatically. These classifications can be tested for retrieval effectiveness by measuring their average information content with information theoretic methods. Additionally a user study is conducted which makes use of activities, dialogue acts, speaker identity and formality to distinguish meetings. The results indicate that non-keyword based methods can be used for indexing spoken communications and some empirical evidence has been developed to show that they are orthogonal or complementary to keyword based techniques. In conjunction the use of non-keyword based features for accessing and documenting spoken communications is a clear advance towards practical systems.

The introduction proceeds to map the landscape of the thesis. The intelligent meeting room project, which serves as a framework for this thesis, the software contributed to that system and the rationale for non-keyword based retrieval in that context is detailed in Sec. 1.2. The importance of indexing for the documentation of spoken communication is addressed in (Sec. 1.3). The underlying linguistic concepts and most important databases are introduced in Sec. 1.4. The organization of the main thesis chapters is presented in Sec. 1.5.

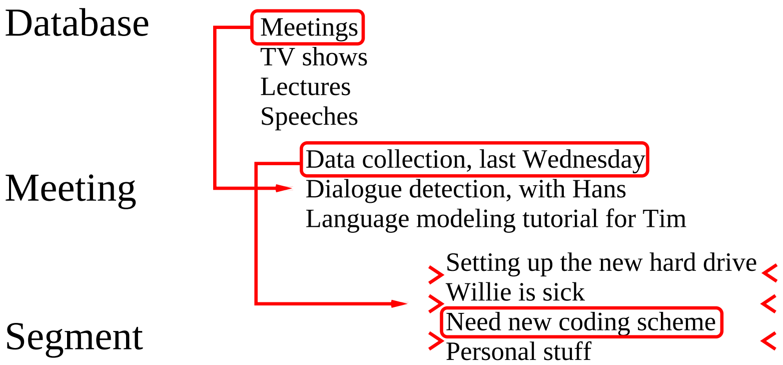

Database
Meeting
Segment
Meetings
TV shows
Lectures
Speeches
Data collection, last Wednesday
Dialogue detection, with Hans
Language modeling tutorial for Tim
Setting up the new hard drive
Willie is sick
Need new coding scheme
Personal stuff

Figure 1.1: **Information Access Hierarchy:** Oral communications take place in very different situations and in very different styles which are often known about a spoken communication or might be inferred easily – the database level. Within a certain (sub-)database a specific meeting needs to be selected. Since meetings might be long they need to be segmented into more consumable entries such as topics. Topics may be indexed using activities such as storytelling and discussion. At the utterance level statements, questions, answers etc. (dialogue acts) might be used to indicate the type of interaction (not shown in figure).

1.2 The Intelligent Meetingroom

Intelligent rooms have been a powerful metaphor for the development of pervasive computing where the computer is part of the background of our environment although it is accessible everywhere, anywhere and seamlessly. The most important aspect of intelligent rooms is so far the natural control of room functions, starting with lighting and heating conditions, presentation and entertainment functionality, room services through robots, information services and so forth. The intelligent meeting room metaphor however expands that vision by making the room self-documenting additionally to easy to control and set up (Fig. 1.2).

My thesis advisor Prof. A. Waibel has therefore started a project which would, additionally to the multimodal recognition capabilities, bring together technologies to enable the documentation of spoken interactions in that room [2]. Obviously

[2] Another research strategy that could have been used is to build a small prototype system for a limited task that could be fielded, tested and expanded. An example of such a system is the SCAN/SCANmail system at AT&T (Whittaker et al., 1999) or the concept of an audio notebook (Stifelman et al., 2001; Whittaker et al., 1994). That research strategy has the advantage that it might shed light on a concrete and full solution of a specific problem. The research strategy in our group has the advantage that a more global perspective can be developed and novel enabling technologies such as dialogue analysis and summarization can be developed. The systems in our

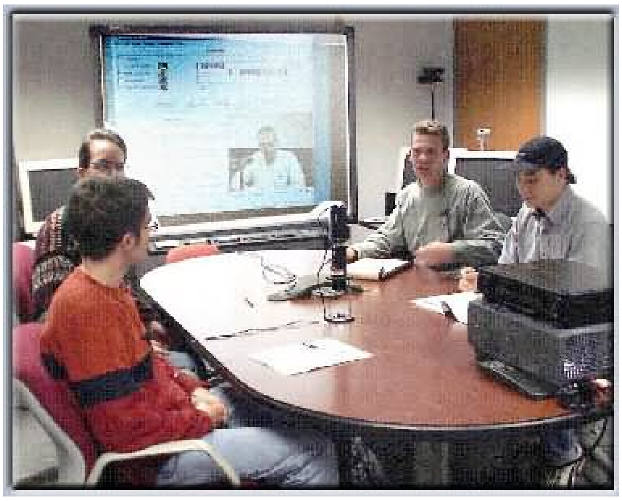


Figure 1.2: **Intelligent Meeting Room:** The idea of the intelligent meeting room scenario is that does not react on user input but that it supports the documentation of meetings that happen in it. For that purpose audio and visual recording are conducted, speech recognition and speaker detection are being performed, people are tracked and identified visually, speech is summarized and the dialogue itself is analyzed. The meeting browser tool serves as an integration as well as a presentation platform for the documentation.

this is a lot more ambitious than the control or adaptation of a room. The system needs to process input which is not only generated by humans using multiple modalities but which is also not directed at the system. The problem was therefore compartmentalized such that different technologies could be developed that would solve individual aspects of the problem:

Meeting browser The results of a documentation need to be accessible in some form and the meeting browser is designed to serve (a) as an integration platform for the multiple modalities and (b) as a graphical access system to the generated record. The meeting browser is developed by Michael Bett (Fig. 2.1). It features the following core graphical components: Direct textual display, summary display, colored bars representing dialogue features, speaker encoding via colors, audio playback synchronized with word level highlighting and video playback (if video is available). The meeting browser also features some limited search capabilities and selection mechanisms for spoken interactions. It does not feature techniques for fast audio

lab are now reaching the threshold where they could be deployed such that those two strategies can be merged soon.

skimming such as audio compression, pause elimination and topic jumping. The system is implemented in Java and runs under Windows.

Speech summarization Speech summarization may be different from text summarization and in work by Klaus Zechner a speech summarization system with a standard text summarizer at its core has been built (Sec. 2.4.6.4,2.2). His system however features a number of additional steps to make speech summaries better by addressing the problems of inaccurate machine transcripts, lack of clause boundaries, distributed information (question/answer pairs), disfluent and unreadable speech and lack of topic boundaries (Zechner, 2001; Zechner and Waibel, 2000a). The system can be called as a separate program from the meeting browser and runs under Windows and UNIX.

Dialogue component To add non-keyword based retrieval methods dialogue analysis as featured in this work is being performed. Specifically the meeting browser can currently request topical segmentations, activity annotations (storytelling, discussion etc.) and dialogue act annotation from the dialogue component. The dialogue component is a large toolkit which enables the construction of statistical classification and annotation models for dialogue acts and activities and for topical segmentations. It is highly flexible and configurable for rapid experimentation and can be extended to similar tasks. It also contains a number of prebuild models, including a dialogue act and activity classifier [3]. Some other non-topical features such as formality might be integrated directly into the meeting browser. The integration is currently not in day to day use since the dialogue component requires to start a separate server process on a UNIX or Windows system.

Auditory scene analysis The analysis of environmental sounds may carry other pieces of information such as “is someone in the room?”, “is a telephone ringing?”, “is the person working on the computer?” etc. (Malkin, 2002) [4]

[3] The flexibility is achieved by using the efficient object oriented language Python (`http://www.python.org/`) on top of our labs speech recognition system Janus (Finke et al., 1996). Janus provides an efficient neural network package, prosodic analysis and language model evaluation. The full system contains more than 45000 lines of Python code which is complemented by an estimated 2.000-10.000 lines of C code added or modified inside of Janus itself. Several other external components have been integrated, especially a part-of-speech tagger based on (Brill, 1994a) and trained by Klaus Zechner, the language model toolkit by the author from earlier work, a robust parser (Gavaldà, 2000) and off the shelve machine learning algorithms (Joachims, 2000; Murthy et al., 1993; Quinlan, 1992). The special Janus-Python and Python class-libraries are available and tested on LINUX, SUN-OS and Windows. The meeting browser can be interfaced via a socket connection.

[4] The project is conducted by Rob Malkin and preliminary system integration is underway.

	Ratios in %				
		Types per Token			
Corpus	Non-Keywords	All	First 100	First 200	$\frac{\mathrm{fof}_2}{\mathrm{fof}_1+\mathrm{fof}_2}$
Meeting corpus	77.97	18.43	71.36	67.75	26.21
Santa Barbara	75.51	29.90	73.93	69.88	24.59
CallHome English	75.48	15.80	78.61	71.66	25.55
CallHome Spanish	75.21	24.82	78.26	74.19	22.96
Switchboard	75.12	5.22	75.95	69.26	27.51
Broadcast News	56.05	7.77	83.81	80.22	30.25
Brown corpus	51.45	17.27	82.27	79.70	32.88

Table 1.2: **Keywords in Conversations:** The first column shows the ratio of stopwords (or non-keywords) – the more stopwords the fewer keywords are available for indexing. The types per token ratio measures the percentage of unique keywords (types) per keyword (token). The values in the “all” column are measured over the whole document length, for the small SantaBarbara and meeting corpora the manually annotated topics are chosen as the base unit. In order to account for document length difference the “first 100” resp. “first 200” columns cut off all documents after the first 100 resp. 200 keywords (tokens). The ratio of keyword types which occur only once (fof_1) or twice (fof_2) would be high for a rich vocabulary. $\frac{\mathrm{fof}_2}{\mathrm{fof}_1+\mathrm{fof}_2}$ is a refinement of this argunment used in the context of language modeling (Ney et al., 1994).

Furthermore there are other multimodal capabilities that have been integrated into the intelligent meetingroom: Speaker identity by audio and video (Bett et al., 2000), speaker tracking in rooms (Bett et al., 2000), gaze tracking (Stiefelhagen et al., 1999, 2000), handwriting (Jaeger, 2000), emotion detection (Polzin, 1999) and speech recognition (Waibel et al., 2001a).

The purpose of the dialogue component in the meeting browser is therefore to deliver non-thematic information that can be used for information access, both for browsing one meeting as well as for picking a meeting. Traditional information retrieval so far focussed on the use of keywords that are supposed to capture the semantic content of the conversation, often called topic. The meeting corpus and the Santa Barbara Corpus represent that situation and Tab. 1.2 compares the keyword distribution of these corpora with other corpora such as CallHome where family members talk on the phone and Switchboard where strangers talk on the phone (see Sec. 1.4.3 for more details on the corpora). Broadcast News – which consists of spoken language as well – is very different and more similar to the Brown corpus which represents a balanced selection of written genre [5]. The overall percentage of keywords is much lower for meeting-like corpora and additionally the average number of keyword types per document as well. These observations are a reflection of several underlying reasons why keyword based access is unlikely to be as effective for dialogue retrieval as for written language. The main meta-reason is that conversations are conducted to be effective for the participants, not for overhearers or later retrieval, and that the participants can resolve any communication problems on the spot. The degree of this "overhearer effect" is depending on the familiarity of the discourse participants and the common ground established (Clark (1996), Fig. 2.2). In that sense conversations differ fundamentally from written language or broadcasts which is also observed in the informal language of conversations (Sec. 2.4.2.3, Heylighen and Dewaele (1999)). But this fundamental difference may also have effects on other aspects of language that are important for standard retrieval methods, especially the use of keywords (Tab. 1.2):

idiosyncratic keyword usage The participants share "insider" knowledge, use idiosyncratic terms and expressions freely, refer to and continue previous discussions which may prevent the occurrence of keywords a searcher would use.

keyword mismatch Keyword based methods rely on the idea that an item in the query can be matched with an item in the document. However the amount of keywords in meeting like documents is a lot lower than in written documents and the percentage of keyword types per keyword is also lower. (Tab. 1.2). One property of written text and likely broadcast media is that a word is not necessarily repeated but a synonym is used since stylistic constraints prohibit the repetition of keywords (Beeferman et al., 1997). Finding the right keywords to retrieve a document might therefore prove difficult.

Traditional techniques to mitigate the mismatch between keywords in queries and documents rely on the availability of semantic hierarchies or related data with similar semantic content. Topics in spoken communications can be very specific to the participants such that typically neither semantic hierarchies or related data are available.

LVCSR performance The performance of automated speech transcription with current LVCSR (large vocabulary continuous speech recognition) systems is around 38% word error on meeting data (Waibel et al. (2001a), Sec. 2.4.5.2) although it is much better for other speech genre. Additionally one might

[5] A sample of the Brown corpus was generated using the topics of Choi (2000)'s "9–11" database which correspond to 9–11 sentence initial segments of Brown corpus documents.

assume that many of the most important keywords may not be in the vocabulary of a general speech recognizer since the topics discussed are often very specific to the group which is communicating. The error regime of the LVCSR system is also likely not good enough to convince users to prefer reading transcripts over listening to audio (Stark et al., 2000).

long term memory Participants of conversations – with certain somewhat uncommon exceptions – remember only the general topic of a conversation and not the keywords used (Sec. 2.4.4). It is therefore very likely that they have forgotten the crucial keyword which might add to the difficulty already posed by the keyword mismatch problem.

1.3 Documentation of Spoken Interactions

Documenting spoken interactions is more complex than just recording it and the transition to a written document is usually more than a mere transcription task. Without any claim for completeness the documentation process modifies the original material by

- eliminating unnecessary, private or embarrassing material.
- extracting, adding and condensing information.
- citing the spoken interaction.
- relating the spoken interaction to other documents.
- retargeting the audience beyond the participants.

If an indexed audiovisual record is available documentation can be done cheaper, faster and/or with higher quality. The options are to either do just a recording and apply minimal (automatic) indexing, to use the record as a memory aid during minute construction or to fabricate a high quality multimedia document which consists of traditional minutes backed up by detailed citations of the original record (Fig. 1.3). Since memorization problems are not as severe if an audiovisual record is available the construction of a high-quality document may also be delayed or be performed on demand. Two indirect effects of audio recordings have been observed if the spoken interactions are documented in a standard manner:

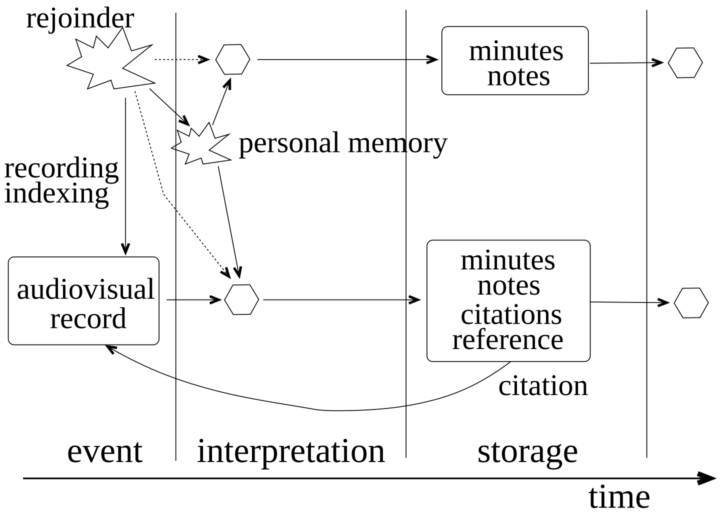


Figure 1.3: **Audiovisual Records:** Audio and visual records can be produced at the time the spoken communication takes place and an automatic indexing process may be applied to add information for finding a spoken interaction or passages of it. This record may be accessed for documentation. If the communication is interpreted by a person she or he might access personal memory and/or the audiovisual record. If a record is available it may be cited or referred to. The interpretation can be stored again and made available for retrieval; further interpretations may be based on these records.

lowered agreement costs Meeting minutes are typically circulated to reach agreement about the outcome. If an audiovisual record is being used during the construction of the minutes and is possibly supplementing the documentation it gains a higher level of credibility which may reduce the agreement efforts. In some cases the participants speak directly "for the record" (Moran et al., 1997) which ensures that the statements are considered to be part of the minutes.

higher attention Details of the meeting which are important don't need to be pinned down but can be accessed later on. Moran et al. (1997) reports that the user marked the audio record during the meeting to indicate sections that need to be listen to later on. Wilcox et al. (1997) reported that people liked the audio recording since it freed their meeting time from taking notes.

There are two interesting tasks: Adding indices automatically in the absence of any indexing (zero-effort documentation) and the support of navigation for manual documentation. Both tasks can be supported by adding indices to the audiovisual record and the indices developed in this work therefore contribute to the solution of that problem by providing:

Scenarios Information access applications are described including the aspects of privacy, manual labels that are available, user interfaces and the reinterpretation of spoken interactions.

Dialogue analysis algorithms To extend the set of available features the dialogue is automatically annotated with high level features such as activities, dialogue acts and games. Those might be more understandable and aggregate than the lower level features and therefore more suitable for a user. Additionally high level database and sub-database information are being inferred and topical segmentation is being performed.

Access performance assessment Information theoretic measures are used to determine the search space reduction of the different features, including microlevel features such as words (Sec. 1.4.2). Furthermore a user study is being performed to test the access performance (Sec. 6).

1.4 Linguistic background

1.4.1 Introduction

Most approaches to information access in written domains don’t use sophisticated techniques to transform the written input into features for the retrieval engine: Only keywords (and phrases) are used which are (sometimes) preprocessed with a stemming algorithm to remove morphological variation. Attempts to use stylistic information in information retrieval of written text showed fairly limited success (Karlgren, 2000; Kessler et al., 1997) and no conclusive results beyond the detection of broad genre were presented. The common observation that spoken language is less formal provides evidence that spoken language is more contextualized since formality is used to decontextualize (Sec. 2.4.2.3, Heylighen and Dewaele (1999)). If contextualization has a direct effect on language one should be able to measure it and it should be more important in spoken language than in written language.

This question is closely related to the question whether spoken language and written language are based on an underlying “langue” or whether there is such a great disparity that they have to be treated separately: The most radical approach is dialogism (Sec. 2.4.3.1) which assumes a great disparity and invites a holistic and highly contextualized interpretation of dialogues. Bahktin (1986) describes dialogue in terms of situation, topic and style. These categories seem to be very intuitive although there is also some dependency between the categories: The situation would definitely restrict the topic of the interaction and the style in which it is carried out. Bahktin (1986) assumes that speech genres – stable yet very flexible forms of language usage – have a profound impact on how language is being used and how it has to be taken. Less radical is the influential (systemic-)functional linguist Halliday (Sec. 2.4.3.2) who reintroduces the function of language as a centerpiece of grammar. In his terminology the functions of *register* (context of the situation) are the *field* (the nature of the event), the *tenor* (the social / relational aspect) and *mode* (the kind of language being used). This categorization is deeply correlated with other grammatically systems such as clause-level grammar. Since these dimensions, however, are highly dependent on each other one may conclude that they are not drawn very well. Additionally they don't reflect the distinctions interesting for an information retrieval application since topic is not a separate dimension and the categories might be hard to characterize for a lay person. These functions however are relatively fixed compared to Bakhtin's genres. Theoreticians like de'Saussure or even Chomsky – which still represent the mainstream way of thinking in modern linguistics even if not all of their analysis have survived – tend to ignore most of the contextual effects on language.

If spoken language is therefore more contextualized we would assume that the kind of interactions surrounding it and therefore their style would have a profound impact on the language which is what we try to measure. Halliday (1994) wouldn't deny this impact and he is trying to relate those "functions" directly to grammatical categories. Bahktin (1986) however goes head on with Saussure since his genres idea cannot be integrated as readily as Halliday's functions:

> Therefore, the single utterance with all its individuality and creativity, can in no way be regarded as a *completely free combination* of forms of language, as is supposed, for example, by Saussure (and by many other linguists after him), who juxtaposed the utterance (*la parole*), as a purely individual act, to the system of language that is purely social and mandatory for the individuum.

It is not crucial for this work whether Halliday or Bakhtin are right in principle – however the analysis does not a priori assume fixed functions and the categories are chosen such that they are intuitive and hopefully effective. This does not exclude a functional view of language, quite contrary, since we exactly try to establish such relationships which allow to identify the context of the conversation using features of the conversation. However the nature of the function and the features used is very different from Halliday (1994). The difference may also be attributed to a difference in the level of analysis: While for example activity-names may have some more universal value the (sub-)database level seems to be more flexible and activities may manifest themselves in very different ways given the situation and the participants (Sec. 2.4.2.1). Given the flexibility of these categories a machine learning approach as offered in this thesis does not only have practical advantages but may also be the appropriate modeling device.

The underlying assumption of this thesis is therefore that strong contextualization effects exist than can be measured and used to characterize a dialogue. The preceding Sec. 1.2 gives additional reasons why keyword based retrieval may not be as effective in spoken interactions as it is in written language. In Sec. 2.4 a large number of theories and computational models is introduced which are aimed at discourse and its categorization.

1.4.2 Stylistic features

One of the big questions is still what is the style or the type of an interaction? The problem of a general discourse typology is currently unsolved although a lot of proposals have been made (Sec. 2.4.2.3). It seems however that sub-categorization can be done fairly well and domain specific categorization can be done. This thesis applies two types of categorizations: At a very high level (the database level) conversations differ a lot and in many feature dimensions at the same time. Those distinctions are either naturally available since the spoken interactions have been filed in different databases or they can be determined using simple classifiers. Within one database the situation is more complex and the decision was made to look at databases such as meetings and define a specific set of activities such as “discussion”, “planning” and so forth. The activities were assumed to be constant in one topical unit, the annotation therefore consists of topical segmentation with an activity annotation.

As important as high level descriptions of spoken interactions are they need to be characterized by basic features. The following basic features are used, either directly, such as formality in the user study, or indirectly, as features for higher level characterizations (for more details see Sec. 2.4.2.3 and Sec. 4.2):

words The most natural feature are just the words by themselves, their parts of speech if they are rare or their semantic category (WordNet features).

simple syntactic features Some simple local syntactic features might be available which are related to register variation (Biber, 1988; Biber et al., 1999; Quirk et al., 1985).

durations Lengths of words, turns and overlaps between speakers are indicative of register variation and are easy to extract from the signal.

formality An abstract formality criterion based on part-of-speech distributions is calculated and exploited in the user study.

prosodic features Power and pitch and histograms or normalizations over those (e.g. the range of the pitch-variation, . . .) are used as features in emotion detection and topic segmentation.

However these features might not yet be rich enough and therefore other features which possibly encode high level non-keyword based indices are detected:

dialogue acts Statements, backchannels (*yeah, right !*), questions, attention directives (*hey, listen!!*), . . .

dialogue games Multiple dialogue acts, e.g. a series of statements with clarifications is an "information-game", questions with answers are "question-games".

activities Story-telling, discussion, informing, . . .

(sub-)database Different databases that are generally available: Phone calls, broadcast news etc.

emotions Neutral, happy, sad, excited, . . . [6]

1.4.3 Corpora Used in Empirical Studies

1.4.3.1 Introduction

One of the most important parts of this work is the selection of corpora – it defines the scope of the study and signifies the possible applications. Unfortunately there is no such thing as a large corpus of meetings that is available for scientific research at the time of this writing. Nor, even if that corpus would be available, it might not be annotated with many annotations. The corpora used are therefore picked to simulate aspects of the retrieval problem and also reflect the availability of annotations on corpora.

The CallHome Spanish annotations are the most throughout and have been done under project Clarity under participation of the author. The meeting corpus has been produced and annotated under project Genoa, the Santa Barbara corpus has been annotated under project Genoa as well and the dialogue act annotation for Switchboard has been done in preparation of the John Hopkins Summer Workshop at the University of Colorado in Boulder under the direction of Dan Jurafsky (Jurafsky and Shriberg, 1997). The author is highly indebted to the many

[6] (Polzin, 1999) annotated emotions on an utterance level. The detectors work in this thesis operate at the same level. For information access however cumulative features over topical segments (an all neutral, somewhat excited, somewhat happy segment) would be more appropriate.

individuals that constructed tagging tools, manuals, finally carried out the annotation itself and transformed the original annotation into products that are usable for other research groups. All of these steps require hard work, creativity and knowledge that can't be valued enough.

1.4.3.2 CallHome Spanish (CHS)

CallHome Spanish (CHS) is a corpus of personal telephone calls between family members that available via the linguistic data consortium (LDC96S35). 120 calls have been recorded and 5-10 mins of each calls are transcribed by the LDC. There are about 100-250 turns per transcribed dialogue. The calls are often very personal and the participants rarely seem to remember that they're being recorded, sometimes with the exception of an initial discussion. One person is calling from the US to a relative in their Spanish speaking home country, the language is Spanish. In project Clarity the database has been annotated in our working group using dialogue acts, games and activities. The coding manual is under preparation for publication (Thymé-Gobbel et al., 2001) and the database is available to the scientific community via the LDC (Waibel et al., 2001b). This corpus is therefore the largest real dialogue corpus in this study that is annotated with dialogue information such as activities (Switchboard contains only dialogue act information but is an order of magnitude larger than CallHome Spanish).

1.4.3.3 Meeting database

Our research group has been recording and transcribing some of our own meetings. Eight of these meetings have been annotated with emotions, segmented and annotated with activities. Most of the meetings are highly informal without a predetermined agenda and the participants may have close personal contacts. A lot of the meetings are from our own data-recording group and indeed some discuss the recording of meetings. This data is not available for outside use since the participants did not agree to that and the data contains both group internal as well as private information – even for inside use the transcripts had to be edited. Meetings usually have 1000-1400 turns and are typically between 45 and 70 min long.

1.4.3.4 Santa Barbara

. The corpus has bee published by the LDC and their documentation (LDC2000S85) states that it

> is based on hundreds of recordings of natural speech from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and ethnic and social back-

> grounds. It reflects many ways that people use language in their lives: conversation, gossip, arguments, on-the-job talk, card games, city council meetings, sales pitches, classroom lectures, political speeches, bedtime stories, sermons, weddings, and more. The three CD-ROM volumes in Part 1 contain 14 speech files of between fifteen and thirty minutes each, from the Santa Barbara Corpus of Spoken American English.

A subcorpus consisting of 7 meeting-like situations has been annotated with topic boundaries and activities [7]

1.4.3.5 TV Genre Corpus

A large corpus of television subtitles from various US stations was recorded along with programming and genre information. The programming information has been obtained by querying `tv.yahoo.com` which also provided the genre of the show. 1067 different shows have been collected over the course of a couple of months, a show typically being 30min long. Both types of information where combined into a single database which consists of the program information and the timestamped subtitles and was recorded semi-automatically.

1.4.3.6 Switchboard

The corpus has bee published by the LDC and their documentation (LDC93S7) states that it

> [...] is a collection of about 2400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven ”robot operator” system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.

Switchboard was also used in the DARPA HUB-4 speech recognition evaluations and the author of this thesis participated as part of our working group in the evaluation in 1996. It has been considered the benchmark for spontaneous speech

[7] The dialogue number are 2, 4,6,8,10,13 and 14.

recognition over the last couple of years and received international attendance reaching beyond the DARPA community. This corpus was also used in the 1997 John Hopkins Summer Workshop on dialogue modeling with the application to spontaneous speech recognition in which the author participated (Stolcke et al., 2000). The dialogue corpus constructed in that effort is by far the largest resource that is annotated with dialogue acts, over 1155 dialogues have been annotated (Jurafsky and Shriberg, 1997). It also presents the first use of the DAMSL annotation scheme which is loosely based on the scheme for the TRAINS corpus (Core and Allen, 1997). Dialogue act models have been trained on Switchboard and applied to the English corpora in this study – all dialogue act annotations in this study have been derived using that dialogue act tagging module. The dialogue act models have not been trained to be optimal on Switchboard but rather to have maximal portability.

1.5 Organization of Thesis

Why is anyone interested in accessing databases of oral communication ? Which scenarios are interesting and what has to be observed ? And what have others done in these applications, what is the related work this thesis builds upon ? Chapter 2 presents a variety of application scenarios and discusses the related previous work along with aspects of privacy, manual versus automatic annotation and reinterpretations of spoken communication. The scenarios discussed are broadcasts, meetings, lectures and speeches, tutorials and vague information as well as entertainment applications and actions accompanied by speech. Additionally other related work is discussed, ranging from linguistic theories to industry standards for multimedia.

Chapter 3 presents algorithmic work on dialogue act and game detection. In contrast to previous work the training uses a fully discriminative procedure and an integrated dialogue act and game detection tagger is introduced. It therefore presents the machinery to detect those features that are used in various ways throughout the thesis and presents background reading on the machine learning techniques used. Continuing the classification approach chapter 4 discusses global variations in dialogue style by determining the membership of a dialogue to a certain class and the membership of a topical segment to a certain activity. Chapter 5 – largely independent of chapter 4 – demonstrates how dialogues can be segmented automatically into topical units. Among other results it is shown that dialogical features such as speaker initiative and speaking style can be used very effectively in this task and the chapter adds weight to the importance of non-keyword based indices. Non-keyword based methods for segmentation have the advantage that they don't necessarily require speech recognition systems. Using activities and dialogue style as well as information about speakers, keywords, dominance and emotions chapter 6 makes an assessment how these features make dialogues accessible. This is proven using information theoretic measures, a user study and a description of user interfaces. The final conclusions and an outlook on attractive continuations are offered in chapter 7.

Chapter 2

Applications and Related Work

2.1 Introduction

This is not the first work to consider the problem of information access to spoken data. Sec. 2.2 features a short introduction of application oriented research groups, systems and approaches. The main discussion of related application oriented research will be in conjunction with the discussion of application scenarios (Sec. 2.3). It will include the aspects for privacy, access, automatic and manual annotation and the possibility of "reinterpretation" (reformulating the results of the rejoinder for future use).

On the other hand there is a large variety of other research that this thesis also draws upon and that might be affected from insight of this work (Sec. 2.4): Linguistics (dialogue, stylometry and grammar) and computational linguistics (prosodic modeling, topic segmentation, dialogue modeling), speech recognition (language modeling, dialogue modeling, vocabulary adaptation, keyword detection), visualization of oral communication and information retrieval. The features used in this thesis should also have an effect on emerging industry standards such as MPEG-7 (Sec. 2.5).

2.2 Systems and Research Groups

There are a lot of research groups which claim to work on information retrieval from speech. However, most of them work on Broadcast News which is significantly different from spontaneous conversations. Some work, however, was also conducted on the access to audio records of meetings, typically combining online-notes or slides with the audio record. Currently the most important applications of indexing for spoken language are:

IR from Broadcast News Garofolo et al. (1999) feature keyword based information retrieval on audio databases. On Broadcasts News it doesn't matter whether a machine or manual transcript of the show is being used. The author however has some doubt that this result would allow meaningful generalizations to rejoinders such as meetings (Sec. 1.2).

Systems for Broadcast News Retrieval Wactlar (2000) and Kubala et al. (1999) access broadcast news using keywords, additional features and support visual browsing, including thumbnails of video shots. Whittaker et al. (1999) is working in the same area and is very relevant to user interface interface design. Their SCAN system features the segmentation of speech into intonational phrases, the segmentation into topical segments and the display of salient words over the time of a conversation.

Voicemail Retrieval The integration of various types of media is an interesting task since it allows us to communicate seamlessly even if environment is changing.. Bacchiani et al. (2001) show that voicemail can in part be converted to email and may be accessed via standard email clients.

Online Note Taking Notes digitized on fly can be synchronized with the audio and help to find information (Landay and Davis, 1999; Stifelman et al., 2001; Whittaker et al., 1994; Wilcox et al., 1997). This technique is very powerful since it allows users to leverage their traditional note-taking instruments. Moran et al. (1997) presents a long term user study and reports that the notes are also used to mark locations in the record for later listening. The downside of this technology is that it requires additional hardware such as a scanner integrated in a notebook.

Synchronization with Slides Multiple modalities (e.g. audio, video, whiteboards and most importantly slides and their headings) can be synchronizing using timestamps for browsing presentations such as lectures (Abowd, 1999; Hürst et al., 2000).

Audio Skimming Arons (1994) first demonstrated fast playback in an audio only device in conjunction with topical segmentation. The disadvantage of audio as a linear medium may become less severe if audio skimming is feasible. The insights of this thesis might also be used be used to construct systems which provide very fast playback (Sec. 6.4) since the features that are important for indexing are identified.

Dialogue Analysis Kristjansson et al. (1999) suggests to do a simple dialogue analysis on meetings and results for monologue/dialogue/chatter distinctions have been presented. His result presentation is very incomplete such that it is hard to judge his work [1].

Readability of Speech Transcripts The speech recognition error rates that are state of the art on Broadcast News are as good as perfect transcripts for humans although the users still complained about the transcripts if they are produced by a human (Stark et al., 2000). The reason may be that Stark et al. (2000) did not remove filled pauses, repetitions and so forth which typically shortens the text significantly and improves the readability at the same time (Zechner and Waibel, 2000a). The error regime for rejoinder data however is much worse such that it is unclear whether this result would be the same – indeed the author would doubt it.

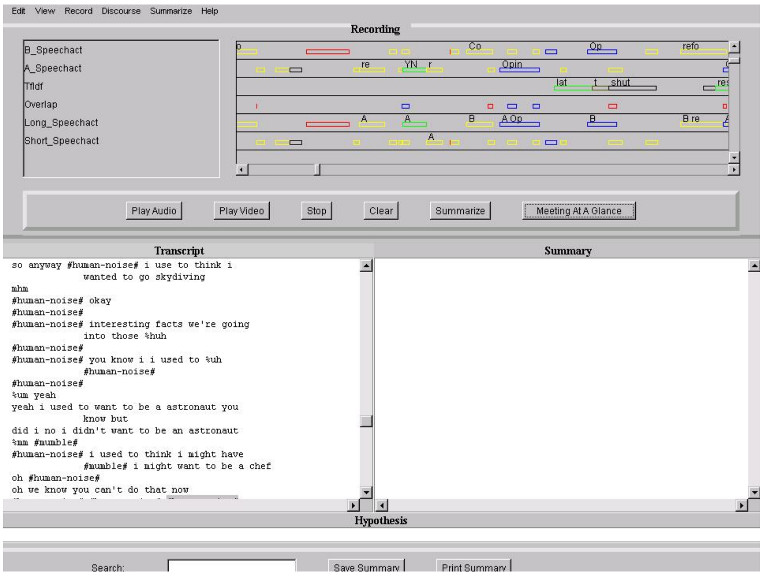
Summarization In our own working group Klaus Zechner adapted statistical summarization techniques to spoken language (Zechner and Waibel, 2000a).

In the commercial world a number of players explore audio and video indexing. The market leader in video indexing is currently Virage (`www.virage.com`). Virage's current key clients include local as well as most major broadcasters that want to syndicate material (Sec. 2.3.3). The introduction of digital production technology might make this market secondary since the indices are more efficiently created at production time. They have expanded their customer base to include organizations and libraries with their own video documentation for speeches, lectures [2] and corporate communication [3]. The features used by Virage's automatic video indexing are mostly video based, interpreting closed captioning, reading on-screen text, recognizing faces and most recently speech recognition. Other applications include music videos and speeches by politicians as well as lectures that are being broadcast (Sec. 2.3.4). The applications do not contain indices that are structured or that represent style and situation unless it is captured in the transcripts. Wordwave (`http://www.wordwave.com/`) on the other hand has grown out of the manual transcription business for entertainment (especially close captioning for the US-TV market), legal proceedings and large corporate customers. In many cases a rough transcript or even worse a rough index as Virage delivers it is not acceptable. The services delivered by Wordwave

[1] The tests were performed on just 4 meetings of 15-45 minutes, intercoder agreement has not been measured, there are no absolute accuracies or entropy reductions available and there is no user study evaluating these categories. Their results are therefore very preliminary.

[2] Virage features the Harvard Business School (`http://video.hbs.edu/`) as one of their customers.

[3] SUN has a web-site dedicated to corporate communication with videotaped interviews, among others with Virage Europe's General Manager (`http://www.sunwebcasts.com/ibc/tech/virage1.html`).



```
so anyway #human-noise# i use to think i
            wanted to go skydiving
mhm
#human-noise# okay
#human-noise#
#human-noise# interesting facts we're going
            into those %huh
#human-noise#
#human-noise# you know i i used to %uh
            #human-noise#
#human-noise#
%um yeah
yeah i used to want to be a astronaut you
            know but
did i no i didn't want to be an astronaut
%mm #mumble#
#human-noise# i used to think i might have
            #mumble# i might want to be a chef
oh #human-noise#
oh we know you can't do that now
```

Figure 2.1: **Meeting Browser (developed by Michael Bett):** The display of the dialogue features can be zoomed in or out in order to get a global view of the meeting or a more local view. Statements are blue, questions are green, negative answers are red, other answers are yellow, backchannels are blue. The display shown here is a reduced display and we can display all of our other features as well. The meeting browser has also been used as a tool to visualize and demonstrate our discourse annotation and to play back segments.

may extend into a high quality broadcasting segment. Interestingly the service of Virage partially depends on close captions provided. One may characterize Virage as the approach to produce rough indices good enough for recall for a large number of documents while Wordwave attempts to create and improve a small number high quality document. Other companies such as Lernout&Houspie, BBN-GTE (now Verizon) and Compaq try to occupy this space as well and demonstrate systems which seem to be aimed at immediate commercialization. AT&T and Xerox (at Xerox-PARC, FX-Xerox and Xerox Research Europe) have prototypes of systems although these systems are more research oriented and analyze conversations rather than broadcasts. All of these markets are also courted by the high-end computermakers since they require enormous investments in equipment.

Information retrieval is an interactive task which seems to be particularly true for the retrieval of oral communication. The problem can be divided into a within conversation navigation and an across conversation navigation problem. Waibel et al. (1998) developed a meeting browser in our working group (Fig. 2.1). The meeting browser supports within conversation access by combining displays of transcripts, audio playback with highlighting of the transcript, video playback and a zoomable display panel of dialogue and emotion features. Across conversation retrieval is supported by a couple of access methods: A conversation can be found like a normal file on a file system but it can also be queried using the names of the participants, the time it took place, dialogue features and keywords that occured in the conversation.

2.3 Information Access Applications

2.3.1 Introduction

We have all lived well without systems to access oral interactions so far, is there really a need for systems that record and index them? Or is it a dangerous development and we are drifting towards an Orwellian scenario? The only way to address these questions successfully is to discuss potential applications. While infrastructure for storage and transmission is largely application neutral the question of information access is more dependent on the actual application: The features that are useful, the features that are of interest, the information that is available about the rejoinder as well as the information need are all dependent on the application scenario.

Applications could be the documentation of *social contracts* as they are often negotiated in meetings; *information* as it is conveyed in tutorial sessions, "tours for a newcomer" or speeches; *fact findings* as exercised in (formal) discourse in courtrooms, meetings, professional/lay interactions or debates; *actions* that are either carried out by reaching a conclusion, issuing a command to someone or *actions that are accompanied by speech*. *Broadcasts* may be important independently since they have been viewed by many people and are therefore part of the socialization.

These applications arise in a number of typical speaking situations and this chapter will analyze some of them: Meetings (Sec. 2.3.2) show elements of most of the abovementioned applications. Broadcasts, lectures and speeches have been studied in the past and are reviewed in Sec. 2.3.3 and Sec.2.3.4. Furthermore vague information and tutorials (Sec. 2.3.5) are another opportunity where the use of audio documents might be very successful. Sec. 2.3.6 will discuss the application of continuous audio and video recording to collect a lifetime of human experience and the recording of action accompanying speech.

For each of these scenarios aspects of *privacy* will be discussed as well as the prospects of collecting additional *manual and automatic annotation* for indexing. The process of transforming an oral communication into a written document does not only result in the memorization of the event but also in an *reinterpretation* of the event. Reinterpretation is a natural and necessary process and it will be discussed for each of the scenarios.

Another application of shallow dialogue processing is pragmatic understanding for automated agents that can enter the personal space of a human. It would be highly desirable for these agents to apply common rules of courtesy and enter a social space only if appropriate. This seems to be important for robots as well as for mobile devices that are always in our personal area, such as cell-phones (see the TEA project `http://tea.starlab.net/`) or personal digital assistants which may notify us in different ways or choose to postpone a notification. These applications may benefit from an analysis of the environment in terms of the social situation and the style of the interaction: properties that are discussed and used in this thesis.

2.3.1.1 Privacy

While there is a tremendous advantage of making certain communications retrievable we might not be happy about making all of them available to anyone. Widespread surveillance as allegedly done by secret services and on corporate systems as well as the tracking of users on the World Wide Web make it obvious that privacy is more and more a construction that has to be negotiated, actively or implicitly (Cranor et al., 2000; XNS, 2000). The recording of conversations, especially of audio recordings, is subject to specific constraints in many legal codes and requires the consent of the individuals involved. Any deployable or even research design has to keep these constraints in mind. This has also been a concern in ubiquitous computing at research centers such as Xerox-PARC and Belotti and Sellen (1993) established some general guidelines and rules that establish a fine grained array of mechanisms which answer to many privacy concerns.

On the other hand we are often willing to give up some privacy if there is an obvious reason to do so or we have no reason to assume abuse: If I order a product over the phone I definitely need to give up my name and address and if I prefer to pay by credit card I need to give that up as well. At the same time I give up the information which products I order or which products I am interested in. The easier it is to give up that information the easier the transaction or information need can be fulfilled. E-commerce sites compete over the simplest possible access methods that would still ensure security. The discussion how privacy on the Internet is preserved while still satisfying commercial and practical demands is undecided. Given the premature status in this high-profile area the situation in the documentation of spoken communication is naturally less mature and this thesis can only scratch the surface of the problem. Similar to privacy on the Internet systems should only require to compromise as little privacy as possible to serve a specific task that is in the interest or the nature of the activity of the participants. As a minimal requirement this entails that the systems have to proof their usefulness, make their presence known and provide means to ensure that unwanted recordings are avoided.

Not all applications however intend to respect the privacy of the individuals, seek consent for recording or make their presence known, whether the application is endorsed by law or not. Law enforcement and intelligence agencies for example want to collect information about individuals, corporations and governments. This data is usually collected without the knowledge of the target group and the privacy of that group is obviously breached. Sensors such as microphones are mounted on virtually every computer and given todays lax computer security measures it seems feasible for individuals to break into selected computers and use those sensors for surveillance: Lax computer security of network connected devices translates into security breaches of private and office spaces. Indexing and filtering techniques for this data could vastly improve the effectiveness of this surveillance. These technique also help to reduce the bandwidth needed for the transmission therefore lowering the footprint of the surveillance and enhancing their effectiveness. The use of cell-phones and mobile computers could pose additional privacy and security threats: A security breach in an organizer with a microphone and a wireless connection could translate in the ability to monitor the microphone continuously. Given the tremendous capabilities of attacks on sensors combined with audio and video information processing technology this issue has to receive special attention. On the other hand all of these fears are just as relevant about the surveillance of email and Web access behavior where the technical requirements are a lot lower.

Applications which indiscriminatively record individuals lives seem to be another problematic application: Firstly only the person who is recording maintains control over the record and random people can appear on it. Other individuals would not always be aware of the recording or they might not be able to avoid it for practical reasons. Secondly the pervasive availability of audio and video sensors opens the door for massive privacy breaches by correlating a range of sensors especially if the security of the recording and the transmission to main servers are not strictly enforced or sensors could be breached.

The analysis of dialogue and the corresponding filtering techniques however can also be used for the good: If I need a sensor for “X” – e.g. the status of a room – I may be able to implement that sensor using a microphone and some analysis algorithm. If the sensors output is only the relevant feature the interesting task has been solved. If however the output is the original audio which is available on a standard device one could make use of it by other means or it could be tapped by someone else. The acceptance of sensors such as cameras and microphones may depend on the control users can exercise over them and compartmentalizing the usage appears to be an option.

2.3.1.2 Manual and Automatic Annotation

As important as it is to generate as much using automatic analysis any reasonable system would attempt to encourage the input of indexing information when convenient and to extract as much information as possible automatically from digitized information in the environment. The user interface of the indexing system and the integration with other systems are important parameters for the success of a realistic indexing system. In a meeting system for example the appointment calendars and notes of the individuals could provide important information, in lectures the slides and the times at which they have been shown could be important. This aspect will therefore also discussed for all application scenarios.

2.3.1.3 Reinterpretation

Oral communication itself is transient such that it needs to be recorded (storage). However there are also many reasons to process a verbatim recording (reinterpretation, see also Sec. 1.3):

- retargeting the audience beyond the participants since it might not be understandable for non-participants (overhearer effect, see Fig. 2.2).
- extracting and condensing information to retain only the relevant portions.
- adding or deleting.
- changing the meaning.
- citing the rejoinder.
- relating the rejoinder to other rejoinders or documents.

The standard documentation of a spoken communication is a written document which entails that both storage as well as reinterpretation are conflated. Audio recordings definitely allow to automate the storage aspect of documentation but they also support reintepretation: Moran et al. (1997); Whittaker et al. (1994) and Wilcox et al. (1997) observe that users of audio-recorders that are coupled with note taking devices tend to annotate less during the rejoinder but improve the notes afterwards. The addition of an (indexed, digital) audiovisual record adds the capability to cite that record and use it for the production of other reinterpretations. The audio record retains many qualities of the original spoken communication that are commonly not reflected in a written record since written language is usually aimed at a more precise and unambiguous description of social contracts, fact findings and so forth.

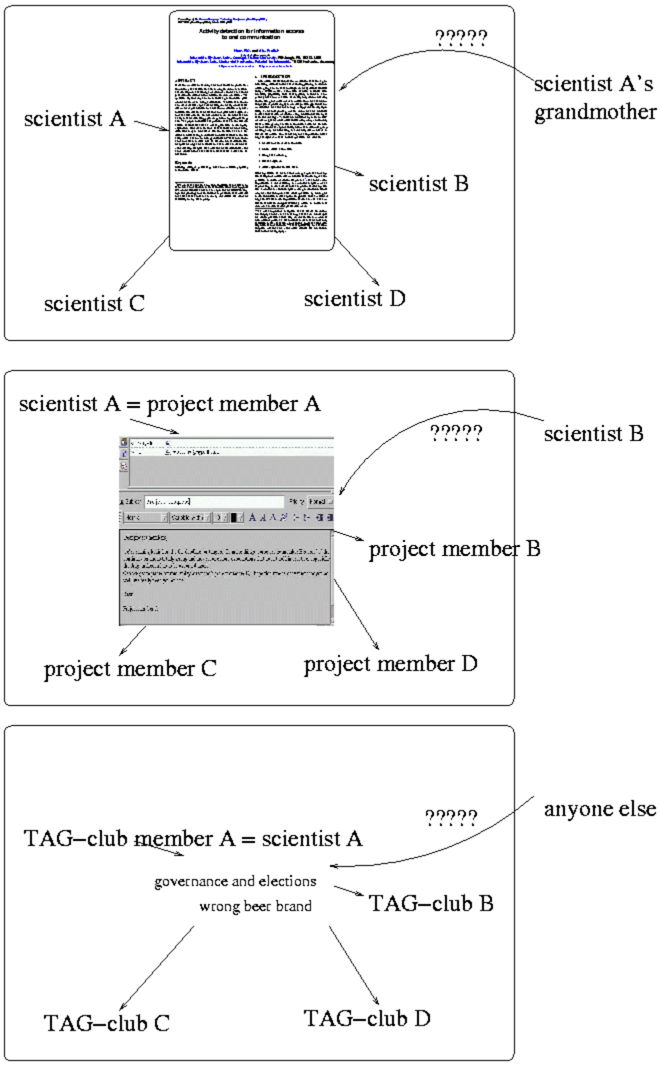


Figure 2.2: **Overhearers of Discourse:** Clark (1996) provides a review of how different groups participate in dialogue. Person A is part of different groups (a scientists, a member of a family, a working group, social circles etc.) and the utterances/emails/papers he produces are designed to be optimal for the communication within that group but may not be understood in other communities. An overhearer, even if part of other dialogue contexts of A, does typically not understand an individual utterance or conversation. If A is using spoken communication in one of his contexts immediate feedback is available to resolve potential misunderstandings which reduces the need for the explicit resolution of references. Written information is usually designed for a larger less specific audience and it is therefore more carefully crafted to be understood by the target audience, especially since no immediate feedback is possible. Reinterpretation is necessary to resolve the references in spoken communication to retarget it for a larger audiences or to maintain the memory for these references for the original group.

2.3.2 Meetings

Meetings are a social encounter of special importance: Companies, administrations, research groups and many other organization hold meetings. The meeting genre is very flexible and is used for fact finding, the construction of social contracts, conveying information and socializing. Meetings are often found worth documenting however their documentation is expensive since it takes the attention of participants during the meeting to make notes, it takes a lot of time to prepare minutes of a meeting and it takes even more time to circulate minutes of notes to achieve agreement. Given the importance of meetings and the inherent difficulty to document them any automated support that can increase the depth and efficiency of documentation, its accuracy and timely delivery is useful. The most interesting empirical study on using audio records is Moran et al. (1997) on meeting minute construction. Their system combines the audio with notes that were time-stamped such that they could be used to navigate in the recording. They presented a long term user study that indicated that their system is effective. However, their system did not add indices automatically, the only available index was the time-stamped annotation of the user.

There are few empirical and very few empirical linguistic studies of meetings itself. Kristjansson et al. (1999) is probably the only study that attempted to use higher order structure on meetings. However many of their claims are not supported by empirical work, the only results that can be related to discourse is a discrimination of monologue/dialogue/chatter – and even that result is hard to interpret. From a more linguistic angle Bargiela-Chiappini (1997) analyzed professional meetings which are very different from the meeting or Santa Barbara database we have been analyzing. It is not obvious that the participants of Bargiela-Chiappini (1997) maintained contact with each other on a day to day bases and shared private information in that context. Their study also involved a significantly smaller number of meetings of related groups such that the diversity of the database is not ensured. The problem of their work is that it is not clear how it may be repeated and the analysis doesn't apply techniques that seem to be generalizable. The most important empirical analysis parameters in their feature set that can be reproduced are long/short turns, pronominal usage (especially I/we) and speaker overlap. Additionally they investigated the generic structure of the meetings and the role of it in organizational communication and ascribe roles of participants given their position in the organization (Bargiela-Chiappini, 1997, Chapter 1.3, "Language and organizational communication").

Meetings may show a certain generic structure, namely introductions of group members, presentations, open discussion, action items, conclusions, setting up follow-up meetings and so forth (see also Bargiela-Chiappini (1997)). A summary of a meeting could therefore be composed of the action item or conclusion sections of the meeting. Many meetings feature multiple media, especially slides in presentations, which might be used to enhance the index and the recording. While this information seems to be important to analyze none of the meetings in the database analyzed in this thesis contained interesting examples such that this aspect could not be analyzed. While this is regrettable it also shows that these seemingly important indices do not exist in all meetings.

2.3.2.1 Automatic and Manual Annotation

Meetings are often organized and handled by support staff, materials are being produced and circulated ahead of time and the times, locations and dates are entered in productivity applications. All of this information could be made available and it could be used to document the rejoinder. A complete commercial grade meeting documentation system would have to make all of this information available. This aspect is not explored in this thesis but it is necessary to realize that other indices may be available – the goal of this thesis however is restricted to the exploration of non-keyword based features. The complexity, software engineering needs and costs of these systems are tremendous making it difficult to explore them in academic or commercial research and development environments. Additionally the availability of these indices varies from one setup to another such that a throughout assessment is difficult. The research strategy adopted in our group was therefore to develop key capabilities which would most certainly add to an overall system (Sec. 1.2).

While the meeting record would also benefit from the ability to access all written notes from the participants, all slides and all drawings on boards, the timing information is crucial to correlate such events with each other and the oral communications surrounding them. The simplest model for attaching time stamps might be to digitize notes, drawings and slide operations on the fly such that the time stamps can be recorded automatically. These ideas are explored by other research groups:

online note taking People often take notes during meetings and one person might be assigned the role of the note-taker. Technologies that allow the fast input on text on small mobile devices [4] and the integration of a “note-taking work-

[4] The PalmPilot is one of the most popular mobile devices in the year 2001. While collaborative note taking is explored by some research groups a major issue – the speed of text entry – has been largely ignored. Speech recognition is not a solution here since the participants would need to talk while another member is presenting which would interfere with the meeting. Fortunately the introduction of small portable QWERTY keyboards either as part of the device or small add-ons and other fast (up to 40-50 words per minute for average users) text input methods (Isokoski, 1999; Textware Solutions, 2000) are about to eliminate the bottleneck seen in the first generation of small portable devices. It can be assumed that in the very near future these devices will be able to form

station" into the scenario would allow to synchronize notes that carry time stamps with the audio signal. The synchronization of audio capture with meeting notes was the focus of many projects in the past such as Moran et al. (1997); Stifelman et al. (2001); Whittaker et al. (1994); Wilcox et al. (1997). Long et al. (1997) in fact shows that PalmPilot and Apple Newton devices are used primarily for the purpose of taking notes. Collaborative note taking has also been explored by Landay and Davis (1999) and is a powerful technique by itself even without audio recording: A note taking application that provides notes with timestamps and a "privacy" checkbox could therefore provide effective and cost efficient indices that could be gathered by a "note-taking workstation".

slides and electronic whiteboards During a meeting slides might be shown or a whiteboard might be used. Abowd (1999); Hürst et al. (2000) record the slide changes and digitize drawings on whiteboards since the presentation aspect is prime in lectures. Lectures and presentations often follow a specific generic progression and that information could be used to identify important information. A simple yet effective example is the use of slide headings with timestamps. Since the replay of slides and white board drawings is not as easy as the display of textual notes the applications need to be able to support both the recording and extraction of important events such as slide transitions and titles as well as a time synchronized replay facility for the slides and white board events themselves.

specialized systems A very specialized meeting room may support the automatic collection of votes, distribution of speaking rights to microphones and so forth. All of these events should be collected since they can either be added to the record as a new event or they may support techniques such as speaker identification.

buttons A meeting recording system should at least support the use of an explicit "off the records" button and additionally "action item" and "conclusion" buttons. Landay and Davis (1999) has – even in the note taking environment without audio recording – observed that users wanted to be able to distinguish shared and private notes and installed a "private" button.

It may not be necessary to install physical buttons but use cue phrases that are natural on one hand but on the other hand serve as commands to the meeting recorder. The system may also restrict the input to certain microphones ensuring that the recorder does only interpret speech that is intended

local wireless networks using infrared technology or radio transmissions such that the hardware for collaborative note taking will be provided by the mobile users themselves.

as a voice command. A physical button can have an on/off state whereas a voice button is a one time command. The system has to interpret when the indicated segment starts or ends. Cue phrases such as "action items", "let's do that", "this is what we have to do", "can you do that ?" also indicate that a section before or after the cue phrase of a specific length may be important. This issue should be handled in the context of a system design in conjunction with the user interface since it has to be seen as a command to the program.

2.3.2.2 Access and Privacy

Access and privacy are very important issues when recording real world meetings. Even for research purposes we have found it hard to gain permission to record meetings and the use of most of the meeting data recorded in our group is restricted to the use within our own group. So far the problem of access and privacy has not played into our own system designs since the intelligent meeting room system has not been fielded (Sec. 1.2). As soon as this happens the author assumes that a number of questions will arise immediately:

- who has access to the meeting other than the participants
- are there active decisions during the meeting what should be recorded
- is postprocessing/reinterpretation done for sure [5]

This technology will be used in different and unforeseeable ways and long term studies of deployed systems in multiple groups would be necessary to precisely describe how they are used and what level of privacy and access control need to be exercised. Four basic strategies however may be assumed to illustrate the space of decisions to be made once these systems would be fielded:

just let it run Once the formal part of the meeting begins the recording is started, probably when the formal introduction occurs. This strategy might be good for very formal meetings where the participants have to maintain a perfect face or maintain that their meetings are open to the public.

let it run and annotate later Meetings that need a high quality documentation could be annotated by hand by a designated minute taker. Exclusion of segments, segmentation and summarization could be handled by that person, possibly supported by automated techniques. The result could also be a written document that contains links into the meeting to provide support for

[5] Postprocessing and reinterpretation are currently not actively supported in the intelligent meeting room system or the capabilities are very rudimentary (Sec. 1.2).

the minutes produced and achieve better agreement when the minutes are circulated (see also Moran et al. (1997)).

record but indicate In many meetings a lot of discussion might be irrelevant for the documentation but the participants know what the relevant items would be. Nevertheless, for the most part, they agree that the recording may be published and in rare cases some parts need to be excluded that should be off the record. Since the whole meeting is recorded the snippets don't have to be captured precisely by the buttons and the automated analysis can be used to enhance those.

record only what is important Informal meetings may contain a lot of irrelevant discussion and information. Subjective, emotional or personal information might be shared that is neither intended for a larger group nor for any record. It therefore seems to be unnecessary to delete those portions later on but rather start the recording with some button. Usually meetings contain a session where the participants draw conclusions from the previous discussion: If they learn to just record that section it could serve as an excellent summary.

2.3.2.3 Reinterpretation

The most common way to document meetings is to write minutes. Minutes are only a reinterpretation of the meeting by a single person and it is therefore common that minutes are circulated and agreed upon. In some meetings the minutes itself might be subject to discussion in a meeting. Yet in other situations the results of meetings, even if they are established as a result in the form of minutes, are being discussed.

The process of minute construction requires knowledge of the meeting and an audiovisual record can support this process. Adding search capabilities for the record and the capability to construct a multimedia document with pointers into the original record (Moran et al., 1997) could be supported by an automated system. Non-keyword based indices might support the minute construction process by providing information where a topic starts and ends (Sec. 5) and what kind of conversation is being carried out (Sec. 4). The identity of the active speakers would also likely be important.

Since minutes may represent only one view of what happens they may need the agreement of the participants. This agreement process – minute circulation – can be very cost intensive especially if delicate or far reaching interpretations of the rejoinders are possible. The experience of Moran et al. (1997) is that the use of audio records of meetings was an assurance for the participants that the minutes would accurately reflect the statements in the meeting. Further trust could be built if the production of the minutes would implicate that important statements are backed up by audiovisual citations – much like citations in written language lend trust to the document. The minute construction process would therefore be more objective and circulation could be more efficient. The role of the reinterpretation in the document creation and circulation process was also researched by Stifelman et al. (2001); Whittaker et al. (1994); Wilcox et al. (1997) who demonstrate the effect of personal minute construction and "after-meeting" editing in the presence of audio recordings.

2.3.3 Broadcasts

The information retrieval community discovered that a large range of interesting documents might be available in audio or video form and therefore addresses the problem of information access to multimedia databases. Implicitly the information retrieval community carried over basic assumptions about the relevant indices for documents which were successful in text based retrieval: The most important and usually the only index is a bag of keywords extracted straight from the data. In the TREC evaluations and conferences audio broadcasts have been indexed using standard keyword based information retrieval technology (Garofolo et al., 1999). The focus has so far been to achieve the same information retrieval results from speech recognizer transcriptions as from manual transcription which can be seen as successful. However it is not clear whether information retrieval should be based on keywords alone and this thesis argues against it.

Kubala et al. (1999) and Whittaker et al. (1999) present work on the access to broadcast news and are more oriented towards the audio aspect of the retrieval problem than Wactlar (2000). BBN-GTE (now Verizon) focus on named entity extraction (Kubala et al., 1999) while AT&T focussed on the segmentation into intonational phrases and topical segments (Whittaker et al., 1999). Companies such as Virage focus on indexing using closed captioning, on-screen text and face-id and apply traditional search techniques to these. Additionally the "encoding" process may be enhanced by manual labeling using a special browser. Virage uses the well established keyframe detection techniques and recently added speech recognition capabilities. The key market for Virage is the syndication of broadcasts which requires the material to be available in indexed form. Indices therefore add significant value to existing material by making it available for sale.

Broadcasts are very different from the meeting genre described earlier, they are neither a rejoinder of people bound to a physical location or medium nor do they necessarily depict them. They don't contain a social contract nor convey information nor engage in fact finding – at least they are not doing it in an interactive fashion with the recipient. There is no backchannel, no possibility to engage in an interactive dialogue, nor can the recipient take control or disagree explicitly. Broadcasts are (with very few exceptions) a one way medium much like newspapers and books, you may enjoy them or switch them off. The language used has to be understandable for a heterogeneous audience (decontextualized) which requires some of the same attributes that make written language what it is. Broadcasts also contain explicit structuring elements for consumers such as “cuts” between different “shots”, schedules, (theme) music and so forth which may be used for automatic indexing.

Given these observations work on broadcast speech – important out of its own right – may have little importance for other spoken genre which are more interactive in nature and which are not highly processed.

2.3.3.1 Analyzing the News Genres: *Informedia*

Wactlar (2000) describes the *Informedia* project which uses the location of a news event to index the event along with the topic of the event, the key players and the time it occurs. It therefore makes a significant step over Garofolo et al. (1999) since generic information is used. Interestingly the dialogue situation that is presented by the media – a presentation of the newscast to the audience – is not very useful for indexing. However the presented situation or event is crucial for generating the index: It is commonly assumed that the event occurs shortly before the report but that depends on the type of the report: Very close to the event initial speculation and factoids may be dominant. Later the reports might become more reflective over time as more information becomes available and would eventually summarize the whole event and put it into perspective. Other types of reports could be the expert discussions, “live on the scene” reports, interviews, speeches, expert statements, press conferences, commentaries, comic depictions and so forth. All of these reports occur in a certain timeline and are related to each other – they form a meta-genre that is associated with news reporting and a system that can access all of these records should be able to collect and organize them. If an analyst would have this information it would support search strategies such as moving from summary judgments to individual factoids or opinion pieces. The genre carries information where a specific type of information can be found and how it has to be evaluated – this aspect of genre is however underappreciated for the broadcast meta-genre.

2.3.3.2 Automatic and Manual Annotation

Broadcasts are very expensive to produce so it is mandatory that they can be reused as much as possible yet many local TV-stations have only limited infrastructure to support effective indexing. Manual and automatic indices might be applied most effectively in the production process (Sec. 2.5): The video camera could record information including the location, angle, time, objects, peoples and events depicted and the information would be passed along from the video camera through production and editing to the archive. Similarly text read by an anchor speaker, the production schedule, the names of the cuts, the names of the speakers featured and so forth are known at production time and could be made available by an archive system.

Still many outlets are incapable of handling a production process which supports indices during production and there are huge historical archives that could be harvested. This holds especially for local TV stations with little infrastructure but it may also hold for heterogenous media organizations with multiple media properties and highly complex production systems. Automatic indexing which is applied ex-ante may therefore still be important for a couple of years to come. Independently the indexing of non-cooperative institutions in commercial, military and intelligence applications will continue to require automatic ex-ante indexing.

Broadcasts however have another important property, they are not isolated but reflected in multiple facets since they address large audiences or even entire nations. A broadcast of news is usually correlated in time with a news report in printmedia or on WWW sites. One simple yet effective search method for videos of news could be to search using a text based search engine and hope to find pages which point to audiovisual material – since this is standard for certain websites one could also restrict the search to those. Another options is to exploit just the temporal aspect of written news reports: One may start a standard search for news stories, evaluate the temporal profile of the stories retrieved and use that as additional information for the retrieval of audiovisual material. Other information such as programming information is available readily, often already in electronic form, for example for all US TV markets at `http://tv.yahoo.com/`. A classification of the sub genre of a broadcast section can also be done automatically with reasonable accuracy (Sec. 4.4). Many broadcasts also contain subtitles, making the indexing without a speech system even easier and more practical [6]. Given the plethora of other information it may be very practical to leverage classic keyword based search engines to solve retrieval problems in Broadcast News.

2.3.3.3 Access and Privacy

Privacy is not an issue when accessing broadcasts – they have been previously found presentable in public on another medium. However the broadcaster would like to be able to receive revenues from their expensive productions and therefore restrict access, collect royalties or couple the access with advertisement. This thesis is not concerned with these issues although the success of a commercial

[6] Even if closed-captioning is available speech recognition might be important since a close caption might be shifted by a couple of seconds relative to the audio event (Wactlar et al., 1998).

product might well depend on it. Producing high quality indices and summaries of news stories is obviously a service that adds to the quality of the product so it is likely that broadcasters themselves will explore this opportunity.

2.3.3.4 Reinterpretation

Broadcasts implicitly reinterpret themselves. An event or a situation is reported upon and over time the event is commented in more abstract categories, generalizing over multiple events and bringing various informations and interpretations together. As indicated earlier generic information might help to structure and qualify the type of reinterpretations that are available yielding a more effective information presentation: If a user wants to access a news event of the past they want to be pointed to a highly condensed neutral reinterpretation of it first, be able to view the different opinions on it and finally trace the developing story as a whole. On the other hand there is no reinterpretation of broadcast information by the viewers. This situation might change over time using interactive feedback mechanisms and automated methods to gather, summarize and manage feedbacks of large audiences. Broadcasters such as CNN recently experiment with chat rooms that they opened for specific political topics and included feedback from those moderated chats into the live broadcast program. One important reinterpretation is currently not being addressed in a suitable way, namely the personal reinterpretation of broadcasts which may consist of annotations, bookmarks etc.

2.3.4 Lectures and Speeches

Lectures and speeches are usually held in public and require a lot of preparation; they are designed to carry importance to a lot of people and often achieve that goal. Both events are oral in nature, have rhetoric qualities, feature intonation, pace, gestures and timbre of the presenter. All of these features might be important to fully understand the presentation. A written summary has to take away most of these qualities and replace them with other devices which is time intensive and changes the original quite dramatically. It might therefore desirable to provide audio and video recordings of the events additionally or instead of written documentation. Low cost recording and indexing would enable the documentation of many speeches and lectures which would not be documented otherwise. Speeches or other semi-public material is also essential for the communication in large cooperations or for maintaining investor relationship. A lot of companies supply audio-visual material of investor conferences over the Web as a new form of documentation that may be more convincing and personable than press releases [7].

Abowd (1999) and Hürst et al. (2000) allow lectures to be presented in multiple modalities, edited, played back and delivered with little effort for the professor and the students. Both approaches rely heavily on the slides and their headings to structure the lectures. Other interesting annotations might be question and answer sections, announcements by the instructors or students, a high noise level in the audience, the overall level of interactivity and so forth.

So far the question of retrieval for lectures has focussed on local navigation within a lecture. However, as more lectures become available online, finding those lectures on the Web and finding the appropriate lectures is an unchartered terrain. Lectures need to be indexed by department, institution, institution type (commercial, academic, ...), audience (semester, program, prerequisites) etc. The author has at least found the following types of lecture presentations on the Web:

- lecture home pages
- book-like presentations, possibly with animations
- powerpoint, pdf and postscript slides
- lecture notes (by students or lecturers)
- multimedia presentations (Georgia Institute of Technology; Multimedia, Teleteaching and Electronic publishing group, Department of Applied Science, University of Freiburg)

Currently it would be fairly hard to find specific lectures that might be related to someones interest using keyword based engines. Table 2.1 shows a distribution of power point lecture slides by web site region that were found using a simplistic method. It indicates that there are already a number of lectures available despite the suboptimal search technique and that a categorization of lectures would be an advantage.

2.3.4.1 Automatic and Manual Annotation

Both lectures and speeches are often transcribed and annotated manually. In many cases they are announced and therefore the announcements may serve as indices. The speakers are often well prepared and have a concept if not the entire presentation ahead of time, often in the form of slides (see also the discussion of

[7] SUN has a web-site dedicated to corporate communication with videotaped interviews (`http://www.sunwebcasts.com/ibc`). Even with a small number of broadcasts they needed a retrieval method and organized the audiovisual documents hierarchical.

site	count	site	count
edu	146	edu.albany.www	17
com	58	ie.econ	10
edu.gatech.cc.minnow	36	nl	9
edu.neu.ptd.www	26	edu.cmu.ecom.www	9
edu.cs	24	edu.sg	7
edu.au	23	edu.ucla.ioa.tyr	7
edu.uk	19	ca	7
edu.albany.rachel	18	edu.gatech.cc.www	6
hk	18	net	5
de	18	org	5

Table 2.1: **Powerpoint Lectures by Origin:** Web locations with powerpoint slides for lectures were found using a query to the search engine http://www.google.com/ with the keywords "lecture slides powerpoint" in September of 2000. The domain name cluster lend themselves to interpretations about which language to expect, what the general area of study is and what kind of presentation has to be anticipated. Using a couple of simple heuristics sites were hand-clustered according to their domain names (groups with less than 5 members are not displayed and represent less than 5% of the 497 power point slides found).

presentations in Sec. 2.3.2). Public speeches by politicians are often accompanied by a press release which might contain the speech and a summary.

Abowd (1999) mentions that there is usually a certain amount of manual post-production editing involved to make the presentation more suitable for later viewing. The instructor of a course can only generate maximal benefit for the students and for his own efficiency if the students don't need to contact the instructor and the students can generate benefit from the recording. The incentive for the instructor to provide a good product is therefore high.

2.3.4.2 Access and Privacy

Speeches and lectures – just as broadcasts – are usually designed for public consumption although there might be some lectures or speeches that are only addressing a certain group or that may carry confidential information. The attendants of the lecture or speech, although their identity can usually not be reconstructed, will have to be informed of the recording. As in broadcasts copyright and licensing issues may also influence access schemes which is not a concern for this work.

2.3.4.3 Reinterpretation

Lectures are often repeated, they may start as seminars or scholarly discussions, become lectures, lecture notes and eventually text books. Lectures are often part of a curriculum, they are repeated and usually belong to a course. Lectures reinterpret themselves over time, e.g., a lecture this year will be reinterpreted next year in the repetition of the lecture. An interesting application would therefore be to compare the same lecture over time, at different schools, by different teachers etc. Students might benefit tremendously since they might get a better perspective on the material by an instructor that is following their line of thought or by multiple perspectives on the same or similar material. The most interesting reinterpretation however would be to see an instructor alongside with the books or written materials that are being used. Students could then read the book or follow the lecture and adjust the instruction modality and speed to their personal preference.

Speeches may also be reinterpreted, often via press releases or in broadcast media. It would therefore be interesting to integrate the speech with those immediate reinterpretations and link those to reinterpretations within a larger news story (see also Sec. 2.3.3).

2.3.5 Vague Information and Tutorials

People communicate vague information and short tutorials on “how to do something” every day and longer tutorials are also fairly frequent. Similar random meetings on the hallway, the office, the coffee-room, over lunch – all of these produce relevant information but we would never start writing minutes about them. Most of this information is therefore lost since and it has to be reproduced over and over again. It may also contain very important information which is completely forgotten and cannot be reproduced. The ability to capture this information could therefore vastly improve individual and organizational memory and reduce the time individuals use to present that information. The written form is also unsuitable for vague information since it implies a certain degree of precision. The quantification of uncertainty is easier in spoken form since more devices such as prosody and hedges exist. The value of vague information and tutorials is often underestimated or it is unclear at the time the information is revealed. In most situation one would just record the information and may either construct a written report when the need arises or search using automatically generated indices (Fig. 1.3) [8].

[8] The author of this work wrote a toolkit for language modeling that was not intended for public use. Over time however it was widely adapted within our lab and even beyond. The colleagues wanted to know what the steps were necessary to build a simple language model for a speech recognizer and what the pitfalls may be. The users also assumed that the author of the tools would

2.3.5.1 Automatic and Manual Annotation

Vague information or tutorials can't be annotated with much additional information since they occur ad hoc and can happen in many environments. In general the need for manual annotation has to be kept fairly low since it is often unclear how valuable the information is going to be. Probably the easiest environment to control is the office environment where a personal computer with a microphone and a keyboard can be assumed – video cameras are currently starting to become more popular as standard computer equipment. A basic system would require information about the meeting such as the participants to be entered and would record the rejoinder along with timing information. The location and time of the rejoinder is given implicitly and can be stored as well. Mobile devices for note taking or a note taking application on the office workstation could be integrated (Sec. 2.3.2).

2.3.5.2 Access and Privacy

When vague information is conveyed, humans might be reluctant to give access to a larger group of people. In the tutorial case the person that is conveying the information will usually determine to whom to present that information. The presenter is becoming more efficient since he does not have to convey the same information multiple times: The presenter would usually initiate the recording and add basic indices. The person that the information is conveyed to would usually have to agree to being recorded. A simple way to register their approval would to record an oral approval at the beginning of the recording.

deliver direct support for the tools. The standard approach is to write a technical document about the features of the software but it was completely unclear at the beginning that this document would be needed. Instead of supporting every user separately by email or in person a "howto" document was constructed that covered the essentials. The documentation was constructed as needed by individual user requests, but the author took a little more time to construct a more general tutorial instead of just answering the specific request. This approach required some foresight and additional time of the documenter. An audiovisual documentation system would have been able to capture the information effectively and with less effort. Also it would have captured more information that did not make it into the "howto" documentation and the documentation would have been available earlier. The "howto" documentation also addressed the user needs better than a standard technical document since they were not interested in getting full insight into the tools or the technical aspects but rather wanted to solve a limited language modeling problem. The author was therefore able to effectively support a large number of users with little effort since the answer to requests were either provided and found in the documentation or were in the documentation but have not been found by the colleague. An audiovisual documentation system would enable this kind of support in more situations, with less effort and at greater depth.

2.3.5.3 Reinterpretation

Vague information and tutorials are usually not reinterpreted, they are actually rarely recalled at all. However when they are recalled the information might be reinterpreted, summarized and commented on. Since systems of this kind are not deployed there is no good way of telling what the usage strategies might be. These strategies will also depend on the people using them, the capabilities of these systems and their user interfaces.

2.3.6 Constant Recording and Action Accompanying Speech

Wactlar et al. (2000) explore how lifelong personal experiences can be recorded and indexed. The idea is to equip a human with a video camera and microphone which are recording constantly. A GPS receiver delivers the location of the individual and the individual can add voice annotations to the record. Additionally the temporal information can be stored easily. One current project focus seems to be the creation of panoramic views by integrating multiple camera shoots. The work in “experience-on-demand” – although a technically viable concept – leaves serious questions on privacy as well as usefulness open.

An interesting twist on the constant recording approach is to record only speech by the user: Fritz and Hundschnur (1994) describes speech as either task-oriented, relationship-building or action-accompanying. The user could build a log-book or record of his or her daily activity by “thinking out aloud” activities. Some actions are accompanied by speech anyways and it may be easy to produce more speech like that. This approach might be interesting for professionals that act in environments that are hard to control: A mechanic in a plant may document his or her actions and would keep a rich record that also contains precise timing information. This record would deliver additional information about the operations in a plant and can be used to analyze day-to-day operations as well as forensic investigations such as professional misconduct or the analysis of catastrophic events.

2.3.6.1 Automatic and Manual Annotation

Wactlar et al. (2000) already use GPS systems to record the location of the individual, the viewing direction could also be recorded and the user might be able to actively annotate situations via speech. Action-accompanying speech might have many correlates in the physical environment. In the mechanic example the record of his or her activity can be correlated with the activities in the rest of the plant.

2.3.6.2 Access and Privacy

Recording oral communication is usually bound by legal rules that require explicit consent which would be hard to achieve in a scenario where recording is done in a continuous way: Consent can be elicited easily if the participants can see a clear benefit from keeping the records. Since the recording is indiscriminate this goal can't be claimed. Recorded action-accompanying speech is only recording the user and may not be as problematic, especially when used in a professional setup where the actions and therefore conversations are part of the professional activity.

2.3.6.3 Reinterpretation

Personal experiences are currently relived by telling stories and showing videos or photographs. However there are quite a few differences to continuous recordings: Storytelling is idealizing a situation and interpreting it, videos and photographs feature very specific angles as well. One has to find the proper perspective to shoot a good photo and it takes an effort to produce a good video. Reinterpretation could be the compilation of a video in conjunction with a story that is being told or the selection of photos with comments. Given the early stages of this project and the open questions on privacy it seems premature to discuss potentials for reinterpretation. It is also unclear how to make recordings from body mounted cameras that are pleasurable for later viewing and reliving the experience. Reinterpretation of action-accompanying speech is a very unclear concept – it is usually lost without this recording technique, therefore no traditional technique achieves a similar result which could be used as a blueprint. Since the goal is to document the actions that are taken it is also unclear how these actions would be reinterpreted: However they could be reconstructed and related to other events in the environment in the following applications:

forensic analysis in case of professional misconduct (police, plant operators and workers, health care professionals)

chain of events analysis analysis of a larger event which leads to catastrophic results (chemical or nuclear plants)

production optimization reconstruction and analysis of the day-to-day operations in professional environments (plants, mobile workers, etc.). Activities that are carried out manually have a potential for optimization if they occur frequently enough and recordings might be reveal those easily.

2.4 Related Work

2.4.1 Introduction

Many disciplines including linguistics, anthropology and sociology wondered about the nature of human discourse and the amount of work is confusing [9]. However there are not too many large scale empirical studies of human discourse such that this work makes a significant contribution out of its own right. This thesis also has the advantage that it presents an independent evaluation criterion for the analysis that is being performed: It should be useful for information access. This work is not only related to discourse theory, it is also related to general information retrieval since that is the task which is being performed, it is related to speech recognition, speaker identification etc. since these technologies are applicable, it is related to machine vision since rejoinders may also be indexed using vision based human id or the visible focus of attention of the individuals.

The section will therefore introduce work which is related to the the consideration of discourse classification (Sec. 2.4.2.3), including activity, microlevel features, topic and generic progression. Sec. 2.4.3 introduces special dialogue theories: Dialogism (Sec. 2.4.3.1), systemic-functional grammar proposed by Halliday (Sec. 2.4.3.2), the theory of dialogue act and games (Sec. 2.4.3.3), Rhetorical Structure Theory (RST) and Gross and Sidner's theory (GST) (Sec. 2.4.3.4). The psychology of long-term and autobiographic memory is also important since it reveals indices of conversations people might remember (Sec. 2.4.4). Speech recognition is important to automatically transcribe meetings at low cost (Sec. 2.4.5) and is presented with dialogue act and emotion detection from speech. Keyword based information retrieval as well as summarization are discussed in Sec. 2.4.6).

2.4.2 Discourse Categorization

The discussion of dialogue categorization is very difficult given the literature, already on the terminology level. This subsection will therefore clarify some of the terminology and present the point of view taken here.

General dialogue typology or taxonomy (Sec. 2.4.2.1) is an endevour to categorize all possible dialogues into a set of classes. While there is no accepted

[9] Even building a taxonomy of work that was carried out in that field would be daring attempt and should not be tried here. Rather than presenting all the work in that field and weighing all the pros and cons a few authors presenting significant advances that lead to the applications provided in this thesis will be discussed. Even linguistic literature like Linell (1994) points back to other literature like Schiffrin (1994) which itself is not more than an attempt to summarize the achievement so far, the authors decision is therefore in good company with linguistic experts of the field. A fairly good tutorial introduction to the seminal work in discourse analysis is Slembrouck (2001).

general solution the discussion is important to understand. Sec. 2.4.2.2 describes one partial solution where the constraints on the way people interact and how it their interaction has to be taken is called activities. Activities are a central term of this thesis and investigated further in Sec. 4.

One option, instead of defining high level style, is to look at micro-level features, e.g. word and part-of-speech distributions as well as dialogue acts. The possible choices are limited at that level and one might assume that many of these choices are correlated to each other (Sec. 2.4.2.3).

Since dialogues can be very long it is important to understand that the style of the conversation is changing which may support information access. Sec. 2.4.2.4 describes the notion of topic and the discussion is later continued in conjunction with algorithmic work in Sec. 5. This thesis assumes that the activity typically stays constant in a topical segment.

A genre could be described as an activity however it has a richer set of properties associated, the key property is *generic progression* (Sec. 2.4.2.5): Separate identifiable stages are taken which serve different functions in establishing the overall goal of the activity. The notion of generic progression would enable to find information by just knowing where the parts of the generic progression are located. However generic constraints are more likely to apply to formal settings and haven't been observed in our databases.

2.4.2.1 Discourse Typology

Which labels can we assign to a dialogue? The author fully agrees with Fritz and Hundschnur (1994) who describe an array of approaches to discourse typology just to conclude that there is really no conclusive structure. However, he also concludes, that sub-taxonomies for specific applications might very well be understandable. Fritz and Hundschnur (1994) isolated the following problem areas for a general taxonomy:

Taxonomy There is no agreed taxonomy in the literature and Fritz and Hundschnur (1994) suggest a minimal hierarchy: task-oriented, relationship-building and action-accompanying dialogue. There are two types of taxonomies used here, one is based on the notion of a "database" of very different speech data types while the other is based on the "activity" of the speakers (Sec. 2.4.2.2). However there might be other options for the top level of the taxonomy such as the government/business/private/other interactions, the number of participants, the place the rejoinder takes places etc. Sec. 2.4.3.1 shows that genres are most sensitive to language change and are therefore a very productive linguistic entity: They are constantly in flux and culture dependent such that the existence of a general taxonomy is unlikely.

This property makes a machine learning approach specifically interesting since it can be retrained easily.

Sub-taxonomies and domain specifity It seems to be a lot easier to build taxonomies for a specific area such as "discourse in sales", however generalizing the dialogue types found might be hard. This problem might be coined domain specifity and the *planning* activity (deciding on a future plan to be agreed upon and carried out between the discourse participants) may be taken as an example: Two soldiers in combat reduce the conversation to the shortest and most precise language, likely using specific constructs for this situation. On the other hand planning a night out can be very wordy and carry many discussions about unrelated events while at the same time the actual plan is barely mentioned on the surface.

Dominant function It is not clear how to treat a complex dialogue that is composed out of sub-dialogues which by themselves belong to different types in a taxonomy.

Dialogue acts Many authors proposed that dialogue acts and sequences of those are strongly related to the type of the dialogue, however the relationship is not fully explored or explained at this point. These dialogue act sequences can be be interpreted as the dialogue games in this work and are used for activity detection (Sec. 4, Franke (1990)).

Lexical information Action verbs (such as planning and storytelling) and other lexical items were used as a basis for dialogue classification but the validation of this approach is still pending according to Fritz and Hundschnur (1994). This thesis supplies empirical evidence to the use of action verbs as labels for discourse types (activities, Sec. 2.4.2.2).

Granularity The number of dialogue types can be arbitrarily high, in fact even TV shows ("a CNN interview") create commonly used dialogue types which may actually exhibit different constraints on the participants.

General taxonomies or classifications are therefore likely only of limited use. There are more options to escape this dilemma, all of them refer to lower level features:

- define the typology by "clustering" along "main dimensions" of microlevel features such as words and parts of speech (Sec. 2.4.2.3). These main dimensions might be found using principle component analysis (Biber, 1988), however for a critical discussion see Lee (2002).

- use activities since there is only a limited number of ways people can actively use language (Sec. 4, especially Sec. 4.5)

- use databases of dialogues as reference points to define how similar the style is to a certain database (Sec. 4, especially Sec. 4.3).

- use micro features directly

- supply multiple independent characterizations (Bakthin's characterization of dialogue (Sec. 2.4.3.1) and systemic-functional grammar (Sec. 2.4.3.2)). This approach can be applied independently and the intelligent meeting room as well as the user study exploit it.

The approach based on clustering is problematic since it is ad-hoc and uses variables that might not appeal to a lay user. The approach using databases as landmarks is very intuitive and simple, however it only applies to large variations in discourse style. It may be seen as a supervised version of the clustering approach. The approach based on activities is much more fine-grained however even the agreement between human coders on activity classification is fairly low. Only some micro level features lend themselves to direct inspection, examples are formality and dominance/initiative of speakers which are investigated in the user study (Sec. 6.3).

2.4.2.2 Activity

Levinson (1979) describes the fact that choices of the discourse participants are restricted by term activity (the *structural* aspect of the activity). In that sense there might be a limited number of possible actions speakers might take which might in turn lead to a limited set of dialogue types. Indeed this work uses activities and the terms to describe them are based on action verbs such as discussing, planning and so forth. The core of his article is (Levinson, 1979, p. 393)

> [...] that types of activities, social episodes if one prefers, play a central role in language usage. They do so in two ways especially. On the one hand they constrain what will count as an allowable contribution to each activity, and on the other hand they help to determine how what one says will be taken [...]

This may be seen in examples such as interrogations where one of the participants – the subject being interrogated – may not be cooperative and therefore purposely violates the Gricean's maximes of being informative. Therefore activity analysis shows that the situated understanding of language is crucial for many types of language. One may assume that interrogation have very typical dialogue act patterns (long statements and questions by one person, short yes/no answers or only backchannels by the other) such that it is evident that the detection of this meta-variable is more natural at the dialogue act level than at any lower level. Our tagging scheme for CallHome Spanish (Sec. 4.5.2, Ries et al. (2000); Thymé-Gobbel et al. (2001) was fairly simple and assigned only one simple major category to each segment, specifically *storytelling*, *advising*, *planning*, *discussion*, *consoling*, *interrogation*, *recording* and *undetermined* have been used [10]. Additionally to these categories our analysis of the CallHome Spanish database suggested to look at gossip which was also investigated by Eggins and Slade (1998). According to their work gossip is mostly characterized as a third-person oriented activity which wasn't supported by how we naturally looked at the data: People effectively gossip about their own lives as well. The function of gossip to form a social group by establishing a common believe system however correlates with our intuition of gossip and it seemed a reasonable definition for CallHome Spanish as well. Instead we attached a positive/negative/divergent/neutral evaluation annotation to each storytelling segment and identified the main person/object of the conversation. Gossip, as defined by Eggins and Slade (1998), would therefore corresponds approximately to a *storytelling* segment with a *negative* evaluation about a *third person*. Results on the automatic detection, manual annotation and activity annotation in the on the meeting and Santa Barbara corpus are obtained in Sec. 4.

Conversation analysis (CA) (see Slembrouck (2001) for a short introduction) is a very phenomena-oriented research direction in Sociology investigating face-to-face encounters. CA attempts to uncover how social order is brought into the conversational context. It uses speaking turns in context and especially the dialogue act sequences to describe higher level structure such as activities. It is therefore very similar to our approach to use dialogue acts and games (Sec. 2.4.3.3)) to detect activities. Franke (1990) also presents a large number of dialogue act and game sequences and analyzes them.

Finally, Allwood (1995) and Allwood and Hagman (1994) present an activity-based approach to pragmatics that attempts to be more inclusive than previous approaches. The participants in an activity are rational agents that can be understood and/or explained by trying to make relevant contribution to the current activity context. The activity context is subdivided into "constraints and enablements" according to Allwood (1995),p.15: Communicators are *physical* and *biological* entities, they are perceptual, understanding, motivated and emotional beings (*psychological enablement*) and they are part of *social* and cultural groups.

[10] The *recording* category is used for artifacts that are introduced due to the recording conditions such as a discussion about the fact that they are being recorded.

However, in the investigation of concrete empirical data he only uses a couple of simple features and compares them across fairly different high level dialogue types (Allwood and Hagman, 1994). We have also found that high level dialogue types vary dramatically along very simple features (Sec. 4, especially Sec. 4.3 and Sec. 4.4). Their work however does not contribute to the interesting question whether one can distinguish more similar categories such as activities within meetings (Sec. 4.5.3). Martinovski (1996) from the same school presents a similar setup however including more features such as dialogue act distributions and turn overlap features. However the same critique applies – the activities of interview and discussion are so distinct that they can be easily told apart on all of these levels. In later work Martinovsky (2000) describes how dialogues differ between Hungarian and Swedish courts including the use of a new feature type, namely the repetition of words and/or larger syntactic constructs. Repeating material from another speaker can be seen as a power device since it introduces or constitutes a reinterpretation of preceding material and seizes the right to do so. It seems that the dialogue act distributions (especially the number of question) can be interpreted by a lay analyst in this situation. Linell et al. (1988) describes different activity types by the distribution of dominance in the discourse. He uses dialogue types and associates dominance and activity contributions to each of those: This dominance feature is relatively simple and intuitive and will therefore be used in Sec. 4 and Sec. 6.

2.4.2.3 Microlevel Features

Classic microlevel features can be formulated as distribution of word classes, syntactic constructions, measures of vocabulary size as well as relationships between those features such as ratios. They are applied in a couple of setups and they are also used in this thesis to characterize higher level distinctions of dialogue. Indeed these microlevel features as well as most others will be used in an implemented form in the database and activity classification chapter (Sec. 4, especially Sec. 4.2).

Holmes (1998) presents an overview of “stylometric” scholarship which has been applied to the attribution of authorship if it was disputed based on historic or other evidence. The measurement of microlevel features has also been advocated as an analysis method for register studies in written (and sometimes spoken texts): Tannen (1984) and Quirk et al. (1985) try to identify to what extent grammatical and other choices vary along “register” or within a discourse typology. This idea was used in corpus based linguistic work (Biber, 1993; Biber et al., 1998) and in information retrieval (Kessler et al., 1997; van Bretan et al., 1998) recently. Another interesting similar aspect and application is presented by Kaufer; Kaufer and Butler (2000) who presents techniques for teaching writing using “representational compositions" [11]. To compose such representations students need to display certain types of phrases, for example in order to involve a reader. To that end he has build a large English phraseology and a text visualization system which helps students to actively and objectively explore basic writing principles without the consultation of a teacher.

Another way to look at the microfeatures is to make use of "semantic fields". McTavish et al. (1995) uses semantic word categories to determine stylistic features in conversations and assumes that they are related to social distance. A similar semantic classification of nouns and verbs in English can be derived from WordNet's lexicographers classes (Fellbaum, 1998).

Among the stylistic features there is definitely one which stands out and it might be called *formality*. As Heylighen and Dewaele (1999) point out, when applying principle component analysis to stylistic features across languages (see also Biber (1993); Biber et al. (1998)), the first dimension seems to always be interpretable as formality. His definition of "deep" formality (as opposed to "surface" formality which is "attention to form for the sake of convention or form itself") is "attention to form for the sake of uniequivocal understanding of precise meaning of the expression". His definition therefore continues to describe formal language as largely context-independent and non-fuzzy. Formality is reflected in many facets, however it can certainly be measured in the part-of-speech distribution distinguishing word types that are "anchored in spatio-temporal context in order to be meaningful" (*deixis*) from those that are not. A simple formality score is defined by counting parts of speech

$$\frac{\text{nouns} + \text{prepositions} + \text{articles} - \text{pronouns} - \text{verbs} - \text{adverbs} - \text{interjections}}{\text{all words}}$$

He proves his point by plotting formality scores for a number of databases and they seem to rank databases as expected.

Additionally there is a set of features which is only relevant for dialogues: Dialogue acts (and games) are certainly a very powerful feature since they represent the actions speakers are taking in a dialogue (Sec. 2.4.3.3). Among others they encode dominance (Linell, 1990; Linell et al., 1988) and they may also be assumed to be highly correlated with activities. Other simple dialogue measure are "interactional" features such as speaker overlap and turn lengths [12]. Features

[11] Writing education is standard in the American undergraduate curriculum while it is not in Germany. It typically involves the training of writing in a couple of genre.

[12] Example applications of these are Allwood and Hagman (1994) who describe a large number of simple features but the study does only relate the features to general databases where all features vary significantly (see also Sec. 2.4.2.4). Linell (1990); Linell et al. (1988) describes the use of a dominance feature to characterize different types of lay/professional interactions and Martinovsky (2000) compares the style of court proceedings in Sweden and Bulgaria using dialogue features.

related to prosody may be derived from the audio channel such as pitch, volume and pause and are used for emotion detection (Sec. 3.8) and topic segmentation (Sec. 5.6.6) and one might add features like emphasize and lexical stress.

2.4.2.4 Topic

The definition of topic in linguistics is all but consistent. A recent literature review (Goutsos, 1997) puts special emphasis on the fact that topic can often be more reliably defined linguistically by not referring to the coherence of the propositional extension of a segment. Halliday and Hasan (1976) presents the classical reference that defines “texture” as a property of real texts as opposed to arbitrary collections of sentences. A *text* is defined by *cohesion* (consistency within the text) as well as consistency with the context of the situation or *register*. Cohesion can be *grammatical* (reference, substitution, and ellipsis), indicated by *conjunction* or *lexical* (repetition as the simplest phenomenon).

There is another trend in empirical linguistics to deal with the definition of topic which is to ask human subjects to code it on a corpus and to measure the agreement. Given different definitions of topic and their agreement results one might claim to have a better or worse definition. It seems however that even the simplest definition appealing to a lay persons intuition seems to work reasonably well.

Another test is to construct algorithms and see how well they are doing with different sets of features and segmentation criteria. Sec. 5 introduces an effective probabilistic algorithm and successfully uses the following features:

keyword repetition can be related to *lexical cohesion*

speaker initiative is the identity of the speaker for each utterance, typically indexed with the information whether the turn was long or short. A topic is therefore *cohesive* with regard to its speaking rights.

word/part of speech distribution are different in the beginning, middle and end of a topic, following standard genre conventions. It therefore follows genre convention and is *cohesive*.

2.4.2.5 Generic Progression

In early work conversations and narratives were characterized by their generic progression or stages (Eggins and Slade, 1998; Gee, 1986; Labov and Waletzky, 1967; Levinson, 1979; Longacre, 1996; Plum, 1988; Tannen, 1993). A regular expression for narratives, the best researched genre, was proposed as (Eggins and Slade, 1998; Labov and Waletzky, 1967):

```
Abstract? Orientation? Complication
    Evaluation Resolution Coda?
```

Different genres are defined as belonging to different patterns, typically expressed in a regular expression syntax. One may assume that this could be taken as a basis for distinguishing activities (Eggins and Slade, 1998). We initially tried to do that as well when we started the annotation of CallHome Spanish but we found their definitions or similar ones hard to apply:

- the different types of storytelling are only distinguishable by looking at very fine distinctions at turning points in the conversation. Inspecting our data we found it difficult to make those fine distinctions by hand and we found the resulting labeling counterintuitive.

- the finite state descriptions as presented in Eggins and Slade (1998) were fairly complicated but would need even more specific information when applied to new databases. The complexity needed for the finite state representation therefore renders it unintuitive unless a more coarse grained classification is adopted.

Since we did not have good intuitions about a specific generic progression within more detailed activities and the intercoder agreement between simplified activities is only moderate this seems to be a prudent decision (see also Sec. 2.4.2.2 and Sec. 4), especially in hindsight.

Nevertheless genre progression may be exploited in many information retrieval and automated summarization scenarios. Generic progression provides information where different types of information can be found and how they have to be interpreted. Newswire texts commonly follows a pyramid style, introducing the core facts in the first couple of sentences and adding more information as the article progresses. Since meetings are often used to strike or enforce social contracts the conclusions have to be presented in a clear and unambiguous form, one might find presentations, action items, conclusions and the scheduling of follow up meetings (Bargiela-Chiappini (1997),see also Sec. 2.3.2). There might also be specific cue phrases for action items and summaries leading into them (“action items”, “so what do we have to do”, “where do we go from here”) or concluding them (“let’s do that”, “that’s the plan”). The cue phrases may also be used as a user interface to the speech system as a “voice button” and be part of the documentation of the system. Our meeting databases however consisted of fairly informal meetings such that these features were not (naturally) present.

Even if detailed structure within a topical segment can’t be obtained automatically there might be a general structure that always exists: A topic is usually introduced (initial segment), it is then explored (middle segment) and it is completed or the transition to the next segment is negotiated (final segment). A summary of a typical segment could therefore consist of just the initial segments of each topic and some salient parts of the middle segment. The initial segment has some interesting properties that make it a good candidate for a summary:

- it is self consistent.
- all entities are introduced (low number of anaphora such as pronouns).
- the issue is introduced which is indicative of the rest of the segment.
- it is just the initial segment of the topic, a user can access more by just continuing to play or browse it.

An application of this standard generic progression is the segmentation of dialogue using the difference of parts of speech distribution in beginning, middle and end of topical segments (Sec. 5.5).

Orlikowski and Yates (1994) present another application of genre, however to email lists: They discuss indicators for audience (addressed to individuals, email lists, . . .), purpose (mostly in the subject line: `FYI`, `RE:`), structure (content contains lists, LISP code) and language (formal, informal, emotions). Additionally structure specific to the discussion forum was taken into account which might be a good idea for other applications as well. As discussed in Sec. 2.4.3.1, genre is one of the most flexible entities of language and there should be no surprise that with the invention of new cyber-media options new cyber-genre evolve (Breure, 2001). Similarly professional/lay interactions, e.g. doctor patient interviews, have been studied in great detail, foremost by Linell (1994). The relationship between both parties is usually well defined and in many situations the professional is engaging in a form filling task: The professional is gathering all the information necessary in their categories to make a determination. Modern genre theory however makes broader claims on what a genre is (shortened after Breure (2001)):

Pattern of communication A genre structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation. This is captured in microlevel features as well as generic progression.

Situatedness Genre is a type of communicative action, associated with a situation. As an institutionalized response to a recurrent situation, it reflects the norms, ideology and habits of the discourse community concerned with respect to such circumstances (see also Sec. 2.4.3.1).

Dynamism Genre is a relatively stable phenomenon subject to evolution (Sec. 2.4.3.1, Sec. 2.4.2.1).

Content, form, and function Not only content and form but also purpose and function are important (Sec. 2.4.3.2 and the preceding discussion of Orlikowski and Yates (1994)).

2.4.3 Selected Linguistic Theories

2.4.3.1 Dialogism

What makes a dialogue? And is dialogue fundamentally different from written language? Key properties of spoken interactions are (Linell, 1994, p. 8f):

> [...] *sequential organization*, *joint (social-interactional) construction*, and *interdependence between acts* (local units) and *activities* (global units and abstract types).

Given these properties this thesis explores features of dialogues other than topic which has been dominant in written language retrieval. A similarity to Clark (1996)'s metaphor of a musical concert which requires the "joint construction" to create a piece successfully is apparent. Distancing himself from the "monologistical" tradition (Linell, 1994, p.22) does not analyzes a dialogue in the "transfer-and-exchange model of communication" where

> The utterances and their meanings are explained by recourse to the speakers communicative intentions, and the listeners task is described as recovering these intentions [...]

In order to understand a dialogue in the "monologistic" tradition an overhearer, for example a computer, has to follow the same reconstruction process: The intentions of the speakers as well as their achievements in the grounding process have to be reconstructed. The terminology suggests that the "monological" tradition analyses speaker/listener interaction as a series of monologues rather than a dialogue. The hypothesis that dialogue is a joint (social-interactional) construction would allow to make the additional hypothesis that this construction can at least be observed partially on the surface [13]. The concept that activities are interdependent with the local dialogue acts is also important to notice since it suggests that the detection of the dialogue act would enable the detection of the activity. The dialogical view on the other hand suggests to analyze the local contributions and their effects in a holistic fashion which is the approach taken in this work. Particularly useful are the dominance relationships which were already proposed

[13] The speakers might be using codes that are not understandable for an overhearer (Clark, 1996) that wasn't part of the conversation since expressions and references that are being used might not be understood (Fig. 2.2).

by Linell et al. (1988). The dialogistic approach therefore suggests that some understanding is possible even without deep semantic reconstruction – this thesis therefore depends deeply on this hypothesis being true and the success supports this underlying hypothesis. Shallow understanding is necessary for current systems since the speech recognition is fairly inaccurate and the semantic analysis of free spontaneous speech is still an unsolved problem.

"In recent years the *term* 'dialogism' has become closely associated, if not identified with the work of the Russian literary scholar Mikhail Bakhtin" (Linell, 1994, p. 49). Bahktin (1986) analyzes the characterizing elements of a dialogue as the topic, situation and style – at least on a high level the analysis of a dialogue can be achieved without the metaphors of humans as complex information processing devices but rather using the metaphor of social beings.

An important aspect of Bahktin (1986) is that societal change seems to first occur at the level of activities: What activities are exercised and how may change rapidly [14]:

> Utterances and their types, that is speech genre, are the drive belts from the history of of society to the history of language.

This observation – the plasticity of genres – should indicate that genre definitions have to remain very general in order to be successful. Indeed this is one of the observations of the experimental Sec. 4. Since genre are a highly variable and productive linguistic entity the ability to learn their properties is very important in order to build practical systems as they need to be adapted over time and to different setups.

Another important related contribution of Bakhtin is the key notion of heteroglossarity: People switch codes or language in order to make a point (e.g. by associating themselves to a certain social group), they cite other people in their code and may even be unable to rephrase professional opinions without recourse to the original. Since the style of the interaction is a feature that is paid specific attention to in this work it would seem natural to investigate this question as well. However style is multifunctional (for example for stressing passages, emphasizing the character of the activity, speaker identity) such that teasing apart these contributions seems to be a task that is too challenging for automated techniques at this point.

Linell (1982, 1994) strongest point might be that traditional Western linguistic has a written language bias (WLB). This bias is specifically visible by distinction between an ideal internalized *langue* and its realizations (*parole*) which allows

[14] Hanks (1988) offers an analysis of a number of written texts produced by native officials in early colonial Maya society (Mexico) as blending Maya and Spanish discourse forms into new genres (Slembrouck, 2001).

to explain many oral phenomena as failures to produce "correct" sentences. As pointed out in Sec. 1.4.1 speech genre (Bahktin, 1986) and activities are highly interdependent with the sentences such that an analysis that ignores them is invalid in their opinion.

In part WLB might be explained by the lack of technical support to analyze spoken language since tape recorders, digital recording and digital speech processing are fairly recent inventions. Automated analysis of spoken conversations, especially on a high level as in this thesis, may be seen as a lackmus test for any theory of language. This thesis can't claim any final results on the matter but there are some indications that activities play a significant role for the indexing of spoken rejoinders which may tilt the evidence to a dialogical interpretation of language.

2.4.3.2 Systemic-Functional Grammar

Halliday has started a major attempt to relate high level categories of the context of a situation to properties in the grammatical system. For the purpose of his analysis he relates separate function of the context of a speaking situation (*register*) with semantics and ultimatively with lexicogrammatical features. The register functions he describes are (Halliday and Hasan, 1976, reformatted from p.22):

field the total event in which the text is happening, together with the purposive activity of the speaker or writer; it includes the subject matter as one element in it.

tenor the function of the text in the event, including therefore both the channel taken by the language – spoken or written, extempore or prepared – and its genre, or rhetorical mode [...]

mode the type of role interaction, the set of relevant social relationships, permanent and temporary.

These functions are implemented by a semantic system which constructs meanings: Ideational meanings realize *field*, interpersonal meanings realize *tenor* and textual meanings realize *mode*. Halliday (1994) describes how these different meanings are constructed by different interpretations of the clause. Probably the most convincing and famous example is the interpretation of the subject which is seems to have a significant impact on his work. Halliday (1994) claims that there are three different types of subjects, psychological (*theme*), grammatical (*subject*) and logical (*actor*). In the sentence

This teapot my aunt was given from the duke.

this teapot is the theme, *my aunt* is the subject and *the duke* is the actor. In many cases however the different types of subject definitions coincide, for example all three coincide here (*the duke*):

The duke gave my aunt this teapot.

Indeed, if we would have had a systemic-functional parser for English which works reliable on meeting data we might have been able to exploit some of the elements which are realized in grammatical relations. So far the computational part of the field wasn't as active in parsing as in generation. Additionally it wasn't absolutely clear whether there were enough features available that a functional parser would uncover which are not available or represented otherwise. Eggins and Slade (1998), from the same department at Sidney University as Halliday, seemingly didn't use a functional parser in their analysis of spoken dialogues such that there are likely some significant barriers.

The author is also not convinced of the clear separation of the proposed register functions. They seem to be highly correlated, for example *field* describes how people interact whereas the *mode* describes their social relationships – social relationships however are often determining how people interact as elaborated in the last section. Similar it is hard to separate the type of event (for example a lecture) from the way language is being used in it such that these indices are not orthogonal. It seems that the *field* function contains most of the interesting aspects and needs to be teased apart.

2.4.3.3 Dialogue Act and Game Theory

Dialogue acts [15] were considered building blocks for spoken language at least since Austin (1962) and Searle (1976), often called "New Rethoric". A dialogue act is an abstraction of the action a speaker takes by uttering it – it should therefore be indicative of the style of the interaction and different situations should also place constraints on the actions the participants may take. The study of dialogue acts pioneered the field since it opened the understanding for language as action. Speaking an utterance is a locution, doing something with it is an illocution and having achieved something is a perlocution. If I speak (illocution) and make a promise (illocution), I assume that my partner believes me (perlocution). Searle (1976) points out that dialogue acts can be indirect such as rhetorical questions. Although dialogue act theory – especially if coupled with universalistic claims – has drawn numerous criticisms (Slembrouck, 2001) its seminal contribution has survived. There are numerous examples of dialogue act annotation schemes that followed the original proposal.

[15] Dialogue acts are also referred to as speech acts by various authors.

communicative status
describes whether an utterance is intelligible and complete (usually unmarked, examples are self-talk, uninterpretable or abandoned utterances).
information level
provides an abstract description of the content of the utterance (doing a task, talking about the task, maintaining the communication, other talk). Most of the utterances in a non-task oriented dialogue would be maintaining the communication and other talk (smalltalk).
forward looking function
constrains the future believes of the participants and influences the dialogue. Examples are statements such as assertions or opinions, influencing the addressee (opening options, giving action directives), commitments and conventionals such as opening or closing
backward looking function
relates the current dialogue act to past dialogue acts (signalling agreement/disagreement, answers, signalling understanding).

Table 2.2: **DAMSL Dialogue Act Dimensions:** The backward looking functions were significantly extended in Jurafsky and Shriberg (1997) to capture variations of backchannels, commitments are expanded in our CallHome Spanish annotation Levin et al. (1999); Thymé-Gobbel et al. (2001) in the control-act category.

A recent attempt to define dialogue act types is (Allen and Core, 1997) — often referred to as dialogue markup in several layers (DAMSL) — introduces a common language in talking about dialogue acts. DAMSL was adopted by a number of discourse researchers loosely organized in the discourse resource initiative (DRI). This annotation scheme was developed for task oriented dialogues and may therefore need extension to capture multi-party discourse as well as non-task oriented dialogues. Allen and Core (1997) feature a number of independent categories that were annotated for each dialogue act (Fig. 2.2). Their scheme was extended for spontaneous non-task oriented speech in Jurafsky and Shriberg (1997) who annotated the largest database of dialogue acts to date (> 1000 conversations on Switchboard). In our project Clarity the coding scheme was extended to capture specifics of spoken Spanish such as attention directives (*Mire*, *y fíjate*) and control acts such as *gives promise or commitment for action*, *speaker asks for permission for action* (Lampert and Ervin-Tripp, 1993) and is explained in full detail in Levin et al. (1999); Thymé-Gobbel et al. (2001). We felt that capturing control acts was important since it would be allow to break down the *statement* category which was so dominant in Jurafsky and Shriberg (1997). Attention directives seemed important to distinguish since they are clearly marked and frequent in Spanish.

The importance of dialogue acts and games has already been discussed above: They are indicative of the activity as a whole and indicate dominance. Dialogue acts have the advantage that the may be detectable with low quality LVCSR or even with simple prosodic features such as the fundamental frequency and volume (Shriberg et al., 1998; Stolcke et al., 2000).

2.4.3.4 Rhetorical Structure Theory and Gross and Sidner's Theory

Both rhetorical structure theory (RST) (Mann and Thomson, 1988) as well as Gross and Sidner's theory (GS) (Grosz and Sidner, 1986) are often referred to as discourse theories. Given that claim one might think that they could contribute directly to the goal of this thesis. These theories were applied with great success in text generation. The most notable application of these theories to unrestricted (text) understanding is Marcu (1997) who showed that rhetorical relations for RST can be found automatically in written texts (mostly some articles of "Scientific American" that he considered well written) and was able to use those for summarization purposes. He used mostly cue-phrases and their position to find the segment relations and a rhetorical parser that would construct valid interpretations. Despite that new ground that was broken by Marcu (1997) applying this general approach to unrestricted discourse presents a number of open questions:

basic unit of analysis The basic unit could be breath groups, turns (the thing between silences), dialogue acts, dialogue games or topical segments.

status of cue phrases The analysis of cue phrases in Marcu (1997) – key to the success of the work – may only be valid for written language. In an informal analysis using the phrases of Marcu (1997) their status on Switchboard was not immediate clear.

tree structure Dialogues commonly exhibit behaviors that are hard to explain in a standard tree-structured RST (Carberry et al., 1993). Indirect responses, abundant in many of our dialogues, would definitely pose a problem to RST since the textual relations are only present on the semantic level. Discontinuous spans – e.g. returning to earlier points, presenting an example that relates to multiple spans – are also fairly frequent in chatting and can't be handled in a tree based model, at least not directly. Even worse there are examples of parallel narratives about different incidents by the discourse participants that don't interact as well as collaborative narratives where each speaker is presenting and relating their views on a topic. The restrictions of dialogue models that have tree structure are therefore too narrow for this application.

importance Given the complexity of the analysis it is unclear how and which kinds of structures can be extracted reliably. If the features are too unreliable or too simple it is rather unlikely that they would be important to include into the feature pool for tasks such as topic segmentation and activity detection.

local effects and lexicalization It is not clear to what extent certain dialogue patterns are completely lexicalized and would need to be handled as such. Simple example of patterns are backchannels that signal understanding in an "information giving" dialogue game. This observation would suggest to use dialogue games as the basic analysis unit.

While GS provides more insight on the informational level in this context it would require at least limited semantic understanding for unrestricted human-human dialogue which is not a capability that will likely be available soon. Instead of extending these theories to handle unrestricted discourse this thesis takes a more direct approach and focuses on surface correlates of discourse events. Those correlates may be explained differently by different theories of discourse but they may be good enough for the applications we want to pursue: Precisely this focus on shallow analysis is what makes this work different from other attempts to understand dialogue as a whole. Nevertheless there is a possibility that summarization techniques based on Marcu (1997) could be used to support information access of spoken language.

2.4.4 Autobiographic Memory

Many applications for rejoinder access require access by the participants. It is therefore important to understand what people might remember of the rejoinders they are looking for since those are the potential indices. The study of personal and autobiographic memory is performed in psychology. Brewer (1993) takes an instructive general tour of autobiographic memory research and contrasts it with survey research (research on the certainty of survey data). Brewer (1988, 1993) suggest that autobiographic memory contains information about location, people, and actions but very limited information about absolute time. (Brewer, 1993, p.15) features the following illustrative anecdote:

> [...] I may recall having a conversation at Alberton House with Norbert Schwarz. I recall that we were standing in a doorway and that we talked about the German academic system. However, if I try

> to recall what the time was, I can only generate an answer by recalling that it was dark and that we were near a table with drinks; therefore it must have been early evening.

Much of the research in autobiographic memory was devoted to the question whether the memory is just a "copy" of the original (copy theories) or whether it is recreated from some underlying representation which allows to reconstruct the event (strong reconstructive theory). One should expect specific types of reconstruction errors depending on what the underlying representation is. Brewer (1988) shows that many recalls are surprisingly accurate such that he rejects a strong reconstruction theory. The copy theory is vulnerable since many memory reconstruction errors are apparent in everyday life and are usually supported by anecdotes. On the other hand humans can recall whole situations at once as in the example of "flashbulb-memory" when a situation seems to be preserved almost as if a snapshot was taken. Sometimes these events may also exist across a larger population as in the case of the entry of World War Two by the Americans and can be surprisingly accurate.

The situation is different however if we consider the retrieval of specific types of situations (scenes) for which stereotypes or "scripts" are available (Anderson and Conway, 1997; Schank, 1977, 1982). It seems that the memorization of the event is achieved by remembering the unusual parts of the event that deviate from the script. In an extension of that work the CYRUS system was constructed which imitates the autobiographic memory of former US secretary of state Cyrus Vance (Kolodner, 1983).

While it is not clear that knowing the general activity in which an event took place is speeding up the recall of the event, experimental evidence suggests that presenting previous steps in temporal order does (Anderson and Conway, 1997, p.234-238). On the other hand there is experimental evidence that autobiographic memory is organized by *lifetime periods* such that priming experiments showed that events can be constructed faster if the lifetime period is known in advance.

Radvansky and Zacks (1997) presents a body of research that suggests that situations and especially their spatio-temporal aspect are represented in long-term memory and back their claim up by memory response time studies. Additionally Keenan and Baillet (1982) suggests that what the subject feels about the event is crucial for the recall of the event. This is in line with much other research although it is not clear whether it is just the strength of the emotion involved that leads to shortened response times in experiments or whether it is important that the emotion was positive or negative (see also Herrmann (1993); Keenan et al. (1982)).

Given the complexity of long-term memory and autobiographic memory these studies have to be taken with a grain of salt – too many systems seem to be interacting and it is hard to separate all conditions sufficiently. This is also evident by the fact that autobiographic memory tends to be a mix of general and specific memory and that retrieval takes rather long. In summary there seem to be evidence to support that:

- people attending are usually remembered
- the location and circumstances are often remembered
- the topic seems to be remembered, rarely verbatim (Keenan et al., 1982)
- the type of event is often remembered
- the exact time of the event is often not remembered directly but has to be reconstructed from context
- emotion and involvement play an important role whether something is remembered or not

The only surprising element in this list is that the exact time of the event is often hard to recall directly. The consequence is that humans are able to recall both thematic, situational and stylistic aspects of the conversation.

2.4.5 Speech Recognition

2.4.5.1 Introduction

Speech recognition is related to this work in many ways: On one hand word based information needs to be detected using a speech recognizer or manually transcribed and we may have to deal with inaccurate machine transcripts. On the other hand the techniques, features and constraints explored in this thesis might be applied to speech recognition proper: If we know that a segment is more formal or we have a high likelihood for a certain dialogue act we can apply that knowledge to adapt the acoustic and language model. There are further issues in speech recognition that this short introduction can't cover such as the automatic segmentation and clustering of speakers (Renals and Robinson, 2000). The techniques covered are automatic speech recognition (Sec. 2.4.5.2), dialogue act detection (Sec. 2.4.5.3), speaker detection (Sec. 2.4.5.4) and emotion detection (Sec. 2.4.5.5).

2.4.5.2 Detection of Words and Keyword Based Retrieval

Large vocabulary continuous speech recognition (LVCSR) was investigated for many years, including work that was carried out with involvement of the author

Baseline System Word Error Rate on Different Tasks [%]	
BN (h4e98_1) F0-condition	9.6
BN (h4e98_1) all F-conditions	18.5
Newshour	20.8
Crossfire	25.6

Adaptation to Meeting Data	
ESST system	54.1
Baseline BN system	43.1
+ acoustic MAP Adaptation (10h meeting data)	40.4
+ language model interpolation (16 meetings)	38.7

Table 2.3: **Speech Recognition Results (Reproduced from Waibel et al. (2001a)):** The upper part evaluates the baseline BN system across different tasks. MAP (Maximum A Posterior) adaptation was used for domain adaptation. The language model was adapted by interpolating the BN model with a small meeting model. The ESST system (Waibel et al., 2000) were trained on clean speech in a travel planning domain and is significantly smaller than the BN system. The author did not directly contribute to the LVCSR results but to other parts of Waibel et al. (2001a).

or by colleagues at the Interactive Systems Labs at CMU (Finke et al., 1998; Zeppenfeld et al., 1997). Recently corpora that contain increasingly unrestricted speech like Switchboard (free discussions over the telephone between strangers about some given topic) and Broadcast News (broadcast radio and TV programs) have been used in the evaluations of LVCSR systems (DARPA HUB-4 and HUB-5 evaluations respectively). The use of meeting data was suggested for the next hub of LVCSR evaluations (DARPA HUB-6) and Waibel et al. (2001a); Yu et al. (2000) present results on meeting data which is currently featuring a word accuracy slightly below 40% using a state-of-the-art speech recognizer (Tab. 2.3). As Stark et al. (2000) suggests this error rate regime usually leads to the preference of the audio over the transcript when humans have to process it.

The best speech system that was used on a Broadcast corpus on the TREC-8 SDR task (Garofolo et al., 1999), a spoken language information retrieval evaluation, had a performance of 20.5% word error rate while running no more than 10 times as long as the acoustic signal. The error rate for human generated TV closed caption transcripts was estimated at 14.5%, the accuracy for radio transcripts was estimated at 7.5%. A rolling language model was used in some experiments which makes use of the fact that news shows are often about the same events that are in printmedia of the same time period. This language model improvement resulted in an approximately 1% absolute reduction in word error rate. One could also hope to expand the dictionary of the speech engine to find all the proper keywords. In fact Garofolo et al. (1999) shows that keyword based retrieval is not degrading a lot in this domain if the transcript is generated by a speech recognizer as compared to a human transcript, especially if similar written corpora are available that can be used for document expansion (Singhal and Pereira, 1999). Garofolo et al. (1999) presents results on English and it has to be expected that for other languages (e.g. Korean and Turkish) a dictionary of 60.000 words – a common number in LVCSR – will lead to a much larger number of unknown words in a collection of similar coverage. It was therefore proposed by Federico (2000); Ng et al. (2000); NG (2000); Wang (2000) to use ngrams of phonemes or syllables rather than words as the basic unit for speech recognition and information retrieval. The work in Garofolo et al. (1999) is therefore only of limited use for information access in a meeting scenario: Reasonably similar documents usually don't exist in written form and the genre is also very different. Specifically the topic of arbitrary meetings – since they are highly specific to the group of individuals attending – may exhibit a larger variety than broadcasts (Sec. 1.2, Sec. 2.3.3).

2.4.5.3 Detection of Dialogue Acts

Dialogue acts were studied by Stolcke et al. (2000) and Wright (1998) in order to reduce the word error rate of a speech recognition system. The idea is that contextual information should help to disambiguate between different dialogue acts which should also reduce the overall word error rate of the recognizer. This seems to be true for narrow domains and a narrow set of dialogue acts since that would reduce the word choice a lot. Wright (1998) was therefore successful in a small domain which also had strong constraints of words given the dialogue act. Stolcke et al. (2000) on the other hand was unsuccessful in reducing the overall word error rate not only because the domain was too general and therefore also the dialogue act set but also since the most frequent dialogue act covered about 80% of the words in the corpus: 80% of the baseline language model would therefore be trained using 80% of the data and the vast amount of the data received proper modeling to begin with. Stolcke et al. (2000) however proved that word accuracy reductions for minor categories might be achievable even on their task. A number of experiments were conducted with the goal to split the majority category in Switchboard, the *statement*. One idea was to define context dependent statement types (Jurafsky et al., 1997b), another one was the annotation of control acts (Levin et al., 1999).

Dialogue act detection from speech as a problem per se, on the other hand, is well covered in Stolcke et al. (2000) and they explains how n-best hypothesis of a speech recognizer can be used to eliminate the influence of speech recognition errors to a large extent and how to combine word based information with prosodic information. In Sec. 3 the topic of dialogue act detection is revisited and improved detection algorithms as well as the detection of dialogue games (short sequences of dialogue acts such as question/answer pairs) will be discussed. The effects of the speech recognition algorithm, however, is not studied further in Sec. 3 since it has been studied exhaustively in Stolcke et al. (2000) under the participation of the author.

2.4.5.4 Detection of Speakers

Speaker detection is an active field and is recently featured in evaluations with international attendance (Przybocki and Martin, 1999). It is used for authentification purposes (speaker verification), it can be used for surveillance to identify a person even if they are in a place or on a phone they don't commonly use. Roy and Malamud (1997) uses speaker detection on a large database to align utterances in the audio signal with a transcript that was available. Speaker information is used in this thesis to describe speaker initiative which can be used for topic segmentation (Sec. 5.5.2).

Delacourt and Wellekens (2000) is an example and review of automatic methods for clustering unknown speakers in discussions. In a meeting situation with a short 30 second enrollment Pan and Waibel (2000) were able to achieve a performance of 98.5%/95.61% correct speaker identity on one second segments with a close/distant talking microphone. Speaker detection is therefore a fairly reliable technique even if spectral information is available only – if additionally location information is available or each speaker carries a separate microphone it might become even more reliable.

Speaker detection however could also be used to improve speech recognition performance since speakers might have different speaking styles. As a matter of fact speakers can be detected within one meeting from the part of speech distribution which indicates that a speaker dependent modification of the language model according to part of speech distributions is desirable. Speaker adaptation for the acoustic model is already a common technique.

2.4.5.5 Emotion Detection

Emotions are a very interesting human expression that can be communicated using gesture, posture, verbal cues and prosody. Furthermore emotion has physiological correlates like blood pressure and heart rate. Emotions could be important in a meeting and rejoinder indexing system since we often remember the fact that something was highly emotional and we also may remember emotional events better (Sec. 2.4.4).

Polzin and Waibel (1998) and Polzin (1999) describe the detection of a wide range of human emotions using prosodic, acoustic and verbal features. The detection of emotions is treated in Sec. 3.8.

2.4.6 Information Retrieval

2.4.6.1 Introduction

Information retrieval is an empirical science and art of finding information. With that overarching definition this work may be seen as a contribution to information retrieval. Classical modern information classical information retrieval relies on keywords that encode the topic of the conversation (Baeza-Yates and Ribeiro-Neto, 1999). In contrast this thesis focuses on non-keyword based features which were also explored in written language retrieval (Kessler et al., 1997; van Bretan et al., 1998). However their approach wasn't adopted by main stream information retrieval systems since the improvements didn't seem to be large enough.

A keyword based information retrieval (IR) model would take a document, remove so-called stopwords (frequent words that are non-topical such as determiners and other closed classed words) and represent a document as the set of words contained in the document. The basic retrieval method converts a document into a vector of the frequency of each word in the vocabulary. The query is also represented as such a vector and the closest document vectors are taken as the result. There are many refinements to this basic methods: The distance measure between vectors can be defined in multiple ways (TFIDF weights for example), the query or the document may be "expanded" to contain related terms (especially synonyms), link-structure of WWW pages may be analyzed, different document types may weighted differently, multiple databases may be searched etc.

A major push behind information retrieval is the development of the world wide web and search engines for it. Consequently the efficient processing of very large databases, updating indices, assigning popularity scores to Web pages and finding the web pages themselves (spidering) are other major concerns.

2.4.6.2 Text Classification

Text classification is a long standing problem in information retrieval which has also been a benchmark for machine learning algorithms (Sebastiani, 2002; Yang, 1999; Yang and Liu, 1999). Commonly text classification has been seen as a topic classification problem based on keywords, common applications are the generation of topical hierarchies such as in OHSUMED or the generation of index terms which had been created manually before (Reuters). The most commonly used database in text classification research is the Reuters database [16]. Since this database can assign multiple topic tags to a single document the classifier has to be able to output multiple values which has to be accomodated by the typical binary classification and turns out to be critical for performance (Yang, 2001). The neural network based classification framework used in this work (without hidden units) was tested on the Reuters database in preexperiments and delivered results comparable to the results published.

Stylistic text classification, as proposed in this thesis, can be seen as a fundamentally different way of classifying documents and parts of documents. The features used in the text classification systems to date are keywords, in contrast to the stylistic features used in this work. The typical measure used to compare classical text classification systems are micro- and macro-averaged F-scores. These numbers especially make sense since the common databases and comparisions feature multiple possible output categories. However it is straightforward to see that accuracy rate ($\frac{\#correct}{\#total}$) of a classifier with a single 1-out-of-n output is the same quantity as the micro-averaged precision, recall and F-score. Another evaluation of topical text classification methods is to measure the entropy of the output categories. The result is that a maximal reduction of 4.2 bit in search space is possible on the the Reuters database [17] and a maximal reduction of 2.3 bit is possible using topic labels on the CallHome database (Tab. 6.6). The best reported micro-averaged F score on Reuters was was 0.86 (Yang and Liu, 1999) while the corresponding score on the CallHome database topic is 0.57 (Tab. 6.6).

2.4.6.3 Spoken Documents

The access to spoken language has been addressed by the information retrieval community. The problem that was identified as interesting was the retrieval of broadcasts, especially Broadcast News. In Sec. 2.3 this topic is discussed at length however it should be noted here that broadcasts are very different from rejoinders such as meetings since they are designed to be understood by a larger audience and they feature a limited general range of topics. Given their communicative goal they may share many properties with written language. The result obtained was that with little effort the existing IR systems could index the database from automatically created transcripts with the same accuracy as from manual transcripts. While this was surprising as it is the result can't be taken for granted to apply

[16] The database is currently available from David Lewis at `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

[17] The evaluation requires to combine the output labels into "combined" categories. This is necessary since the categories are highly corrolated. A leaving-one-out estimate was used on the Reuters 21578 database. All documents with with topic attribute set to "yes" were used which corresponds to the training and test set of the famous ModApte Split used in most publication.

to rejoinders at all: The keyword distribution is different, the speech recognition problem is hard and other genre and retrieval differences may make this approach a lot less effective (see also Sec. 1.2).

2.4.6.4 Automatic Summarization

Automatic summarization is a topic that is often treated in conjunction with information retrieval. Indeed most practical summarization techniques today use empirical techniques, measure the importance of a segment using topicality and most recently eliminate redundant sentences (Goldstein and Carbonell, 1999). Summarization is also commonly used in user interfaces to search engines. Zechner and Waibel (2000a) in our own working group apply summarization to oral communication and address the problems of inaccurate machine transcripts, lack of clause boundaries, distributed information (question/answer pairs), disfluent and unreadable speech and lack of topic boundaries. Within the information retrieval community the issue of multi-document summarization became a research topic of its own (Baldwin et al., 1999; Goldstein and Carbonell, 1999; Mani and Bloedern, 1997; McKenna and Liddy, 1999; Radev and McKeown, 1998; Radev et al., 2000; Stein et al., 1999). Those techniques are often related to the keyword based approach and make use of the semantic content of words. Some of the approaches refine the keyword based approach by information extraction and filling out forms to represent the documents (Radev and McKeown, 1998). Multi-document methods might become important for summarizing meetings since they are often continuations of the same discussion (see also Sec. 2.4.6.5).

2.4.6.5 Multi Conversation Indexing

Retrieving rejoinders such as meetings is a relatively new research problem. Rejoinders however are not isolated from each other. They are bearing relationships to each other in many ways, by the participants that attend, the topics, reporting and continuation relationships and so forth. The interesting question is whether those relationships can be exploited.

This thesis falls short of exploring this question in experimental work due to lack of data resources that are very expensive and would have to be prepared by ourselves. There are relationships between conversations that may be exploited (time, participants, topic, location), between the participants (social status, relationship) and by the role of the participant in the conversation (judge, meeting leader, . . .). If for example meeting A is strongly related to meeting B by the participants, time and location one could hypothesize that the topics of the two meetings are related or that one meeting is a continuation of the previous one. Or in a more complex scenario if person A is visiting some company and is back at its mother company it may be hypothesized that A is reporting about the meeting at the first company. Databases that would allow these kinds of studies would need in the order of hundreds of conversations over time from a varying group of people that are somewhat connected to each other.

Social network analysis provides the tools to analyze such relationships and the INSNA Web site (INS) provides a listing of resources including a refereed online journal (Journal of Social Structure). Jensen (1997) applied machine learning to social networks with applications to fraud detection and surveillance. A definition of social network analysis is given by Friedman:

> Social network analysis is focused on uncovering the patterning of people's interaction. It is about the kind of patterning that Roger Brown described when he wrote:
>
> ”Social structure becomes actually visible in an anthill; the movements and contacts one sees are not random but patterned. We should also be able to see structure in the life of an American community if we had a sufficiently remote vantage point, a point from which persons would appear to be small moving dots. [...] We should see that these dots do not randomly approach one another, that some are usually together, some meet often, some never. [...] If one could get far enough away from it human life would become pure pattern.”
>
> Network analysis is based on the intuitive notion that these patterns are important features of the lives of the individuals who display them. Network analysts believe that how an individual lives depends in large part on how that individual is tied into the larger web of social connections. [...]

In our current retrieval scenario the user of a retrieval system may chose to retrieve all meetings in which John was talking a lot, all meetings that happened in one place versus another one and during a certain time-frame: The analysis is therefore shifted to the user instead of actively employed by the system which might be easier to understand and sufficient for smaller databases. Simple first analysis steps could be to cluster meetings by their participants which can be done easily on our database: A small experiment immediately showed meetings of the various subgroups of Interactive Systems Labs at CMU. Clearly this field needs further analysis, given the novelty of the component techniques and the lack of data resources no direct studies could be performed.

2.4.7 Machine Vision

Most current machine vision approaches are very problem specific. For that matter only applied work aimed at information access in oral communication scenarios will be referenced. Sec. 2.3.3 discusses work by the *Informedia* project and some of their user interfaces build on machine vision algorithms that create summaries using “key-shots” (Xiong et al., 1997). Other related algorithms try to find shot-cuts in video sequences and Ford et al. (2000) is a good introduction to techniques and questions of evaluation. The shot-boundaries deliver a simple and straightforward segmentation which might be sufficient in some video genre. In some ways they correspond to topic segment boundaries while key shots correspond to the summary of a topical segment. Eickeler and Mueller (1999); Saraceno (1999) are indexing videos and make use of generic restrictions of standard video productions.

These solutions however assume that the video sequence was recorded and edited for human consumption and therefore contains shots, are accompanied by music etc. For the use in systems that process meetings one can’t assume to have an edited video and those approaches are therefore irrelevant. Clarkson and Pentland (1999) presents a very interesting approach that features unsupervised clustering of auditory and visual information to find different scenes a user is attending. This approach is not relying on shot-boundaries but rather on significant changes in the environment.

Gross et al. (2000) and Bett et al. (2000) address the problem of finding a person in a room or around a table, track that person and identify it. The identification can be improved when audio and video information are evaluated together. The question of finding and positioning a person in a room using vision is interesting since not everyone is actually participating verbally in a meeting or may only participate for part of the meeting. The participants may also indicate to belong to certain group or subgroup by the way they are sitting or standing together in a room. Stiefelhagen et al. (2000) uses unintrusive methods to measure the gaze of humans using video camera shots and is therefore able to make more precise predictions on the speaking situation. Mikic et al. (2000) monitor the activities in a meeting room environment by tracking the individuals with four cameras, allowing localization. This information is combined with face-id and gaze detection such that for each participant the location and the gaze are known at all times. They claim that this could be used to decide which camera provides the most meaningful pictures of the meeting room. Quek et al. (2000) is an early attempt explore the joint analysis of gesture, speech and gaze to structure discourse however the system hasn’t been evaluated fully. Finally the body posture (Takahashi, 2000) may be important to understand and may indicate whether someone is paying attention, opposing someone else or trying to get the floor.

One may also construct a simple system which determines the identity of people in a room (using badges) and takes snapshots every couple of seconds to allow for a recollection of the situation (Eldridge et al., 1991). They report that the most useful property of the snapshots was to see the participants and objects in the room. This may support a view that video is mostly supporting situational information like the speaker identities and objects. There may be other ways of capturing and displaying this information and the usefulness of the snapshot approach may make the database size manageable for recording large databases.

2.5 Industry standards: MPEG-7 and RDF

The resource description formalism (RDF) (Brickley and Guha, 2000) is an object oriented format which was proposed by the W3C as a meta-description for content and is being used by the popular WWW browser netscape to describe addressbooks. In the digital library project RDF was used as a mediation language and active research is done on the object oriented semantic and execution of the model. The RDF syntax – which is based on XML (Bray et al., 2000) – is mostly agreed upon while there is an active process of extending it, e.g. by topic maps to facilitate semantic browsing. On the other hand no special provisions for audio or dialogue data were made.

MPEG-7 (Martínez, 2000) on the other hand was intended to be an extension of the largely successful MPEG-2,3,4 standards for audio and video and adds meta-level descriptions to multimedia documents. It was agreed to use XML as the basis of the description language however special extensions are proposed to handle the representation of binary data. MPEG-7 however is not integrated with RDF. MPEG-7 is a standard of gigantic proportions with hundreds of pages of documents. To support the work in this thesis the basic abstractions of dialogue acts, games, topical segments and activities need to be represented as well as further annotations and features over these special segment types. While MPEG-7 supports a segment representation for audio these segment types are not explicitly supported which is a lack in the standard especially since many of the abstractions used in this work have been known. Some of these segment types would also need special attributes, a dialogue act for example is produced by a speaker, carries a dialogue act type and an emotion is attached to it. While XML allows to represent those in informal descriptions it would be useful to use explicit fields for it. MPEG-7 also doesn't support the sufficient statistics for summarization algorithms like Zechner and Waibel (2000a,b) which would require to store relevance scores at the level of dialogue acts or sequences of dialogue acts. MPEG-7 is therefore missing core-capabilities for fairly basic dialogue analysis and summarization which should be part of a standard that is aiming to describe all kinds of multimedia content.

Part of the reason that new requirements are being mandated is by this work is that MPEG-7 was born out of the need to describe broadcast media such as movies. Broadcast media is usually cut into sequences by the producers and it is rather straightforward to find cuts and pick frames that represent the time between cuts. For our applications we can't assume that the audio (and any video) is presegmented by a human, rather we want to derive that information. On the other hand the situation itself is somewhat static as it is a rejoinder which is bound by physical constraints.

Another shortcoming of MPEG-7 is that it can't handle data stored in annotation graphs, which is being developed as a new standard to describe linguistic resources by the LDC and is being backed up by tools in project ATLAS (Bird et al. (2000), see also `http://www.ldc.upenn.edu/atlas/`). Annotation graphs allow for events to be described on multiple timescales even if those timescales are only weakly synchronized. This is necessary when either the recording or the annotation doesn't allow to state exact timestamps for each event on a single time scale. While it may be assumed that all recordings will be digital in the nearby future and that exact time stamps can be attached for each sample the problem of annotating the begin and end of events is still present. Given the size and importance of both efforts with fairly related goals it would seem natural to investigate more cross-fertilization than currently. Similarly the complex representations possible in RDF cannot be used in MPEG-7 such that it is missing abstract querying and mediation capabilities. Still MPEG-7 is the most comprehensive standard for indexed multimedia documents to date.

Chapter 3

Dialogue Act and Game Detection

3.1 Introduction

Dialogue acts are units about as long as an utterance conveying a (basic) speaker intention such as a statement, backchannel or question. A dialogue game is a short sequence of dialogue acts constituting an elementary exchange pattern such as a question/answer pair or statements with backchannels (for more details see Sec. 2.4.3.3). They are therefore indicative of the activity that is being performed (Sec. 2.4.2.2 and Sec. 4), encode dominance (Sec. 4.2.7) and may be important units in speech summarization (Zechner, 2001). Some of the information encoded in dialogue acts may also be reflected in microlevel features (e.g. words, parts-of-speech) however dialogue acts are an interesting abstraction since they may be detected efficiently from speech and statistics based on them may be more portable across domains than statistics based on microlevel features.

The most thorough past work on dialogue act detection was performed at the John Hopkins LVCSR Summerworkshop in 1997 (Stolcke et al., 2000) with the participation of the author. This part of the thesis builds on the results of this workshop and extends them with the detection and segmentation of dialogue acts, the discriminative training of the classifier and the joint detection of dialogue acts and games.

This chapter first introduces classifiers for the detection of dialogue acts and games independently of the context with a given segmentation. The standard ngram backoff language model approach to this problem is described in Sec. 3.2 and is compared to a maximum entropy (Sec. 3.2.5) and a neural network (Sec. 3.3) approach. Ngram sequence induction is discussed in Sec. 3.4 to aid the construction of higher order models for the neural network approach. The necessary extensions for context dependent dialogue act and game detection are discussed in Sec. 3.5. Sec. 3.6 presents the results of these techniques. The chapter concludes by addressing the problem of the integration with a speech recognition system (Sec. 3.7), the application of the techniques presented in this chapter to parsing and emotion detection (Sec. 3.9, Sec.3.8). A conclusion for this chapter is given in Sec. 3.10.

The basic idea for the classifer approach was already presented in Finke et al. (1998) and elaborated in Ries (1999a) and early results on dialogue game annotation have been reported in Levin et al. (1999). The chapter is featuring a lot of material going beyond those publications and the comprehensive presentation of the material is new. Some early work has been carried out by the author in conjunction with the LVCSR summer workshop in Baltimore in 1997 and it is clearly marked as such (Stolcke et al., 2000).

3.2 Language Modeling and Language Model Classifiers

3.2.1 Introduction

In a speech recognition system acoustic and language knowledge are modeled independently. The fundamental analysis is that in order to determine the words sequence $\mathbf{W}$ given the acoustic evidence $\mathbf{A}$ a generative model can be constructed using Bayes' rule (Jelinek, 1989):

$$\mathbf{W}^* = \mathrm{argmax}_{\mathbf{W}} p(\mathbf{W}|\mathbf{A}) = \mathrm{argmax}_{\mathbf{W}} p(\mathbf{A}|\mathbf{W}) \cdot p(\mathbf{W})$$

In order to build a speech recognition engine the problems of estimating the acoustic model ($p(\mathbf{A}|\mathbf{W})$) and the language model ($p(\mathbf{W})$) and searching over the models (argmax) have to be solved. This model allows for a high level division of labor between different components both on the modeling as well as on the software side. Language models have been researched for a long time in speech recognition and may be useful for dialogue modeling as well.

Language models for speech recognition handle sequence information in a natural way. If w_0 is always a special "begin of sentence symbol" and w_n is always a special "end of sentence symbol", the probability of a sentence $\mathbf{W} = w_0, \dots, w_n$ can be decomposed as:

$$p(\mathbf{W}) = \prod_{i=1}^{n} p(w_i|w_0, \dots, w_{i-1})$$

Estimating $p(w_i|h)$ for an arbitrary history $h = (w_0, \dots, w_{i-1})$ directly is not practical since the vocabulary of speech recognition systems can be very large and the number of possible histories is exponential in the maximum length of the histories allowed. A common simplification are ngram models (also called n-th order Markov model or k-testable languages) which use the approximation

$$p(w_i|w_0,\ldots,w_{i-1}) \approx p(w_i|w_{i-n+1},\ldots,w_{i-1})$$

Common sizes for n are one (unigram), two (bigram) or three (trigram). Ngram models are a good approximations for practical language models in speech recognition and are common in commercial and research systems. The application of bi- and trigram models in the search procedures is also well understood and has been under continuous investigation.

Language models and classifiers based on them are popular in the discourse community since they are easy and fast to built, tools for their construction are available, and the individuals have been acquainted with them. Additionally language models can be used to model both the sequence of words as well as the sequence of dialogue acts. In collaboration [1] the author conducted experiments (Stolcke et al., 2000) that showed that other known techniques in language modeling did not outperform a simple bigram approach for the modeling of dialogue act sequences [2]. The other use of ngram backoff models is the modeling of the word sequences in dialogue acts. This model can be improved by training these “language model classifiers” in a discriminative fashion.

This section introduces the standard ngram backoff model (Sec. 3.2.2) and describes how it can be used for classification (Sec. 3.2.3). In Sec. 3.2.4 work on the discriminative training of ngram classifiers will be shown. Furthermore maximum entropy and exponential models will be discussed in more detail in Sec. 3.2.5.

3.2.2 Ngram Backoff Models

A lot of data may be needed for the reliable estimation of ngram models. If, for example, a vocabulary of 100.000 words is assumed, the text has to be at least 10^{15} words long such that every trigram is seen at least once. In a typical text corpus, however, a large amount of the trigrams tokens is covered by a fairly small set of trigram types which are repeated a lot (Zipf’s law, (Zipf, 1935)). This observation indicates that the corpus size needs to be much larger than the first lower bound given above and that most possible trigrams haven’t been even observed.

[1] The collaboration was in conjunction with the LVCSR summer workshop in 1997 at Baltimore’s John Hopkins University. Especially Noah Coccaro and Andreas Stolcke were close collaborators in these experiments.

[2] Tab. 3.3 shows that the joined detection of dialogue acts and games improves the dialogue act detection: dialogue games therefore provide improved dialogue act detection as opposed to other techniques explored earlier, which is a significant new finding.

The most widely used model to address this problem is backing off to a bigram distribution if a trigram hasn't been seen, and then to the unigram if the bigram hasn't been seen either ("backoff-models"). Another problem of maximum likelihood estimates for ngrams is that they drastically overestimate the likelihood of ngrams in an independent test text if the ngram has been observed in a training text, especially for types with small observed counts (Chen and Goodman, 1996; Chen and Rosenfeld, 1999). This problem is addressed by smoothing techniques which try to obtain better estimates than maximum likelihood could offer. More formally, for a fixed vocabulary V and an observed context h, a backoff model for $p(w|h)$ can be defined as:

1. determine a backoff distribution $p(w|h')$ for a context h' that is more general than h, for example $h' = w_1, \ldots, w_{n-2}$ for $h = w_1, \ldots, w_{n-1}$ in the case of a standard ngram backoff model. If h is already the null context, choose the uniform distribution over V.

2. estimate $\hat{p}(w|h) > 0$ for all words w that have been observed in context h, and let $\hat{p}(w|h) := 0$ for all other words, such that $0 \leq \sum_{w \in V} \hat{p}(w|h) \leq 1$ [3].

3. $$\begin{aligned} p(w|h) &:= \begin{cases} \hat{p}(w|h) & \text{if } \hat{p}(w|h) > 0 \\ \beta(h)p(w|h') & \text{otherwise} \end{cases} \\ \beta(h) &:= \frac{1-\sum_{w \in V \wedge \hat{p}(w|h)>0} p(w|h')}{1-\sum_{w \in V \wedge \hat{p}(w|h)>0} \hat{p}(w|h)} \end{aligned}$$

Various discounting techniques can be used for calculating $\hat{p}(w|h)$ for ngram models. A fairly effective yet simple technique is called absolute discounting. The absolute discounting estimate for the set of words W that have been seen in context h is $\hat{p}(w|h) = \frac{count_w - d}{\sum_{w \in V} count_w}$ with $0 < d < 1$ [4]. The construction of an ngram backoff model is therefore fairly easy and will be illustrated for a bigram model:

1. estimate a unigram model

2. count all bigrams in the text

3. estimate d for the bigram distribution

4. discount the bigrams and calculate the probabilities $\hat{p}(w|h)$ for each context h.

5. estimate β using the unigram model

[3] Typically $0 < \sum_{w \in V} \hat{p}(w|h) < 1$ such that $\hat{p}$ is not a probability distribution. $1-\sum_w \hat{p}(w|h)$ is the probability mass that will be used for backoff.

[4] A commonly used value for d is $d = \frac{\mathrm{fof}(1)}{\mathrm{fof}(1)+2*\mathrm{fof}(2)}$ where $\mathrm{fof}(n)$ is the number of (w,h) types that occur n times in the corpus (Ney et al., 1994).

In practice just counting and estimating the model can be computationally expensive such that simplicity is an important property of language models.

The backoff distribution has the property that it is only used if an actual backoff occurs. However the standard approach assumes that the backoff distribution of a n-gram model is the a standard (n-1)-gram model. This observation has been exploited by Kneser and Ney (1995) and is implemented in the language model toolkit of the author. The only modification necessary to implement their procedure is to modify the counts of the ngram tables for backoff distributions. However it is not used for construction language model classifiers since it does not improve the classification results in practice while it may be useful in speech recognition.

3.2.3 Language Model Classifier

Classifiers based on n-gram backoff models are generative classifiers: They model the full input distribution instead of just modeling the output given the input. To determine the best dialogue act l given a sequence of words $W = w_0, \dots, w_n$, Bayes' formula is used:

$$L = \text{argmax}_l\ p(l|W) = \text{argmax}_l\ p(W|l) \cdot p(l)/Z(W) = \text{argmax}_l\ p_l(W) \cdot p(l)$$

using a normalization term $Z(W) = p(W) = \sum_l p_l(W) \cdot p(l)$. We use a different language model $p_l(w) := p(w|l)$ for all dialogue act types l and estimate them separately. If bigram models are used for $p_s(W)$ the classifier reduces to:

$$L = \text{argmax}_l\ p(l|W) = \text{argmax}_l \prod_i p_l(w_i|w_{i-1}) \cdot p(l)$$

This argument obviously generalizes to higher order ngram models for $p_l(W)$. It will be extended to include the automatic segmentation and modeling of dialogue context (Sec. 3.5). Neural networks and exponential models will be presented as alternative classification methods.

3.2.4 Discriminative Training for Language Model Classifiers

Language model classifiers are generative models which leads to a simple and fast model estimation algorithm. However generative models are often ineffective, especially when a large number of irrelevant features need to be modeled and the amount of training data is limited.

However there are some aspects which are specific to language model classifiers which make discriminative training attractive: If in a context h some models might backoff while others don't; this may introduce a systematic bias since backing off is more frequent for categories for which little training data exist.

Discriminative language model classifiers have another advantage: If two words tend to cooccur and present evidence for one and the same class a discriminative classifier is able to weight cooccurence instead of duplicating the evidence (see also Nigam et al. (1999)). An example from dialogue classification could be

`BEGIN_OF_SENTENCE yeah right`

which is a strong indicator for a backchannel, however so is either `yeah` or `right` by itself. A generative unigram classifier would have to assume that each of the words `BEGIN_OF_SENTENCE yeah right` occur independently which they don't. A bigram model on the other hand would not be fooled since it would capture the local correlation. A discriminatively trained classifier would weigh `yeah` strongly as an indicator for backchannels but `right` could be weighted lower since it is only a consequence of the presence of `yeah`. Especially members of very frequent ngrams obviously frequently cooccur, causing this to be an issue. Consequently Sec. 3.6 shows that discriminatively trained classifiers usually don't make good use of higher order ngram models whereas the generative models do.

It is well know in the HMM literature that generative classifiers often perform worse than discriminative classifiers with the same parameter set since they have to model the full input distribution. However generative classifier are *much* easier to train and therefore "discriminative training" algorithms have been proposed that offer higher performance while keeping some of the simplicity of the generative training procedure. This is often achieved by either using simplified optimization criteria or by taking smaller steps towards the solution. Recently extensions for n-gram model estimation have been proposed and evaluated experimentally (Ohler et al., 1999; Warnke et al., 1999).

Another implementation of discriminative training for language model classifiers can be achieved using exponential models. Sec. 3.2.5.4 shows how the language model classifier can be embedded exactly into that framework such that the parameterization of the models are identical. The model can then be trained either using the highly efficient iterative scaling framework or using gradient descent (Sec. 3.2.5.3). The traditional discriminative training techniques in comparison suffer in a number of ways and will therefore not be reviewed in detail:

- they require constraints on the parameter space which are unnecessary in the exponential model and don't seem to make sense for the classifier.

- both the optimization criterion and algorithm are suboptimal compared to iterative scaling.

- discounting techniques can't be easily introduced in a principled way (can be realized by Gaussian priors in the exponential model).

- the formulation of Warnke et al. (1999) and Ohler et al. (1999) requires the separate training of language model interpolation parameters and ngram entries.

3.2.5 Maximum Entropy and Exponential Models

3.2.5.1 Introduction

Maximum entropy models have been popular in language modeling for speech recognition as an alternative to ngram backoff models since they provide a more principled way of feature integration and a clean theoretic background. In this section it will be explored how maximum entropy models are related to language model classifiers, especially for higher order ngrams. These observations have practical importance since they deliver recipes for the selection of features for the exponential model. First the maximum entropy principle and the relation to exponential models will be introduced. The parameter estimation methods and their limitations as well as an efficient reduction of ngram models to exponential models will be presented. Later neural networks will be discussed which are an extension of the exponential modeling approach discussed here (Sec. 3.3).

3.2.5.2 The Maximum Entropy Principle and Exponential Models

The maximum entropy principle is the rule to pick – among a set of probability distributions – the one with the maximum entropy or in other words the one which has the smallest distance to the uniform model while fulfilling a set of constraints. The constraints have the form

$$C_i = E_p \;\; f_i(c, \mathrm{data})$$

where c is the classification of data in the example at hand. It is common to estimate the values C_i for the constraints from a training data set as the empirical expectation of the features. The theory of maximum entropy modeling shows that for all consistent sets of constraints there exists an exponential model which is a maximum entropy solution under these constraints:

$$p(c|\mathrm{data}) = \frac{e^{\sum_i \lambda_i \cdot f_i(c,\mathrm{data})}}{Z(\mathrm{data})}$$

where λ_i are the parameters of the model. Additionally it follows immediately that the exponential model of the maximum entropy solution is also the maximum likelihood solution in the family of exponential models (for more complete introductions to maximum entropy modeling and the iterative scaling algorithm see Berger (1998); Csiszar (1996)).

3.2.5.3 Parameter Estimation

Exponential models can be estimated very efficiently using *iterative scaling* if the sum of the features $\sum_i f_i$ is a small integer for all instances of the training data set. More precisely $\sum_i f_i$ has to be a constant integer for the *generalized iterative scaling algorithm* (GIS) and a (small) integer for each individual instance in the training set for practical implementations of the *improved iterative scaling algorithm* (IIS) to apply. Otherwise gradient descent can be used which converges to the globally optimal solution since the objective function is convex in the parameters.

Iterative scaling usually converges after a small number of iterations for language modeling applications whereas gradient decent may often need many iterations. It is therefore important to use the iterative scaling algorithm in many situations. The proof of the iterative scaling algorithms uses Jensen's inequality and auxiliary functions in similar vein as the EM algorithm. Although the algorithm itself is rather simple it introduces a notational complexity that is unnecessary especially since it is not used directly in this work [5]. All models are solved using the RPROP algorithm since that algorithm can also be applied to neural networks with hidden units (Sec. 3.3) and the restrictions to the feature space do not apply. As long as the problem is still tractable this solution seems to be prudent since it makes the implementation easier and increases the flexibility of the system.

3.2.5.4 Embedding Language Model Classifiers into Exponential Models

In order to embed a language model classifier into an exponential model it is first shown that an ngram backoff model can be rewritten as a single term representation model. Using single term representation models a language model classifier can be embedded easily into an exponential model. We will only present an effective transformation for the ngram backoff models but it is much easier for ngram models based on linear interpolation (Jelinek, 1989).

A single term representation model $S = \langle V, S_1, \ldots, S_n \rangle$ consists of a vocabulary V and functions $S_1 : V \mapsto \mathcal{R}$ and $S_{i>1} : V^i \mapsto \mathcal{R} \cup \{\perp\}$. The intuition is to pick the longest ngram which is available in the model and take the score of that ngram. The single term score $\hat{S}_j : V^j \mapsto \mathcal{R}$ is obtained using the highest matching entry in S:

$$\hat{S}_j(w_1, \ldots, w_j) \quad := \quad \begin{cases} S_j(w_1, \ldots, w_j) & \text{if } j = 1 \vee S_j(w_1, \ldots, w_j) \neq \perp \\ \hat{S}_{j-1}(w_2, \ldots, w_j) & \text{otherwise} \end{cases} \tag{3.1}$$

[5] The experiments that the author performed on dialogue act sequence modeling, which are documented in Stolcke et al. (2000) but are not detailed in this document, use the iterative scaling algorithm. The work is not documented here since it has been part of a collaborative effort.

Unlike the language model classifier the values retrieved don't need to be probabilities and for each ngram the value of exactly one table is consulted. The probability of the sequence however can be computed the same way as for the ngram language model classifiers as discussed earlier [6]:

$$p(w_0,\ldots,w_m) = \prod_{i=1}^{m} \hat{S}(w_{i+1-n},\ldots,w_i)$$

An ngram-backoff model in general requires to sum over both ngrams and backoff weights therefore using more than one term per w_i in the product. By calculating all ngram probabilities any language model could be represented immediately but the goal is to store only as many (or fewer) parameter as the original model. Under the reasonable assumption that iff word w_j has been observed in context $h = (w_1,\ldots,w_{j-1})$ then word w_{j-1} has been observed in context $h = (w_1,\ldots,w_{j-2})$, the backoff weight calculation for the current word can be systematically distributed over the current and the last word. In practice this assumption is almost always fulfilled: In a trigram model this would mean that for every trigram $w_1,\ldots,w_3$ in the model the bigram $w_2,\ldots,w_3$ has to be part of the model as well. More formally for a backoff model $\hat{p}$ with backoff-weights β we define for word w_j in context $h = (w_1,\ldots w_{j-1})$

$$S_j(w_1,\ldots w_j) := \hat{p}(w|h)\frac{\prod_{k=2}^{j}\beta(w_k,\ldots w_j)}{\prod_{k=1}^{j-1}\beta(w_k,\ldots w_{j-1})}$$

otherwise $S_j(\cdots) := \perp$ [7]. Given a single term representation model S^c for each class $c \in C$ and a class prior $p(c)$ the language model classifier can be written as:

$$p(c|w_0,\ldots w_m) = p(c)\cdot\prod_{i=1}^{m}\hat{S}^c(w_{i+1-n},\ldots,w_i)/Z(w_0,\ldots w_m)$$

which can be interpreted as an exponential model: For each class c and m-gram

$$\{(c',w_1,\ldots,w_m)|S_m^{c'}(w_1,\ldots,w_m)\neq\perp\}$$

a feature $f_{(c',w_1,\ldots,w_m)}$ can be chosen with

$$f_{(c',w_1,\ldots,w_m)}(c,data) = \begin{cases} \# & \text{if } c = c' \\ 0 & \text{otherwise} \end{cases}$$

[6] The case of the begin and end of sentence are dropped since they complicate the notation.

[7] Minor modifications for the beginning and end of the sentence are necessary but would complicate the notation unnecessarily without adding to the understanding.

where $\#$ is the number of times the ngram $w_1, \ldots, w_m$ is chosen by the single term representation model S^c when calculating the sentence probability and $\lambda_{(c', w_1, \ldots, w_m)} = \log(S^{c'}_m(w_1, \ldots, w_m))$. Similarly we can choose features corresponding to $p(c)$. The features of this model sum up to $\mathrm{length}(\mathrm{sentence}) + 1$ which may be large and varied (although finite) for a fixed database; it is thus rather impractical to train this model using either GIS or IIS. However, if all features are normalized to sum to 1, both GIS and IIS can be applied (Nigam et al., 1999). On the flipside the second model can't be interpreted directly as a language model anymore and the information about the length of the data gets lost unless it is introduced explicitly as a feature. In dialogue act recognition however this information is very important, for example backchannels are very short whereas statements are very long. Additionally in the original formulation the expected number of occurrences of a certain ngram tokens per class is restricted, in the modified model the expected percentage of the ngrams is restricted. This is not an intuitive restriction if we think of features such as WH-particles that occur once in the beginning of a question: We want to condition on the fact that the WH-particle occured once and not in a certain percentage. In some applications such as topical text classification this might be acceptable, in others such as dialogue act classification it might not be. Another solution to this dilemma is to retain the second model but to encode the utterance/text length explicitly as a feature (for example features that are one for utterances of length one, two, three, ...). Training the exponential model (for example using iterative scaling) can be seen as a generalization of the discriminative maximum mutual information (MMI) training of Ohler et al. (1999); Warnke et al. (1999), since the exponential models are not required to be interpretable as a language model classifier. However one either needs to use gradient decent or general optimization algorithms to be able to reproduce the length modeling of the language model classifier accurately. Using the iterative scaling algorithm one has to either sacrifice the length information or model it explicitly. Since the training times are reasonable with gradient descent it is used and modeling techniques for iterative scaling are left for future investigation.

3.2.5.5 Multigram Models

Ngram backoff models are not the only higher order language model that is discussed in the literature. However, there seems to be only one competing approach to the incremental prediction model, namely the multigram model (Deligne and Bimbot, 1995). A multigram model calculates the probability of a sentence by the summation over all phrasal segmentations S where p' is a probability distribution over multigrams:

$$p(w_1, \dots w_m) = \sum_S \prod_i p'(w_{S_{i-1}}, \dots w_{S_i})$$

A phrasal segmentation for $w_1, \dots w_m$ is a list of integers

$$S = (S_0, S_1, S_2, \dots, S_{k-1})$$

with $S_0 = 0$, $S_i < S_{i+1}$ and $S_{k-1} = m$. It is easy to see that the induced probability over sentences is a probability distribution again. Furthermore one may assume that only for a limited number of phrases $p'(w_{S_{i-1}}, \dots w_{S_i}) \neq 0$ such that the summation may not be as expensive as it seems at first.

The summation however is still a very expensive operation and it does not allow a direct interpretation of the model. The training of the model requires the application of the EM algorithm and relies on a good phrase induction procedure. This model is also hard to include in a search procedure since it is not incremental – in order to simulate an incremental model one might therefore be tempted to use a Viterbi approximation. The Viterbi approximation would lead to a histogram of words and phrases that could be used as the input of an exponential model classifier and it may be an alternative to the input representation induced by the ngram backoff model. In prestudies this model never yielded good performance, though.

3.2.5.6 Conclusion

The language model classifiers based on the most common language model types can be embedded in exponential models containing the same set of parameters. These models can be interpreted as solution of a maximum entropy problem and efficient algorithms exist to solve them, especially if the input vector is normalized. However it is not always advisable to do that, neither in dialogue modeling nor in topic classification. In dialogue modeling we would like to capture crucial information about turn length and single term occurrence, in topic classification we would like to capture term repetition in an appropriate manner [8]. The new inventory of algorithms allows one to use the fast and simple generalized iterative scaling at the expense of a modeling deficiency which may be alleviated with

[8] It is well known for text classification that a normalized histogram is a suboptimal feature representation (Nigam et al., 1999; Wiener et al., 1995). For topic classification the author himself has experimented with exponential models on the Reuters database (Wiener et al., 1995) and found that sub-linear mappings of the counts (for example the square root) are important to apply. Using such mapping the author obtained similar results to Wiener et al. (1995) on the Reuters database with a standard network, however without using hidden units but using Gaussian priors.

additional modeling effort. The use of gradient decent allows for more general models and non-linearities and will be discussed in the next section. The discriminative training of language model classifiers (Ohler et al., 1999; Warnke et al., 1999) therefore seem to be an unnecessary exercise since (a) the models are a proper subclass of the models presented here, (b) the algorithms are idiosyncratic to the problem, and (c) the new method inventory allows for very fast and/or very rich modeling.

3.3 Neural Networks

In the previous Sec. 3.2.4 an exponential model was constructed that is able to train a language model classifier which had the general form:

$$p(c|f) = exp(A \cdot \vec{f} + B)/Z(\vec{f})$$

where the features f are encoded as a vector $\vec{f}$ (for example a histogram of the words in a segment), A is the parameter vector and B is a bias (corresponding to a prior or the $p(c)$ parameter in the language model classifier). A neural network of this structure is called a two-layer network using the softmax function as the output and short-cut connections. The natural pairing for this output function is cross-entropy as the error function (Bishop, 1995):

$$E = -\sum_n \sum_c t_c^n \cdot \ln \frac{p(c|f_n)}{t_c^n}$$

where the training examples have features f_n, a probability t_c^n that example n is associated with class c and the solution corresponds to the maximum likelihood solution of the exponential model. This pairing of error function and activation function can be derived in a very general manner by assuming that the class conditional distributions of the input vectors are generated by an exponential model (Bishop, 1995; Bridle, 1990). The structure of error function and output function are such that the backpropagation algorithm can be readily applied. This model can be seen as a generalization of the sigmoid logistic function to the multiclass case. One extension of the exponential model would be to use non-linearities such as $\tanh(\cdot)$:

$$p(c|f) = exp(A \cdot \vec{f} + B + C \cdot \tanh(D \cdot \vec{f}))/Z(\vec{f})$$

To speed up and improve the quality of convergence over the standard gradient descent the RPROP (resilient backpropagation) algorithm (Riedmiller and Braun, 1993) has been implemented, a very efficient search technique to speed up convergence for most network types which does not require much fine tuning. A

Gaussian parameter prior is used on the network weights; the prior has also been used with great success in text classification and language modeling (Sec. 3.2.5). The Gaussian parameter prior can be integrated easily in RPROP and has also been implemented in Zell (1993)[9]. RPROP is a local adaptive learning scheme that is determining the size of the learning step in an adaptive fashion. For each weight w_{ij} an "update value" $\Delta_{ij}^{(t)}$ is calculated in an iterative fashion [10]. The direction of the update is determined directly by the derivation of the error function whereas the size is determined by $\Delta_{ij}^{(t)}$:

$$\Delta w_{ij}^{(t)} = -\Delta_{ij}^{(t)} \cdot \text{sign} \frac{\partial E^{(t)}}{\partial w_{ij}}$$

After each weight update the "update value" is adapted depending on whether the sign of the derivative of the error function changed.

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} * \Delta_{ij}^{(t-1)} & \text{if} \quad \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta_{ij}^{(t-1)} & \text{if} \quad \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(t-1)} & \text{otherwise} \end{cases}$$

Additionally upper and lower bounds on the $\Delta_{ij}^{(t)}$ are used (50 and $10e^{-}6$) and $\eta^{+} = 1.2$ as well as $\eta^{-} = 1e^{-6}$ are chosen as suggested by Riedmiller and Braun (1993); Zell (1993). The initial update value is $\Delta_{ij}^{(0)} = 0.1$. One may interpret these values as "speed up a little", as long as your going in the right direction, and "really hit the breaks" once your going too fast. Upper and lower bounds make sure there is not too much oscillation due to "breaking down". This basic update rule is extended with a momentum term that is manually determined but kept fixed for every class of experiments (e.g. activity detection for CallHome Spanish). It is depends mostly on the size of the database. Training with a momentum term corresponds to the assumption that the weights have a Gaussian distribution with zero mean and a fixed variance. Unless noted otherwise the networks presented are three-layer networks with shortcut connections ($B \neq 0$) trained with RPROP and momentum.

[9] The neural network simulator used in this work has been implemented in JANUS and was originally developed by Jürgen Fritsch but has subsequently been improved by the author, specifically by RPROP and some efficiency optimizations. Additionally it has been tightly embedded into the Python class library that allows for the effective specification and training of classifiers along with other algorithms such as C4.5, SVM, k-NN, SGNG. The simulator allows highly effective treatment of feedforward networks with full (and partial) connections between layers (Bishop, 1995).

[10] Using a constant Δ the Manhattan rule is derived.

3.4 Feature and Ngram Sequence Induction

In the previous section we explored classifiers based on ngrams and it would be possible — as in the generative classification approach — to take all ngrams occuring in the text and hope that the model prior would discount those that cannot be estimated reliably. Although this argument is correct in theory the computational burden alone would justify a preprocessing step and some classifiers also work better if they are not swamped with irrelevant information. Instead it would be better to select only a small number of highly salient ngrams ahead of time which discriminate effectively between dialogue acts.

The simplest approach would be to just find phrases discarding the classification problem at hand. This problem has been studied for language modeling in speech recognition, e.g., earlier work by the author Ries et al. (1996). A second approach is to use the targets of the classification but not to take into account that a simple classifier could be built without any phrases. This approach has been pioneered by Gorin (1995), and Samuel et al. (1999) delivers a comparison of various metrics on the dialogue act detection problem. A third approach assumes a classifier for the problem which is enhanced with additional phrasal features. The most direct version of this approach, boosting using decision stumps (Schapire and Singer, 1999), performs this kind of feature induction. An alternative implementation of the third approach has been suggested by Pietra et al. (1997) and is based on maximum entropy modeling: The existing classifier serves as a prior and the improvement of the Kullback-Leibler divergence of the training data is measured for all potential features. Berger and Printz (1998) present a comparison of this technique to other induction criteria for exponential models and conclude that it is preferable since it is a natural match. Since our models are identical or similar to this model class this algorithm is a natural choice. We follow the fast implementation by Printz (1998) but add a Gaussian parameter prior to avoid overtraining through the inclusion of idiosyncratic phrases that may occur in only a few utterances. The addition of a parameter prior is designed to avoid overtraining by limiting the size of parameters while measuring the potential improvement and makes the model used for induction more similar to the exponential model (which includes the same prior). The advantage of this approach to ngram induction is that it solves problems such as the weighting of long vs. short and specific vs. unspecific ngrams in a theoretically sound way unlike Samuel et al. (1999). Fig. 3.1 displays word/part of speech phrases that are discriminative of emotions in our meeting database. We observed only small improvements from the addition of ngrams to dialogue act and game detection.

```
<s> <laugh>/Nc
<s> oh/UH
<s> mhm/UH
<laugh>/Nc <b>/Nc
<s> <laugh>/Nc <b>/Nc
<s> yeah/AFF
<b>/Nc <laugh>/Nc
<s> right/UH
NN <laugh>/Nc
<laugh>/Nc i/PRP
<s> <hm>/Nc
Nc <laugh>/Nc
if/CC we/PRP
at/PREP the/DT
<laugh>/Nc <p>/Nc
<uhm>/Nc <p>/Nc
<s> so/CC
i/UH mean/UH
yeah/AFF <laugh>/Nc
<s> well/UH
<b>/Nc the/DT
okay/UH <laugh>/Nc
NNS <laugh>/Nc
so/CC <b>/Nc
<laugh>/Nc <b>/Nc <laugh>/Nc
NNP <laugh>/Nc
<s> yeah/AFF <laugh>/Nc
<b>/Nc <uhm>/Nc
i/PRP have/VBP
<laugh>/Nc i/PRP don’t/AUX-N
<s> <laugh>/Nc <p>/Nc
hours/NNS <b>/Nc
this/DT NN
but/CC if/CC
NN is/VBZ
<laugh>/Nc yeah/AFF
and/CC then/UH
<b>/Nc <laugh>/Nc <b>/Nc
the/DT time/NN
is/VBZ that/WDT
<laugh>/Nc a/DT
<s> Nc <laugh>/Nc
<p>/Nc the/DT
<uh>/Nc <p>/Nc
<s> yeah/AFF <p>/Nc
```

```
<s> <b>/Nc <laugh>/Nc
this/DT guy/NN
<s> <laugh>/Nc i/PRP
that/WDT the/DT
we/PRP have/VBP
<laugh>/Nc he/PRP
<laugh>/Nc <>/Nc
no/NEG i/UH mean/UH
four/CD hours/NNS
no/NEG <laugh>/Nc
i/PRP don’t/AUX-N know/VB </s>
<b>/Nc but/CC
<s> okay/UH
i/PRP think/VBP
<s> <smack>/Nc
RB <laugh>/Nc
yeah/AFF <p>/Nc
big/JJ VB
<s> yeah/AFF <laugh>/Nc </s>
<s> mhm/UH <p>/Nc
no/NEG i/UH
<s> but/CC
oh/UH my/PRP\$
a/DT JJ
a/DT problem/NN
right/UH now/RB
<p>/Nc and/CC
<s> no/NEG
<noise>/Nc <laugh>/Nc
<laugh>/Nc <b>/Nc yeah/AFF
NN <p>/Nc
well/UH we/PRP
of/PREP the/DT
no/NEG but/CC
in/PREP the/DT
but/CC <laugh>/Nc
<uh>/Nc the/DT
<s> i/PRP i/PRP
for/PREP this/DT
to/TO use/VB
this/DT <laugh>/Nc
my/PRP\$ NNP
oh/UH okay/UH
just/RB a/DT
```

Figure 3.1: **Emotion Discriminative Word Ngrams:** The most discriminative word ngrams for emotion detection on the meeting task are shown. By eyeballing the phrases many phrases contain laughter and the begin-of-sentence symbol <s> as well as filled pauses. One of the sequences is a complete utterance which is a backchannel with a laughter. Emotion discriminating phrases were chosen for visual inspection since it is the most immediately visible class of ngrams that we may have intuitions about.

3.5 Context Dependent Detection

3.5.1 Introduction

This section describes the general dialogue act detection model which extends the simple classification model presented above which didn't need to be concerned about the segmentation of the discourse into dialogue acts and the inclusion of context. The dialogue game detection model works essentially the same.

The task is – given a word sequence – to determine a segmentation and a labeling with dialogue acts. Using a *maximum a-posterior* approach we need to find a segmentation S and labeling L of the sequence of words W such that

$$S, L = \text{argmax}_{S,L} p(S, L|W) = \text{argmax}_{S,L} p(L|S, W) \cdot p(S|W) \cdot p(W)$$

First, we notice that while one could first predict W, segment it using S and finally label it using L, an alternative modeling approach is to combine W and S into the new random variable $Y = y_0, \ldots y_{n-1}$ such that each segment is represented by one element in Y. The last token of each y_i is a special "end of segment" symbol and $Y' = y'_0, \ldots y'_{n-1}$ is the mapping of Y onto the words without the end of segment symbols. The concatenation of Y' yields W again. We make two crucial assumptions that are also made in related work in dialogue modeling (Stolcke et al., 2000).

independence of utterances utterances (segments in Y) are independent of each other if the labels of the dialogue acts L are given: $p(Y|L) = \prod_i p(y_i|L)$

locality of dialogue information y_i is only dependent on the label of the dialogue act l_i, not on the other dialogue act label: $p(y_i|L) = p(y_i|l_i)$

Both assumptions are invalid in general: The first assumption is most certainly violated since utterances in a conversation are related by topic (e.g. keyword repetition) which is not captured in the dialogue model at all and similarly anaphora are not modeled in a dialogue act model. It is also violated in named entity tagging: If "`Mr. Salad`" occurs in a text `Salad` may need to be tagged as a named entity but the information that it is a named entity may stem from `Mr.` which is outside of the named entity. Unless there is a special segment for `Mr.` which captures this constraint it would be lost [11]. The second assumption is violated if the dialogue act tags are not rich enough to capture all information about

[11] It is possible to change the analysis above such that word level information from earlier segments is available for classification as well. However this is not relevant for dialogue act modeling while there are other modeling problems such as named entity tagging where it might be crucial. There may also be other discourse phenomena which could benefit from the knowledge of previous words but they won't be discussed here.

the dialogue act in the dialogue annotation scheme but the information could be inferred from context. Using these simplifications the following model can be derived:

$$Y, L = \mathrm{argmax}_{Y,L,Y'=W} p(Y|L) \cdot p(L) = \mathrm{argmax}_{Y,L,Y'=W} p(L) \cdot \prod_i p(y_i|l_i)$$

The task – besides the search – is therefore reduced to the estimation of the dialogue model $p(L)$ and $p_{l_i}(y_i) := p(y_i|l_i)$ which will be discussed now.

3.5.2 Estimating the Dialogue Model $p(L)$

$p(L)$ describes which sequences of dialogue acts make sense. It could therefore be called *dialogue* model since it describes the dialogue dependencies. In the case of context independent classification we fix the segmentation ahead of time but also assume a unigram model for $p(L)$ denying contextual effects on individual dialogue acts. The author has experimented with various types of $p(L)$ on the Switchboard database in collaborative work on the Summer Workshop at John Hopkins University (Jurafsky et al., 1997b; Stolcke et al., 2000). The results should also be valid on the CallHome Spanish and may be summarized as follows: As a baseline $p(L)$ may be modeled by an ngram backoff model. It turns out that one has to include the channel information into the ngram such the model is aware of speaker changes. Most of the effects have been observed using bigrams with small improvements from trigrams [12]. A maximum entropy model (Sec. 3.2.5) that is capable of including ngrams, distant ngrams, dialogue acts from the same/the other channel, speaker-change and dialogue act triggers was built. While the model restricted to ngram features performed as well as the backoff model we haven't observed improvements from the more sophisticated models. In another attempt to improve $p(L)$ an interpolation between a cache model (Kuhn and de Mori, 1990) and an ngram backoff model was tested. The assumption was that global dialogue constraints such as dominance in one conversation or male/female dialogue strategies would be captured by this model. As we will see in Chapter 5 the intuition that the speaker distribution are constrained by an activity is valid but the dialogue act model may take care of this constraint at a local level. Ngram models, usually bigrams including channel information or unigrams, will therefore be used to model $p(L)$.

[12] This part of the work has been mostly carried out by Noah Cocarro from the University of Boulder.

3.5.3 Estimating $p_{l_i}(y_i)$

If we use ngram backoff models to estimate $p_{l_i}(y_i)$ as in most previous approaches this paragraph would be unnecessary since ngram backoff models can be used to estimate this quantity directly. However if we want to be able to make use of discriminative methods a different approach has to be taken:

$$p_{l_i}(y_i) = p(l_i|y_i) \cdot p(y_i)/p(l_i)$$

and one may observe that replacing $p(l_i|y_i)$ with a language model classifier would yield the respective individual language model. Estimating $p(l_i)$ is easy since the number of dialogue acts is typically small. $p(l_i|y_i)$ may use any of the dialogue act classifiers described in Sec. 3.2 and Sec. 3.3.

$p(y_i)$ is responsible for the segmentation and at first glance one would try to use a (ngram backoff) language model to approximate it directly — the resulting procedure is the p_{all} *segmentation*. A better alternative could be to model $p(y_i)$ by a mixture of dialogue acts $p(y_i|l_j)$ such that long range information about the dialogue act can be propagated. This modeling approach might be called *dialogue aware segmentation* since it preserves dialogue act information lost in the ngram approximation for segmentation:

$$p(y_i) = \sum_{l'} p(l') \cdot p_{l'}(y_i) \approx p(l_i) \cdot p_{l_i}(y_i|l_i)$$

The direct approximation of $p(y_i)$ by an ngram model is the p_{all} segmentation, the ngram approximation of the second term is called the *dialogue aware* segmentation. The last approximation reduces the sum over all ngram models to the component of the "correct" model, the ngram approximation of this model will be called is the *Viterbi* segmentation model. The *Viterbi* segmentation model assumes that the "correct" dialogue act (l_i) contributes most to the summation which might be especially justified when the global optimization is performed. The results for dialogue act detection seem to indicate that the p_{all} segmentation performs as well as the dialogue aware segmentation model and that the *Viterbi* model performs significantly worse (Tab. 3.3, p. 99).

3.5.4 Segmental Modeling and Multi Channel Search

Dialogue act classification and a number of related problems that make use of a "chunk and label" paradigm have been studied by various authors in the recent past (Finke et al., 1998; Jurafsky et al., 1997a; Nagata and Morimoto, 1994; Reithinger et al., 1996; Stolcke et al., 2000; Taylor et al., 1997; Warnke et al., 1997; Wright, 1998). The dialogue act detection problem can be seen as a two-level

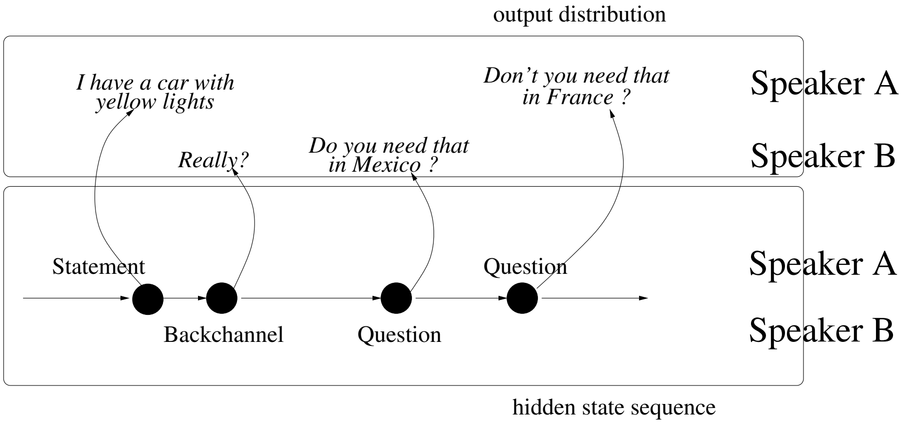


Figure 3.2: **HMM of Dialogue Acts:** The underlying structure of the dialogue model is a probabilistic finite state automaton (PFSA) of dialogue acts which emits sentences rather than words. The dialogue acts are uttered by speaker **A** and **B**, the dialogue model operates on the sequence of dialogue acts from both speakers. Reprinted from Ries (1999a)

HMM where the states are dialogue acts and the symbols emitted are sentences. If the dialogue acts were non-overlapping and we did not want to benefit from direct classification approaches the respective Viterbi-search algorithms would apply directly (Fig. 3.2).

The problem of segmentation (and labeling) is simple when the dialogue acts are treated separately for both channels since there is no overlap between channels to consider and the segments are adjacent. The possible segmentations and labelings are searched by an A^* algorithm (Hart et al., 1968) which can use ngram backoff models for $p(L)$ and $p_{l_i}(y_i)$. The A^* search state contains the information when the current segment has been opened and what the labels of the previous dialogue acts were. The lookahead of the A^* search is a standard language model. The search proceeds by extending a segment by a word or closing a segment. If a segment is closed a special lookahead is used since the next word is known to be the beginning of a dialogue act. A closed segment can be extended by the search by opening a new segment for each dialogue act type. The A^* search procedure is also capable of generating a lattice [13] of segments by enumerating all segments that have been closed during the A^* search.

Modeling multiple channels is more complicated, especially if there is no natural segmentation as in CallHome Spanish. In Sec. 3.5.2 the model for dialogue act sequences $p(L)$ has been discussed and it has been argued that the model needs to take the dialogue segments on both channels in their temporal order into account. The implemented A^* is able to handle this case by extending the search state with information about multiple channels and starting new segments as the dialogue act model prescribes. A standard dynamic programming approach would become significantly more difficult since the lattice that can be searched has become significantly larger and more complex in structure. An alternative is to proceed in three separate steps which allows to apply neural networks efficiently (Fig. 3.3):

generate segment lattice cluster all dialogue acts into one class and generate one general ngram backoff model to segment the lattice into dialogue acts [14]. It is a reasonable approximation of all likely dialogue act segments (Fig. 3.3, ①) and the segmentation can be done separately for all channels.

generate dialogue act lattice given the segment lattices for each channel, replace each segment z in the segment lattice by segments for all dialogue act types s with weight $p_s(z)$ (Fig. 3.3, ②).

extract best sequence search over the dialogue act lattice with a dialogue model $p(L)$ taking into account the weights on the segments (Fig. 3.3, ③). The A^* search state contains the information at what timepoint the last segment ended for each channel and what the last dialogue acts were. The search expands the segments such that the resulting sequence of segments is ordered by the start time of the segments. This is achieved by always expanding only the segments on channel(s) that have the earliest end time for the previous segment. The score is a combination of the dialogue model $p(L)$ and the weight $p_l(z)$ of the dialogue act l on segment z.

No degradation has been observed with this less direct approach when compared to the direct approach on a single channel and it is therefore used throughout

[13] The term lattice is used for a directed acyclic graph with a start node that can reach all nodes in the lattice and an end node that can be reached from all nodes. This terminology is common in speech recognition.

[14] This model corresponds to the p_{all} model which is used for dialogue act segmentation in the overall model.

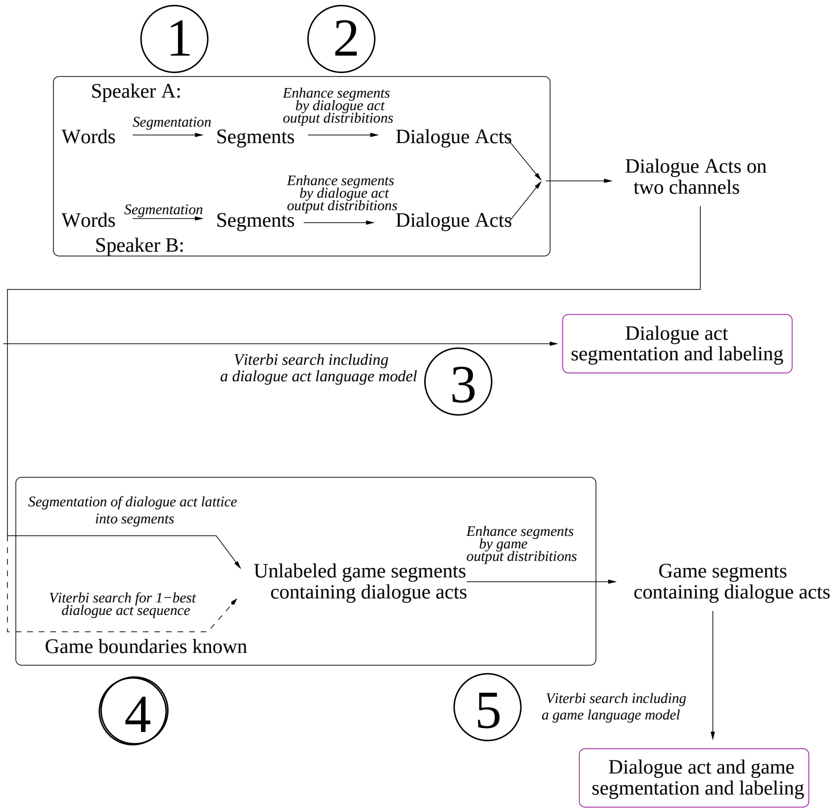

1
2
Speaker A:
Words
Segmentation
Segments
Enhance segments by dialogue act output distribitions
Dialogue Acts
Words
Segmentation
Segments
Enhance segments by dialogue act output distribitions
Dialogue Acts
Speaker B:
Dialogue Acts on two channels
Dialogue act segmentation and labeling
Viterbi search including a dialogue act language model
3
Segmentation of dialogue act lattice into segments
Viterbi search for 1−best dialogue act sequence
Unlabeled game segments containing dialogue acts
Enhance segments by game output distribitions
Game segments containing dialogue acts
Game boundaries known
4
5
Viterbi search including a game language model
Dialogue act and game segmentation and labeling

Figure 3.3: **Incremental Lattice Construction:** The input to the dialogue model are words (or a lattice of words), e.g. produced by a speech recognizer. The words are segmented into a segment lattice ① which is annotated with weights for the dialogue act types ②. This lattice can be searched for the best dialogue act sequences ③ or used for dialogue game detection. Similarly a dialogue game lattice is constructed ④ and the best combined dialogue act and game sequence can be extracted ⑤. A similar picture excluding the dialogye game level has been presented in Ries (1999a).

this work. The use of a segment lattice also allows for a simple integration of discriminative models at the segment level since the A^* algorithm creates only a small number of segments which can be passed to a classifier explicitly.

Dialogue games are a temporally ordered sequences of dialogue acts and the algorithmic framework presented is able to optimize dialogue act and game tagging simultaneously which even improves the detection of dialogue acts (Tab. 3.3, p. 99). However the A^* search is not capable of solving this problem in one pass and an intermediate dialogue act lattice has to be created. The problem is again broken down into three phases after the dialogue act lattice has been created: Segmenting the dialogue acts (the lattice of dialogue acts) into dialogue games (Fig. 3.3, ④), annotating it with dialogue game labels and searching over it with a dialogue game language model (Fig. 3.3, ⑤). Basically the process of dialogue game lattice construction (Fig. 3.3, ④,⑤) is almost identical to the dialogue act lattice creation. The only real difference is that the segmentation into dialogue acts was only on one channel whereas the segmentation into dialogue games requires picking elements of both channels.

3.6 Experiments in Dialogue Act and Game Tagging

3.6.1 Introduction

The input of a dialogue act classifier that is based on language models is a sequence of words (Sec. 3.2.3). Klaus Zechner in our group has provided a part of speech (POS) tagger based on Brill (1994b)'s approach and software for CallHome Spanish and Switchboard. The input string is first processed by the POS tagger and the word/POS pairs that are not in a vocabulary are mapped onto their part of speech. If the part of speech is not in the vocabulary the token is mapped to an unknown word token (Sec. 4.2.4). The vocabulary consists of all part of speech plus the most frequent word/POS pairs and it was shown in Finke et al. (1998) that only a small number of word/POS pairs are necessary. This preprocessing step ensures that information that is reliable to estimate is maintained in the input and that important syntactic information is included [15]. This input representation was also successful and compared well to other approaches in Marcus Munk's masters thesis work (Munk, 1999) which used the dialogue act and game classifiers introduced here to built a stochastic parser for a limited domain speech translation task (Sec. 3.9). Specifically, semantic classes pertaining to the parsing task have been available and have been compared to the POS based approach but

[15] In preliminary experiments the words or word/part of speech pairs for the vocabulary were selected using their mutual information with the output classes but this approach not able to achieve better results.

Classifier	Classification Accuracy per Dialogue Act in %
baseline	
pick the most likely dialogue act	43.1%
NN, shortcuts, 3 hidden units, length of dialogue act	47.8%
ngram models	
unigram model	73.4%
bigram backoff model	75.0%
trigram backoff model	66.0%
neural networks, no hidden units, shortcuts	
unigram features	76.3%
+ utterance length	76.4%
+ 100 phrases	75.9%
+ a bigram prior from ngram model	76.3%
neural networks, 3 hidden units, shortcuts	
unigram features	76.2%
+ utterance length	76.4%
+ 100 phrases	76.0%

Table 3.1: **Dialogue Act Classification Results:** A simple neural network outperforms the ngram backoff model when the segmentation is given, adding more features or hidden units to the neural network does not increase the performance.

were not able to improve the results. The more general and domain independent POS based approach was therefore selected. Preliminary results on dialogue act tagging based on the technologies described have been published as (Finke et al., 1998; Ries, 1999a).

Another potential class of features for dialogue act classification is prosodic information. The work with prosody in the context of the Switchboard (SWBD) corpus has not been able to improve much over word based information (Shriberg et al., 1998; Stolcke et al., 2000). Shriberg et al. (2000) however have shown potential for prosody in many related applications such as sentence boundary detection. One may assume that either dialogue act classification is too difficult for prosodic modeling, or that the acoustic conditions in SWBD were in fact too poor or were overlayed with other information to make reliable prosodic feature extraction impossible.

The application of prosodic information is interesting for a practical system implementation if it can be used reliably to estimate higher level information. Prosodic information and classifiers based on it tend to be simple in structure and

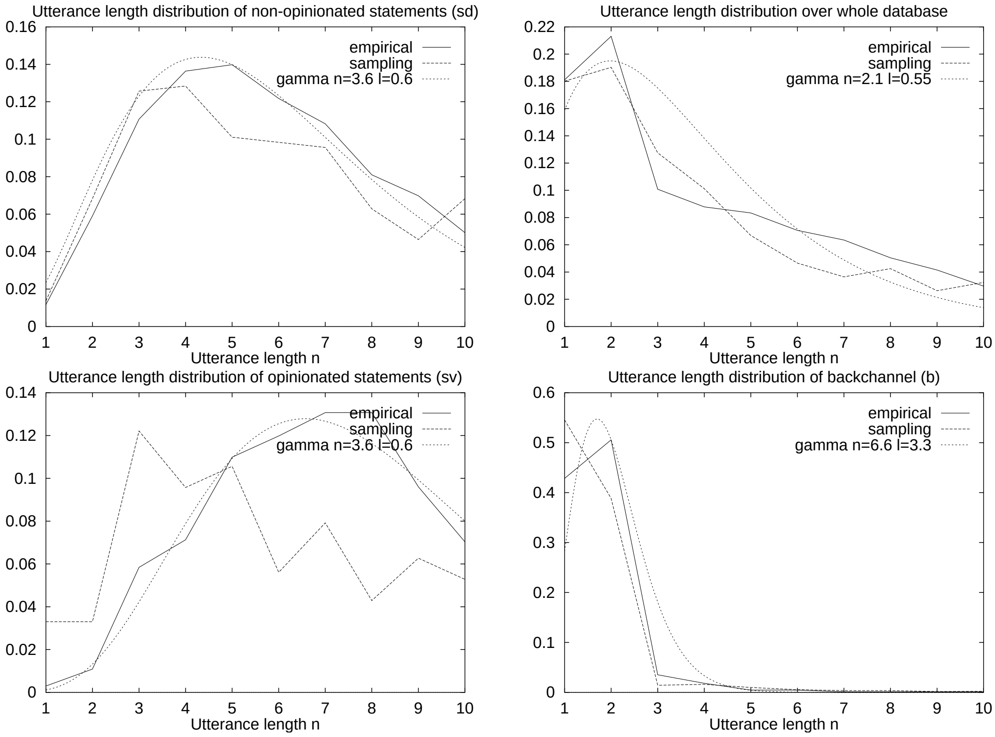


Figure 3.4: **Length Distribution of Dialog Acts:** The length distribution of dialogue acts as generated by the language models (determined by sampling) follow the distribution of the empirical distribution as shown for the whole database, non-opinionated statements, opinion statements and backchannels. The empirical distributions seem to follow gamma-distributions as illustrated by the gamma-fits to the distributions above (reproduced from Finke et al. (1998))

.

the feature extraction could also be done with little computational burden such that systems may be fielded on small devices with low memory and processing capabilities. Therefore prosody might actually play an important role in the design of practical applications where compromises between the computational demands, the complexity of the solution, and the accuracy need to be made. The approach to dialogue act detection using neural networks could also offer the potential to integrate prosody directly and weigh the contributions in a natural way, especially since neural networks have been tested successfully by the author on the task of dialogue act detection from prosodic features (Shriberg et al., 1998; Stolcke et al., 2000).

In either case, we believe that this clearly speaks against the application of prosody to the task of dialogue act detection in CallHome, since the acoustic conditions on this corpus are worse than on SWBD, the speech is more spontaneous and the discourse phenomena may be more complex. Furthermore the improvements seem to be small to begin with such that it will not be attempted here.

3.6.2 Dialogue Act Classification

The pure classification performance for dialogue act detectors without context is presented in Tab. 3.1. It can be seen that while the standard ngram backoff model approach delivers acceptable results, the neural network based method can improve over it while relies only on unigram features. This seems to underline the hypothesis of Sec. 3.2.4 that the local correlation of unigram features, rather than the lack of ngram information, is a problem for the unigram language model approach. Furthermore it can be observed that the introduction of hidden units has no positive effect for the neural network classifier and that the additions of phrases wasn't very successful either. The utterance length alone delivers results better than the baseline but the results using word level information are significantly better. Fig. 3.4 also reemphasizes that the utterance length is modeled reasonably well by ngram language models which verifies that these models have no fundamental deficiencies for segmentation. One can also observe that the length distributions for the different types are fairly different. Tab. 3.3 presents results using context dependent modeling and integrates the segmentation into the evaluation. Similarly the neural network based approach outperforms the ngram modeling approach and the bigram dialogue model does not have a positive impact. Also it should be noted that the alternative segmentation models did not have a great impact. The standard context modeling does not improve the performance, on the other hand the joint detection and segmentation into dialogue acts and games had a significant positive impact on the results.

Overall the results seem to be that using discriminative training of the language model classifiers is important and that once discriminative training is applied neither context modeling nor higher order models seem to be important. The reason may be that local correlations are captured by generative bigram models which represent violations of the independence assumption in the generative unigram model (Sec. 3.2.4). Most interestingly the results on the combination of dialogue act and game tagging indicate that there are significant improvements in detection accuracy if the two are combine.

Dialogue Act	Count	Description
s	19016	Statement
b	8134	Backchannel (aha, yeah!, is that so, . . .)
x	2392	Human Noise
%	2215	Incomplete
qy	1811	Yes/No question
qw	1087	WH-question
atd	1018	Attention directive
ca	1018	Control acts: Reference to future actions (e.g. a commitment)
fe	929	Exclamation
ny	923	Answer: Yes, uh-huh
p	790	Verbal pause (Let’s see, . . .)
qo	540	Open ended question
aa	451	Accept / Believe (Sure, I’ll do that)
qyˆd	448	Echo question in statement form (So the mail has not come yet ?)
na	437	Answer: Descriptive affirmative
b+	406	Backchannel with positive emotional value
br	380	Repetition request
nn	326	Answer: No, hu-huh
bh	283	Verification request
bˆm	222	Mimic of other speaker as backchannel
qr	140	Alternative ’Or’ question
ng	138	Answer: descriptive negative
qh	137	Rhetorical question
fp	136	Open conversation
b-	127	Backchannel with negative emotional value
ˆq	124	Direct quotation
h	99	Dedge
fc	90	Closing conversation
ar	87	Answer: Reject or Disbelieve
ft	74	Thanking hearer
b!	47	Backchannel with surprise value
no	43	Answer: Don’t know
qwˆd	34	Question marker as object holder (15% of what?)
ff	25	Formulaic forward function (e.g. well-wishes)
ˆg	22	Statement with a tag question
l	16	Link (sooooo, AAAAANYYYYway)
nd	11	Descriptive no
fa	8	Apologizing
bc	4	Correction by hearer

Table 3.2: **Dialogue Act Distribution:** Dialogue acts are annotated according to Thymé-Gobbel et al. (2001) and the corpus is published (Waibel et al., 2001b). The distribution displayed above corresponds to the clustering done for this work. In general the effect of clustering dialogue acts together is rather limited since the distribution is so skewed that the effect can be neglected.

		Classification Accuracy per Word in %	
Classifier	Segmentation	Dialogue Model	
		Unigram	Bigram
baselines			
pick the most likely dialogue act		67.7%	
ngram models			
unigram model	implicit	71.5%	70.2%
bigram backoff model	implicit	73.1%	73.1%
neural networks, shortcuts, no hidden units			
unigram features	p_{all}	73.9%	72.8%
+ utterance length	p_{all}	73.9%	72.7%
+ 100 phrases	p_{all}	73.8%	72.9%
unigram features	non-viterbi	73.9%	73.0%
unigram features	viterbi	72.2%	72.4%
Combined game neural-net dialogue act/game detection		78.4%	
neural networks, shortcuts, 3 hidden units			
unigram features	p_{all}	73.4%	72.5%
+ times	p_{all}	73.5%	72.4%

Table 3.3: **Dialogue Act Classification Including Segmentation:** The results including segmentation and the modeling of context mirror the results in Tab. 3.1. The use of alternative segmentation methods does not have an impact on the accuracy but the combined dialogue act and game detection is very successful. The classification results are per word, annotating each word with the dialogue act it belongs to.

Game	Count	Description
info	6001	Speaker provides information to hearer (exluding label elab).
quest	3790	Acquire information from hearer.
opinion	1044	Provide opinion.
elab	1044	Speaker provides information to hearer but elaborates beyond what was anticipated by the hearer.
direct	926	Give directive for action.
express	774	Speaker expresses her/his personal state.
filler	330	No furthering of dialogue; thinking out loud, self talk.
seek	222	Seek confirmation about a previous statement.
commit	130	Speaker commits to future action.

Table 3.4: **Dialogue Game Distribution:** Dialogue games are annotated according to Thymé-Gobbel et al. (2001) and the corpus is published (Waibel et al., 2001b). For automatic dialogue game annotation only the major categories and the frequent minor category **elab** are used.

3.6.3 Dialogue Game Classification

A dialogue game classifier was constructed using neural networks and language model classifiers. Tab. 3.5 presents the results for a classifier where the dialogue game boundaries are known, Tab. 3.6 includes the segmentation: Language models deliver decent results when applied to the dialogue act sequence but deliver only baseline performance when applied to the word level. The neural networks outperform the language models and are able to extract information directly from the word level as well. This is a clear proof of the power of the discriminative training procedure (Sec. 3.2.4). However dialogue act information is still more important and ngram information is not useful for the word level. As can been seen it is important to swap the channels such that one defined channel is always the more active one (channels normalized). Higher order models of dialogue game sequences however have no positive impact.

In Tab. 3.7 the dialogue act and game detection are integrated representing the full model shown in Fig. 3.3. As presented in the preceding Sec. 3.6.2 and in Tab. 3.3 dialogue act detection profits significantly from the integration. There is however a significant change for classifiers that work best in the integrated framework: The best dialogue game classifier is the language model classifier. This may be due to the fact that uncertainty is introduced by the dialogue act classifier. Indeed a combined classifier that features language models for both levels is worse but not a lot. The combined classifier using only language models

	Classification Accuracy per Dialogue Game in %	
Classifier	dialogue model	
	unigram	bigram
baseline		
pick the most likely game	41.3	
dialogue act ngram models		
unigram model	71.2	72.0
bigram backoff model	74.5	75.6
trigram backoff model	69.4	69.7
word ngram models		
unigram model	41.3	
bigram backoff model	41.3	
trigram backoff model	41.3	
neural networks with shortcuts and no hidden units		
word unigram features	59.3	57.4
+ 200 word ngrams	58.9	57.6
dialogue act unigram features	75.1	74.1
+ word unigram feature	74.7	74.7
+ 200 word ngrams	74.3	74.8
neural networks with shortcuts and no hidden units channels normalized		
word unigram features	59.2	57.4
+ 200 word ngrams	58.9	57.6
dialogue act unigram features	74.4	74.4
+ word unigram feature	74.9	75.2
+ 200 word ngrams	74.7	75.3

Table 3.5: **Dialogue Game Classification:** A simple neural network outperforms the ngram backoff model for dialogue game classification. Here the segmentation into dialogue games and the dialogue acts labels are manually annotated and the result is calculated for each dialogue game. If the two channels are separated into a more active and less active channel and word and dialogue act features are integrated the performance is improved. Most notably detection from the word level was infeasible using ngram models but reasonable using the direct classification approach.

	Classification Accuracy per Dialogue Act in %	
Classifier	dialogue model	
	unigram	bigram
baseline		
pick the most likely game	47.0	
dialogue act ngram models		
unigram model	60.1	57.0
bigram backoff model	60.9	60.6
trigram backoff model	57.7	58.2
neural networks with shortcuts and no hidden units		
word unigram features	54.1	52.8
+ 200 word ngrams	53.7	52.5
dialogue act unigram features	61.5	60.1
+ word unigram feature	61.5	60.4
+ 200 word ngrams	61.3	60.1
neural networks with shortcuts and no hidden units channels normalized		
word unigram features	54.1	52.8
+ 200 word ngrams	53.6	52.5
dialogue act unigram features	62.6	60.9
+ word unigram feature	62.2	61.1
+ 200 word ngrams	60.9	60.4

Table 3.6: **Dialogue Game Classification:** The segmentation of a dialogue act sequence into dialogue games and the annotation of the segments with dialogue game labels is done automatically. The dialogue act annotation however has been done manually. The results are similar to Tab. 3.5.

	Classification Accuracy per Word in %	
Classifier	dialogue model	
	unigram	bigram
baseline		
pick the most likely game	52.0	
dialogue act ngram models		
bigram, dialogue acts neural network	56.6	55.6
bigram, dialogue acts bigram model	55.8	54.5
dialogue act ngram models trained on dialogue act detected data		
bigram, dialogue acts neural network	55.5	55.5
bigram, dialogue acts bigram model	55.2	55.1
neural networks with shortcuts and no hidden units dialogue acts detected using neural net		
word unigram features	55.4	53.4
+ 200 word ngrams	55.0	54.1
dialogue act unigram features	55.8	53.6
+ word unigram feature	54.7	54.2
+ 200 word ngrams	54.0	53.7
neural networks with shortcuts and no hidden units channels normalized dialogue acts detected using neural net		
word unigram features	55.4	53.4
+ 200 word ngrams	55.1	54.1
dialogue act unigram features	56.0	54.9
+ word unigram feature	55.7	55.1
+ 200 word ngrams	54.4	54.0
neural networks with shortcuts and no hidden units channels normalized dialogue acts detected using neural net trained on dialogue act detected data		
dialogue act unigram features	55.8	55.5

Table 3.7: **Integrated Dialogue Act and Game Classification:** The dialogue act and game classification have been integrated using the incremental lattice generation approach (Fig. 3.3). The procedure therefore gets the words on different channels as its input, constructs a dialogue act lattice and finds the dialogue games from there.

is used in the experiments on activity detection in Sec. 4.5.4 since it is simpler to reproduce and computationally more efficient.

3.7 Combination with Speech Recognition

In project Clarity we have not integrated dialogue processing with speech recognition as this has been done in other contexts, e.g., the John Hopkins Summer Workshop of 97 with participation of the author. Stolcke et al. (2000) present a detailed description of the techniques that are necessary and effective in the integration with a speech recognition system which apply largely to the Clarity problem as well. Specifically it is very important to use the lattice output of a speech recognizer and integrate the classifiers over the whole lattice. In a nutshell instead of estimating $p_{l_i}(y_i)$ we try to estimate

$$p_{l_i}(a_i) = \sum_{y_i} p_{l_i}(y_i) \cdot p(a_i|y_i)$$

where a_i is the acoustic evidence used by the speech recognizer. The sum can be approximated by summing over an n-best list or hypothesis lattice of the speech recognizer assuming that most of the unlikely hypotheses contribute little to the sum. A number of further important technical details are discussed in Stolcke et al. (2000).

Stolcke et al. (2000) prove that the resulting algorithm is fairly effective and the degradation due to recognition errors can be reduced to a small factor. Another observation is that the standard language model algorithm can be used very effectively in a search over a lattice and one forward pass for each one of the dialogue acts types l yields an approximation of $p_l(a_i)$. For the neural networks the situations would be more complex, however if they can be reformulated as language models or at least single term representation models (Sec. 3.2.5.4) the same procedure could be used. Otherwise the lattice would have to be decomposed in a more complex way.

It should be noted that in Stolcke et al. (2000), like other published work, the segment boundaries should be given beforehand. The resulting optimization formula that integrates the acoustic information would become a lot more complicated to solve. Although this is an interesting problem out of its own right it wasn't attacked in this thesis in order to make progress on higher level structures and information access. The algorithm presented here would be able to work from a word lattice however the summation of acoustical evidence over all possible segmentations is not implemented. In similar vein – although it might be important for commercial systems – prosody only detection of dialogue acts was not explored which could have resulted in low footprint detection algorithms.

3.8 Emotion Detection

Emotions are displayed in a variety of gestures, some of which are oral and can be detected via automated methods from the audio channel (Polzin, 1999). Using only the information whether laughter was transcribed in the utterances the emotions *happy*, *excited* and *neutral* can be detected on the meeting database 88.1% correct (actually it was basically predicting *happy* correctly), while always picking the neutral emotion yields 83.6%. This result can be improved to 88.6% by adding pitch and power information. Clustering the histograms of emotions over topical segments categories which could be coined *neutral*, *a little happy* and *somewhat excited* evolve. Using the histogram generated by detected emotions a classifier on the topic level can be obtained which is 83.3% correct while the baseline is 68.9%. The entropy reduction by emotional activities is 1.3bit of which 0.3 can used by automatic detection.

The results are compared to the Woggles database (Tab. 3.8, Polzin (1999)) which shows that the prosodic module and classifiers are operating in a desirable accuracy range. The database was accurately reconstructed and consists of emotions enacted by actors. The spectral features of (Polzin, 1999) are derived by adapting speech recognizers to the emotions, which wasn't done since it is a computationally intensive task. All other features used in this work seem to be as good or better than Polzin (1999) however the F_0 normalization of Polzin (1999) features much better results (Exp. 2 in 3.8) which is also reflected in the overall performance. However normalizations are often very domain dependent such that it is likely not portable; it should specifically be noted that the last row presents a classifier that is based only information local to the segment with no normalization and should therefore be easy to port to new domains. Overall both classifiers seem to operate in the same accuracy range.

3.9 Application to Parsing

A machine learning approach to dialogue modeling for conversational speech has been presented which annotated segments of words with labels. It should therefore be applicable to similarly structured domains. Marcus Munk used some of the software and methods presented in this thesis in his masters thesis project and has applied it to parsing and translation in restricted speech domains (Munk, 1999). The approach was to learn the high level parse structure from bracketed data and have a parser analyze the remainder in the leaves. The high level structure was captured in two levels, one corresponding to a "speech act" according to the interchange format and the lower one corresponding to the "arguments" of the interlingua. Due to time constraints the direct classification approach was not

Features	Experiment	Detection Accuracy in %	
		Polzin (1999)	Neural Network
Baseline		25.0	
Spectral	1	68.8	
F_0, mean/variance	2	55.8	43.3
F_0, jitter	3	40.4	41.7
Intensity, mean/variance	4	33.0	43.3
Intensity, tremor	5	45.4	45.8
Duration phonemes	6	42.0	
All prosody (2,3,4,5,6)	7	60.4	56.3
tremor, jitter context-independent			55.0
Human		69.0	

Table 3.8: **Emotion Detection Results on Woggles:** The table corresponds to the experiments in (Polzin, 1999, p. 58) but it is using the detection accuracy measure instead of F-scores. The experiment column refers to the feature sets reported there and approximations of those have been attempted for the neural network based classifier. The last experiment refers to a portable featureset that does not need a normalization over multiple utterances of the same speaker.

applied; only ngram backoff language models were used. The speech acts are usually complex and only the main speech act was detected directly by an architecture that corresponds to the game detector described above. The "concepts" that are attached to the speech act are detected separately using a neural network that has access to the speech act as detected, the arguments, and the word level information. The argument slots structurally correspond to the dialogue act component of the combined dialogue game detector. The argument slots are – after being bracketed and annotated – passed to the parser with the restriction that they have to be generated using the argument that the stochastic approach provides. A separate module assembles a parse that looks like as if it came from the parser and passes it through the various components of the speech translation system developed in our working group (Woszczyna et al., 1998).

It should be mentioned that the bracketed data was generated by running a parser with the very same grammar that was used as a comparison such that the advantages of the learning approach could not be fully exploited. The results were slightly above a baseline performance but it should be noted that the parser that was used to create the bracketed data was doing very poorly on that database as well. Very small improvements could be demonstrated by adopting a multi-engine approach interpolating the stochastic engine with the original parser (Munk, 1999). The results are therefore overall inconclusive but the approach is still seen as very promising within our group. Similar work was reported by Buø (1996); Buø and Waibel (1996); Wermter and Weber (1997) and related ideas are used in current projects.

3.10 Conclusion

This chapter has presented dialogue act and game detection technology. A multilevel search procedure was proposed and tested which enables to model dialogue acts and games simultaneously and is able to handle overlap of speakers and dialogue games. A discriminative training approach for the dialogue act and game detection based on the embedding in exponential models and neural networks was tested successfully and the integrated detection of dialogue acts and games improves the dialogue act detection significantly. The best results were achieved by using a neural network for dialogue act detection and a language model for dialogue game detection. The same methods and implementations have also been applied successfully to parsing for speech to speech translation and emotion detection. Dialogue act and game detectors are used to facilitate the detection of activities in the next chapter (Sec. 4) and are useful for visualization purposes (Sec. 6.4) including the user study (Sec. 6.3).

Chapter 4

Database and Activity Detection

4.1 Introduction

There are plenty of microlevel features that describe the register of a conversation or the activity of a topical segment (Sec. 2.4.2). Which user of an information retrieval system, however, would want to enter part of speech distributions? And which user remembers such distributions of meetings he or she attended half a year or just half a day ago? Or which speaker has an intuitive understanding of stylistic feature dimensions, probably developed by a principle component analysis [1]?

This chapter is concerned with the detection of high-level labels for dialogues and topics in dialogues that are understandable for users of information access systems and that are known to be remembered (Sec. 2.4.4). The derivation of types is important since types appeal to users which can be even seen at the otherwise difficult activity level (Sec. 6.3). The detection is achieved by using neural networks (Sec. 3.3) which map microlevel features (Sec. 4.2) on the high-level labels.

Several different types of high-level labels have been investigated and their selection in part reflects the relevant databases we were able to obtain. It also represents labels for different levels in the information access hierarchy (Fig. 1.1).

databases One may assume that there are a number of different databases such as “Broadcast News”, “private telephone calls” and “other phone calls”. This kind of information is typically called register and it turns out that the automatic detection is very easy (Sec. 4.3).

sub-databases Depending on the database there might be a number of sub-databases available and indeed those sub-databases can be distinguished rea-

[1] Actually it is very unlikely that users understand these dimensions at all. In the user study (Sec. 6.3) it can be seen that formality, which is a very important dimension, is not very useful.

sonably well as can be seen in the TV-showtype task (Sec. 4.4).

activities Within one topic of a conversation people likely follow one activity: They are telling stories, discussing, planning and so forth. However, human activity judgments are rather inconsistent as it turns out and it is even more difficult for the machine learning approach to replicate those human judgments. Nevertheless the classifiers show performance significantly over the baseline (Sec. 4.5).

The overall conclusion of this chapter (Sec. 4.6) is that the detection of database and sub-database information is fairly easy while it is surprisingly difficult to reach agreement between humans on activity labels. However the detection algorithms still deliver a reasonable performance for activity detection. Users should certainly be able to make use of database and sub-database information to constrain the search for information since the indices are straightforward. Indeed it is shown in a user study that activities are a very good tool to discriminate between different rejoinders (Sec. 6.3).

4.2 Microlevel Features

4.2.1 Introduction

This section describes the features used for (sub-)database and activity detection in more detail. While features were already introduced in Sec. 2.4.2.3 this presentation describes them in a more implementation oriented fashion. These features might not be the only ones to encode style but they are the ones used in the experiments. Since the feature sets are similar for all experiments they will be introduced at once.

The immediately following Sec. 4.2.2 describes some pitfalls to be avoided while the remainder of the section is devoted to individual feature sets. In Sec. 4.2.3 very basic interactional features are presented, Sec. 4.2.4 describes different word and wordclass based features and Sec. 4.2.5 describes a set of syntactic features. Dialogue acts and games can be used as features (Sec. 4.2.6) as well as speaker dominance (Sec. 4.2.7).

4.2.2 Safeguards

Numerous features could be used for activity classification. One has to be cautious about unexpected correlations in the database. If every conversation with `Mr. Schmidt` is about the `German Reinheitsgebot` and is an informing activity one could infer both from the keywords `German` and `Reinheits-`

Name	Description	Example	Application	Section
Features of a Spoken Documents				
Database	The set of databases depends on the application.	Broadcast-news, personal phone calls, phone calls between strangers.	Easy to remember and understand by humans, easy to detect automatically since almost all lower level features differ significantly depending on the database.	4.3
Aggregates	Histograms of features from the topical segment or utterance level as well as microlevel features. Change significantly when the database changes.	A formality score for a document, a histogram of frequent words and parts of speech.	Input to database identification, maybe visualized or soundified via examples.	4.3
Subdatabase	Finer distinction than database types such as TV show types.	Talk-shows, Movies, Newscasts, etc.	Is likely undestood and known. relatively easy to detect.	4.4
Features of a Topical Segments				
Activity	The type or style of an interaction.	Storytelling, Discussing, Planning, etc.	Is likely remembered by participants, can be detected, successful in user study.	4.5, 6.3
Aggregates	Histograms of microlevel and utterance features.	Emotions, dialogue acts, interactional, participants, etc.	Input to detectors, can be visualized.	4.5, 6.3.2
Conversational Dominance	Average dominance rating of a person, based on dialogue act types Linell et al. (1988).	Very dominant, not dominant at all.	Input to activity classifier, implicit in other features visualized.	4.2.7
Formality	Can be defined formally using parts of speech.	Spoken interactions are fairly informal, written one are more formal.	Use is unclear, was unsuccessful in user study.	2.4.2.3, 6.3
Features of an Utterances				
Dialogue Act	Dialogue function of an utterance.	Question, answer, backchannel, statement, etc.	Can be detected automatically and is a good abstraction for activities. Visualization possible and specific intonations (e.g. questions) may be useful in fast playback. May be useful in summarization.	3, 4.2.6, 6.3.2
Dialogue Games	Short “typical” sequences of dialogue acts.	Question-answer pairs	Similar to dialogue acts.	3, 4.2.6
Speaker Identity	Name of the speaker of the utterance, can be detected automatically.	Henry, Jeff, Klaus, Michael	Useful for topic segmentation, likely useful for information access but unsuccessful in user study.	2.4.5.4, 6.3.2
Emotion	Displayed emotion of a speaker.	Happy, sad, angry, neutral.	Hard to detect in meetings.	3.8
Features of Words / Microlevel features				
Frequent words & Parts of Speech	The most frequent words (often ≈100) and parts of speech (Verb, Noun, Pronoun).	“I”, “want”, “go”, Adverb, Verb	Histograms of frequent words and parts of speech are style and dialog act discrimators, but the raw histograms are “meaningless” to a human.	4.2.4, 3.6
Semantic fields	Categorization of nouns and verbs, assumed to be corrolated with social roles.	“Verb Motion” “Noun Bodypart”	Input to activities, potentially an index by itself.	4.2.4
Syntactic features	Collection of features, especially syntactic constructs characterizable by regular expressions are implemented.	Suasive verbs, WH-questions, . . . (Biber cites about 80 features)	Input for activity detection.	4.2.5, 4.5, 4.3
Keywords	Rare words which index topic.	Spoken interactions may not be indexed as well using keywords as written language.	The words that you would type into Google.	1.2
Durations	Lengths of items.	Lenghts of word, syllables, etc.	Useful in histograms for discriminating databases.	4.3
Prosody	Suprasegmental features of language.	Pitch and power statistics, especially (normalized) histograms.	Histograms, dialog act and emotion detection.	3.8, 3.6, 5.6.6

Table 4.1: **Overview of Feature Categories:** Features are detectable properties of spoken communication. The table gives an overview of the features which are detected or used in this work. It makes obevious that many features can be described but only few are likely useful to an information seeker.

`gebot` as well from the speaker identity that an informing activity is likely. Since the stylistic features at the word and syntactic levels may be very indicative of the speaker identity (Sec. 2.4.5.4, Sec. 6.2.4, Sec. 6.2.4) this information may be even available if the speaker identity is not explicitly given but just a parts of speech distribution. Similarly, the correlation of activities with topic might be more subtle as for example females might discuss different topics with other females than with males and might use different terms (for example females often refer to their spouses as their "husband") (see Tab. 4.4). The author assumes that dialogue act or dominance distributions are less prone to such indirect effects.

This short discussion illustrates that the experimental design has to pay attention to avoid these problems. An important safeguard is that the training- and test-sets should never contain segments from the same dialogue for any of the experiments considered here [2]: This avoids some simple problems, if a certain activity is highly correlated with a conversation which could entail unwanted correlations of the activity with the speaker identity and the topic. A second safeguard is to avoid the use of too many different word types in the classification since this might induce correlation with topic [3]. A third safeguard is the manual inspection of the features if unwanted correlations are suspected. This has been the case for the gender discrimination task (Sec. 4.3) and led to the exclusion of certain "family affairs" keywords.

4.2.3 Interactional Features

Some simple measures indicative of the dialogue type can be derived if timing information is available about which person is speaking when. The detection of the timing information may be fairly trivial if separate channels for each speaker are available and there is little cross-talk or background noise. The idea is to identify *active segments* where one speaker is staying active, often called a speaker turn. We define *active segments* as the longest segments for each speaker which don't contain pauses longer than 0.3 sec and where each pause is surrounded by segments that are at least 4 times longer than the pause (the last provision excludes very short elements that would extend the active turn). The length of the *active segments* provides an idea how much uninterrupted and single sided speech is present in the conversation. The distribution of *active segment* lengths may also have a similar function as the distribution of dialogue acts since short active segments are likely backchannels, long ones are often statements and other are

[2] For the detection of meeting identity segments from the same dialogue obviously have to be in the training as well as in the testset (Sec. 6).

[3] Indeed the number of word types is varied for the TV show genre classification to show that the addition of topic-like features does not change the classification results by too much in that task.

usually some other dialogue act. Another aspect might be captured by a measure that quantifies how much the speakers overlap. We distinguish two types of overlap: If speaker A is active and speaker B is overlapping but the segment of B is wholly contained in A's segment it is a *coverlap*, otherwise it is a real *overlap*. This distinction might be important since coverlaps are likely backchannels by one speaker while the other one is making a long statement or telling a story. The function is therefore supportive to the speaker that is holding the other channel while a real overlap can indicate that speakers are competing over the channel.

Active segments, *overlap* and *coverlap* are lengths of segments and they are presented to the classifier as histograms. The histogram is typically calculated over 10 buckets and the bucket boundaries are set to ensure that each bucket is equally likely. This entails that these features need to be trained.

4.2.4 Words and Wordclasses

The most straightforward feature of dialogue style might be word level information that could be available as a manual or machine transcript. Since we do not want too many actual words to restrict the size of the feature space and avoid the influence of topicality the same processing is applied as in chapter 3: The text is tagged with parts of speech [4] and the most frequent n word / part of speech pairs are used directly, all other word / part of speech pairs are mapped onto their parts of speech. A typical size for n is 50 which ensures that many closed class words are covered, the tokens are frequent and that topical factors play no role. It also ensures that special tokens that are present in the transcript such as human noises and laughters are taken into account. Word level information including parts of speech can be represented as a histogram over the frequent words and the parts of speech. It should be noted that the definition of formality in Sec. 2.4.2.3 is based on a linear combination of parts of speech frequencies. Since the neural network based classifier is able to calculate linear combinations of input features the addition of formality to the feature set would be redundant.

"Semantic fields" are another type of information which might be available at the word level: If we have a categorization of words (or phrases) into semantic fields we could derive a histogram representation of the "semantic field" of a text. McTavish et al. (1995) created the MCCA toolkit to measure social features and social distances based on the distance of semantic fields. However his database was not available for our research. The "unique beginners" of Fellbaum (1998)'s WordNet are an attempt to partition verb and noun meanings into 40 disjunct

[4] Klaus Zechner trained an English part of speech tagger tagger on Switchboard that has been used as well as a Spanish part-of-speech tagger. The tagger uses the code by Brill (1994a).

word classes	
nouns	verbs
act animal artifact attribute body cognition communication event feeling food group location motive object person phenomenon plant possession process quantity relation shape state substance time	body change cognition communication competition consumption contact creation emotion motion perception possession social stative weather

Table 4.2: **Verb and Noun Classes:** WordNet provides a total of 40 verb and noun classes (Fellbaum, 1998). Each word meaning is associated with a class. A full-form word is mapped to a class meaning by picking the most frequent verb or noun class among all senses of the full-form word. (reproduced from Ries (1999b))

classes (Fig. 4.2) [5], seem to be a proper substitution and histograms over “unique beginners” are used as the “WordNet features”.

Kaufer and Butler (2000) and Kaufer represent another candidate for a word/ phrasal feature. Kaufer’s “docuscope” tool is used in his creative writing classes at Carnegie Mellon University and it contains a large collection of phrases that are tied to certain text functions. Students are taught that they need to employ the right text function for a specific part of a text. The use of phrases allows to disambiguate between different meanings of common verbs and nouns which frequently appear in common phrases. In that sense they might provide a functional description of the dialogue. Initial experiments with features derived by the docuscope tool were inconclusive such that this option was no longer explored. It may be that the functional categories – although they seem to occur in our texts – were just not optimally designed for our discrimination tasks.

The word length in and by itself might already be a good features since it is likely an indicator of the part of speech of a word (short words tend to be function words, long ones tend to be nouns). Word length histograms might also be available even if speech recognition would be really bad such that it is an interesting feature.

[5] It should be noted here that these are not the standard WordNet semantic categorizations. Rather any synset in WordNet is classified as belonging to one of these categories. Synsets are elementary semantic categories in WordNet.

English	Spanish
Tense and Aspect Markers	
PastParticiple (past tense) Have + Adverb* + PastParticiple VerbPresentTense	VerbSubjunctive Habere + Adverb* + PastParticiple VerbIndicative
Place and Time Adverbials (examples)	
aboard above abroad across ahead alongside afterwards again earlier early eventually formerly	aqui’ ahi’ alli’ aca’ alla’ dentro ya recie’n ahora antes anteriormente despue’s en
Pronouns and Proverbs	
First,Second,Third Person, Impersonal Pronouns	
	Familiar, personal, demonstrative, indefinite
WH-questions	
	BeginOfSent + (que’ cua’ndo co’mo quie’n do’nde cuanto cuantos)
Nominal forms	
Nominalization suffixes, all other nouns	
	Infinives used as nouns
Passives	
Be+ Something + PastParticiple	similar
Stative Forms	
BeAsAMainVerb, existential there	hay (without que), complex form
Subordination	
Adjective+That, infinitives, that-relative clause, preposition+Wh clear and potential subordinators	similar rules
Adjectives and Adverbs	
Prepositional Phrases, Attributive Adjectives	
	Predicative Adjectives, Adverbs
Specialized Lexcial Classes	
Conjuncts, Downtoners, Hedges, Amplifiers, Emphatics, Discourse Paricles, Demonstratives	
Public, Private, Suasive Verbs	Augmentive, Dimunitive and Perjotive Suffixes
Modals	
can, ought, shall	
Reduced or Dispreferred Forms	
Contractions (’ve ’d ’re), Subordinator That Deletion, Stranded Prepositions, Split Infinitives and Auxiliaries	Split Auxiliary, Estar+Prespart
Coordination	
	Phrasal coordination (patterns containing conjunctions)
Negation	
Not or Other Negation	Many Negation Words
Filled Pause	
Some words	

Table 4.3: **Biber Features:** The features listed in Biber (1988) which were easy to replicate have been implemented. The implementation relies mostly on pattern matching of part-of-speech tagged input. Lexicial specifity is also calculated and measures the type token ratio and the length of the words in the text.

4.2.5 Syntactic Features

Style and register have been studied by grammarians in the past. Some accessible, highly computational and frequently cited work is Biber (1988) [6] Not only did Doug Biber provide interesting insights in the correlations of features by providing factor analysis, he has also provided fairly accurate instructions how to annotate these basic features. Most recently Biber et al. (1999) presents an encyclopedic account of a grammar that features a large number of register dependent usages, however Biber (1988) remains the most operational basis. Out of the original 67 features 46 have been selected and implemented using regular expressions based on part of speech detection, a simple morphological analysis and the wordclasses from Quirk et al. (1985). Spanish features have been developed to correspond with Biber (1988) and for each of his category some corresponding features have been suggested. Spanish also offers many forms of social hints such as dimunitive suffixes which also appear frequently (see also Kattán-Ibarra (1991)). [7] The features used for English and Spanish are presented in Tab. 4.3 and they were presented to the classifier in the form of histograms. The features are called “Biber” or “SBiber” in their English or Spanish forms. The table features a rough look at the different feature classes and which features have been implemented. The implementation is mostly based on simple regular expressions of parts-of-speech annotated text – they therefore represent heuristics. The Spanish features have probably been tested better since they were developed by a linguist.

Additionally shallow grammars for English and Spanish, developed by Klaus Zechner, have been tested. The idea was to use the top level slots of the grammar, the phrase heads and the depth of the tree. However these grammar based features were not successful after initial tests and have not been used in the final experiments.

4.2.6 Dialogue Acts and Games

The detection of dialogue acts and games was done in order to describe dialogue in a more abstract way and to condense important information such that it can be interpreted more easily (Sec. 2.4.3.3). The use of dialogue acts and types of dialogue act sequences for the characterization of larger segments such as activities is a common idea (Carletta et al., 1997; Carlson, 1983; Franke, 1990; Levin and Moore, 1977). Dialogue acts are definitely useful to describe dialogue style since

[6] Among the stylistic features by Biber (1988) were also a small number of word classes similar to the ones discussed in Sec, 4.2.4 such as public/private/suasive verbs (Quirk et al., 1985), amplifiers and downtowners.

[7] The author is indebted for the help of Donna Gates. She has developed the Spanish features in the form of regular expressions after the categories of Biber (1988).

they encode dominance in conversations (Sec. 2.4.2.4, Sec. 4.2.7). Dialogue acts are also useful beyond the definition of dominance since they are often assumed to encode activities (Sec. 2.4.2.2). An obvious example is that questions are very frequent in interrogations however it is not immediately clear if this holds likewise for other types of activities. However some of this information may also be extracted easily from the word level, or the information is available from the segment length itself (backchannels are fairly short compared to statements). The problem with any fixed definition of dialogue acts and games is that they can only capture certain aspects of the dialogue information.

Dialogue games are short sequences of dialogue acts such as question/answer pairs or information giving (statements interspersed with backchannels). The game type is usually determined by the first dialogue act – also called move – in the game. The underlying idea is to focus the attention on the initiating move in a dialogue game and ignore the remaining moves since they don't advance the knowledge state materially. In a question/answer game for example the fact that the question is followed by an answer or that there might be clarification questions would not be represented since only the "question"-game label would remain.

In the English experiments the dialogue acts are detected using a model trained on Switchboard similar to Stolcke et al. (2000). The model was trained to be very portable and therefore the following choices were taken: (a) the dialogue model is context-independent and (b) only the part of speech are taken as the input to the model plus the 50 most likely word/part of speech types. In the Spanish experiments the dialogue acts were either hand annotated or annotated using a combined dialogue act and game detector described in Sec. 3. However – to make the replication easier and use less complicated models – dialogue acts and games were detected using a combined classifier that used only ngram backoff models as classifiers. The results of this classifier are not the very best in the tests but they are very close to the best performing system and the system is simpler overall.

4.2.7 Dominance

The dominance of one speaker over another seems to be a concept that is very important to understand in a conversation. If we, for example, want to describe a dialogue as either a storytelling or a discussion activity it might be useful to know whether there is only one dominant speaker or whether there are several dominant speakers. However dominance may not be appropriately captured via simple measures, such as just counting the length of the contributions only. Linell et al. (1988) associated a dominance value with each dialogue act and have shown that there is a good correlation of the average dominance for each speaker with the actual perceived dominance. In Linell et al. (1988) dialogue act types that restrict the options of the conversation partners have very high dominance (questions);

dialogue acts that make contributions but don't restrict the conversation partner have high dominance (statements); dialogue acts that signal understanding carry low dominance (backchannels). We follow that definition and represent the dominance distribution in a conversation by a the histogram of speaker dominance. It should also be noted that the bucketing of the histogram is learned.

The formula we used used a weight of 5 for questions, 3 for statements, 1 for answers and -2 for backchannels; the dominance was the average dominance per turn. Dominance could therefore also be inferred visually in a graphical representation of the conversation which shows the dialogue act type for each speaker.

4.3 Databases of Documents

4.3.1 Introduction

Two prototypical situations for the discrimination of large dialogue style differences are demonstrated: The first task is a cross-corpus discrimination task in which we included four different spoken language corpora (*cross-corpus task*). The second task is a *within-corpus task* where we try to distinguish speakers with different features (gender) in one corpus, namely Switchboard. The idea is to train a classifier for these tasks and see which features are important for classification performance. This content of this section is reproducing the authors results in Ries (1999b)).

The first task is to discriminate between four spontaneous speech corpora: CallHome English, CallHome Spanish, Broadcast News and Switchboard, which are all published by Linguistic Data Consortium (LDC). All corpora but Broadcast news are telephone conversations and all corpora but CallHome Spanish are in English. Unless we want a trivial 100% result we have to exclude a couple of potential features: We cannot rely directly on word identities (since we have multiple languages and transcription conventions) and on features that would immediately uncover the fact that Broadcast News contains music, a variety of speakers and speaking situations et cetera. The same argument could be made for the fact that CallHome speakers are relatives. The basic result is that even simple features allow almost perfect discrimination. This has two important consequences: (a) if we don't know which database an item belongs to we can make a good guess and (b) even the most simple features will work. Similarly it would a total waste of energy to do any dialogue processing for making this kind of determination, rendering all but the simplest features unimportant. Such high level types also seem to be more reasonable to understand for a user of an information access system than the dimensions of a factor analysis (c.f. Biber (1988)). One may therefore represent documents rather by their membership to databases than by the values husband watch children seen family she T uh-huh her usually Texas care wasn't either Yes kids feel um haven't fact Do We nice ago find

Table 4.4: **Salient Words for Gender Discrimination:** Keywords for talking about family life are very salient for gender discrimination. Salience was determined using Gorin (1995)'s salience measure which is related to mutual information.

of stylistic dimensions found by a factor analysis.

It could be interesting, e.g., to measure what kind of social relationships people have that talk to one another. The distinction "well acquainted"/"not acquainted" was already made in the *cross-corpus task* (CallHome/Switchboard). While we don't have more data the experiment we have decided to use the Switchboard corpus and try to distinguish between male and female speakers since we assume that unacquainted male and female speakers will show stereotypical gender specific discourse behaviors (*within-corpus task*). In Switchboard the topic of the conversation is also given in advance by the system that connects "random" people (see Sec. 1.4.3.3 for more details). However, as can be seen from Tab. 4.4, the participants did not stick to their assigned topics and drifted off to private discussions. Some of the best gender discriminating keywords therefore relate to private "family affair" topics as revealed by the salience analysis for keywords (Gorin (1995), Tab. 4.4).

4.3.2 Experiments

The first set of experiments was carried out with a reduced set of features such that both the cross-corpus task as well as the with within-corpus task could be tested. All experiments used a neural network with shortcut connections and 5 hidden units (Sec. 3.3). The first interesting result from Tab. 4.5 is that the cross-corpus task is extremely simple. If we actually look at the means and variances of these features it becomes immediately obvious that the discrimination should be trivial.

The results for the within-corpus task are a little harder to interpret. Using just the fact that a word was said on one channel versus the other ("only word" in Tab. 4.5) resulted in a good baseline of 66.3%. If we compare this to the results of the pause and length measurements most of them seem to do worse than that, with the exception of the active segment length and interestingly enough the word length. The first suspicion was that the speaking rate is different for male and female speakers but in the database of Shriberg et al. (1998) we have found no significant difference.

The next observation is that the frequent 300 words, from which a lot seem to

	Detection Accuracy	
Feature	cross	within
	corpus task	
baseline	25.0%	50.0%
length of active segments	100.0%	68.8%
length of overlap of active segments	100.0%	61.3%
length of words	95.5%	71.3%
all of the above	97.5%	67.5%
+ most freq. 300 word		83.0%
most freq. 300 word		83.0%
most freq. 300 words (excluding “family affair” terms)		80.0%
40 most salient words		75.0%
20 most salient words		72.5%
10 most salient words		68.8%
only “word”		66.3%
WordNet		65.0%
parts-of-speech histogram		72.5%

Table 4.5: **Database Detection:** The cross corpus task is the discrimination of CallHome English and Spanish, Broadcast News, and Switchboard, a task which turns out to be very easy. The within-corpus task is the discrimination of gender in Switchboard. The word information that was used did not include part-of-speech information, a word that was not in the vocabulary was mapped to a default token. (reproduced from Ries (1999b))

be keying towards family affairs (Tab. 4.4), can give us a good performance. Eliminating the “family affair” words from the list still resulted in good performance. However, restricting the size of the list considerably – even using a salience analysis – did not result in much better performance. We may therefore conclude that the discrimination is done mostly by the non-topical words and that all of these words contribute to the classification.

We have had only limited success using (Biber, 1988) as well as using WordNet features. However, the part of speech distribution is a fairly successful feature. Overall gender specific style difference do exist and manifest themselves in the features presented here. It is therefore likely that other salient social roles may also be detectable if we had training data for them.

	#		#		#
Talk	344	Edu	25	Finance	8
News	217	Scifi	24	Religious	5
Sitcom	97	Series	24	Series-Old	3
Soap	87	Cartoon	23	Infotain	3
Game	46	Movies	22	Music	2
Law	32	Crafts	17	Horror	1
Sports	32	Specials	15		
Drama	31	Comedy	9		

Table 4.6: **TV Show Types:** The distribution of show types in a large database of TV shows (1067 shows) recorded over the period of a couple of months until April 2000 in Pittsburgh, PA

4.4 Sub-database: TV Show Types

While the database detection task was almost trivial the question is whether the detection of sub-databases is just as easy and which features might be important; the results in this subsection are reproduced from Ries et al. (2000). A user may for example not only remember that something was a TV-show but also that it was a game show. This type of distinction may also be available in other databases such as meetings: A board-of-directors meeting is certainly different from an office meeting or from a budget meeting and so forth. However we didn't have data for this kind of distinction such that the TV-show task was created to fill in the gap.

We set up a recording environment for TV shows which continuously records the subtitles with timestamps from one TV channel and the channel is being switched every other day. At the same time the TV program was downloaded from `http://tv.yahoo.com/` to obtain programming information including the type of the show. Yahoo assigns primary and secondary show types and unless the combination of primary/secondary show-type is frequent enough the primary showtype is used (Tab. 4.6). The TV show database has the advantage that we were able to collect a large and varied database with little effort.

The same neural network classifier as in Sec. 4.3 has been used however dialogue acts have not been detected since the data contains a lot of noise, is not necessarily conversational and speaker identities can't be determined easily.

Detection results for TV shows can be seen in Tab. 4.7. Both WordNet categories as well as word level features deliver a solid baseline performance and feature combinations can improve detection results slightly. To measure whether this task is topic sensitive different vocabulary sizes were used. Indeed the sub-database detection accuracy goes up significantly if the vocabulary is expanded

Feature			Accuracy	Entropy
Wordnet	Biber	Words	in %	in Dit
most likely show type			32.2	3.31
●			56.9	2.41
	●		50.9	2.73
		50	61.3	2.35
●		50	61.5	2.25
	●	50	62.2	2.33
●	●	50	60.0	2.29
●	●		61.2	2.28
		250	62.7	2.17
		500	66.0	2.14
●	●	500	64.9	2.13
		5000	67.2	2.08

Table 4.7: **Show Type Detection:** Using the neural network described in Sec. 4.3 the show type was detected. If there is a number in the word column the word feature is being used. The number indicates how many word/part of speech pairs are in the vocabulary additionally to the parts of speech. (Table reproduced from Ries et al. (2000))

which hints at a limited dependency between topic and genre. This isn't really a surprise since there are many shows with weekly sequels and there may be some true repeats. On the other hand it should be noted that the resulting change in entropy is not great which indicates little effect in information access performance (Sec. 6.2). The results indicate that finer grained sub-databases can be detected with reasonable accuracy.

4.5 Activities

4.5.1 Introduction

The database and the TV show type detection task are located at the (sub-)database level of the information access hierarchy (Fig. 1.1). This section is devoted to the detection of activities which are located at the topic segment level. For this purpose three databases have been annotated manually with topic segmentations and activity annotations (see Sec. 1.4.3 for a more detailed description of the corpora):

meetings have been collected at Interactive Systems Labs at CMU (Waibel et al., 1998). A subset of 8 meetings has been annotated. Most of the

Measure	Meeting		SBC		CallHome Spanish	
	all	inter	all	inter	all	no story
κ	0.41	0.51	0.49	0.56	0.59	0.47
Mutual information	0.35	0.25	0.65	0.32	0.61	0.95

Table 4.8: **Intercoder Agreement for Activities:** The meeting dialogues and Santa Barbara corpus have been annotated by a semi-naive coder and the author. The κ-coefficient is determined as in Carletta et al. (1997) and mutual information measures how much one label "informs" the other (see Sec. 6.2). For CallHome Spanish 3 dialogues were coded for activities by two coders and the result seems to indicate that the task was easier or the training of the second annotator was better. The no-story column contains the results if the storytelling activities in the "official" annotation are excluded. This table is slightly extended from the authors work in (Ries and Waibel, 2001).

meetings are from the data annotation group itself and are fairly informal in style. The participants are often well acquainted and meet each other a lot socially.

Santa Barbara (SBC) is a corpus released by the LDC. 7 out of 12 rejoinders have been annotated.

CallHome Spanish is a corpus released by the LDC. All 120 dialogues have been annotated. Since the corpus is in Spanish different features are available and more annotations were done on this corpus (dialogue acts and games). The results are presented separately in Sec. 4.5.4.

Activities are described by action verbs and their distribution can be seen in Tab. 4.9. They have been annotated according to a scheme that was devised initially by Ries et al. (2000); Thymé-Gobbel et al. (2001) and is discussed in Sec. 4.5.2. The annotator has been instructed to segment the rejoinders into units that are coherent with respect to their topic and activity and annotate them with an activity which follows the intuitive definition of the action-verb such as discussing, planning, etc. The set of activities can be clustered into "interactive" activities of equal contribution rights (discussion, planning), one person being active (advising, information giving, storytelling), interrogations, and all others [8]. The annotation for CallHome Spanish has been done with great care and the first coder instructed and practiced with the second coder such that relatively high agreement

[8] It would seem that the dominance feature would be perfectly suited to distinguish between these interactive categories. Indeed it performs rather well on the Santa Barbara Corpus which (at least in our subset) contains meetings between people that don't know each other extremely well.

Activity	SBC	Meeting	CallHome
Discussing	35	58	8
Informing	25	23	n/a
Storytelling	24	10	901
Planning	7	19	87
Undetermined	6	8	83
Advising	5	17	81
Recording	3	2	73
Interrogating	2	1	86
Closing	0	1	25
Consoling	0	0	9

Table 4.9: **Distribution of Activity Types:** The meeting database and the Santa Barbara corpus (SBC) contain a lot of discussing, informing and storytelling activities however the meeting data contains a lot more planning and advising. The CallHome database on the other hand contains a significant amount of storytelling and is therefore a very different database. The informing category was not available to the taggers of CallHome while it seems to make sense to distinguish informing from storytelling in the other databases (table published by the author in similar form in Ries et al. (2000) and Ries and Waibel (2001)).

was possible despite the difficulty of the task. The instructions for the annotation of the meeting database did not go beyond the instructions given in the tagging manual and no comparison of tags was done before the intercoder data was annotated. The results for intercoder agreement (Tab. 4.8) are therefore naturally lower but fairly good overall given the complexity of the annotation problem. Especially the intercoder agreement for the interactive activities and CallHome Spanish are fairly good. It can be compared to Carletta et al. (1997)'s agreement results for transactions ($\kappa = 0.59$) which are the equivalent to activities in a task-oriented scenario. Given these agreement results the definition of the activities proposed seems to be reasonable given the complexity of the task and the intercoder agreement obtained. For classification a neural network without hidden units as in Sec. 4.3 was trained (hidden units did not improve the detection results).

4.5.2 Annotation of Activity Types

The annotation of activities has been conducted first on the CallHome Spanish database. On CallHome Spanish the main activity label (Sec. 4.5.2.1) was annotated as well as a possible evaluation by the speakers of the topic (Sec. 4.5.2.2.1) and the main topic / object / person being discussed (Sec. 4.5.2.2.2). These orthogonal annotations could for example yield a storytelling activity with a negative evaluation about other people (which may be called gossip). For the meeting and the Santa Barbara Corpus only activities have been annotated. This annotation scheme has been developed mostly by the author. It is an extended version of Ries et al. (2000) and will be included in Thymé-Gobbel et al. (2001).

4.5.2.1 Basic Activity Types

The list of basic activities types annotated on CallHome Spanish, meeting and Santa Barbara corpus is being presented. The storytelling activity was split into storytelling and informing for the meeting and the Santa Barbara corpus. These were also the instructions that were given to coders (orally).

4.5.2.1.1 Storytelling and Informing A strong cue for a storytelling activity are the following subparts in a story, usually in this default sequence, where all elements are optional or repeatable (Eggins and Slade, 1998; Labov and Waletzky, 1967):

1. abstract/introduction
2. orientation (initial part of story)
3. complication
4. evaluation
5. resolution
6. coda (a final wrap-up section containing relating this story to other things, finding the next topic to talk about etc.)

In many situations we also find appraisals in storytelling activities. However the abovementioned sequence is not necessary and serves only as an illustration. In most storytelling activities one speaker is dominant and assumes the role of the storyteller. There are two other options: Both are telling the story collaboratively or one speaker basically triggers the other all the time to continue the story and might therefore use a lot of the channel while not being the storyteller.

The informing activity is different from the storytelling activity since active person is just presenting information without telling a story or giving explicit advise. On CallHome Spanish informing wasn't coded as a separate category yet.

4.5.2.1.2 Planning Planning is a activity where people try to figure out the course of some future events they are intending to engage in. Planning also typically entails a mutual commitment to carry out the plan that was agreed upon. In CallHome Spanish planning typically relates to trips/visits, career changes and moving homes. Evaluations are very rare in planning and we have not identified substructure in planning activities.

4.5.2.1.3 Discussion Discussions are mutual exchanges of information on a certain topic, often coupled with appraisals. The discussion is different from the storytelling activity in that there is not just one central story that is being told and that the exchange is usual mutual. Topics of discussions in CallHome Spanish are usually news, sports and politics, rarely acquaintances.

4.5.2.1.4 Advising In an advising segment one speaker is giving – solicited or unsolicited – advice to the listener about a specific situation, usually a personal or professional matter. It usually includes instructions (weak or strong, commands and recommendations). The specific function of this activity is to express the speakers opinion about a rather personal or a professional issue and try to make the other person follow that advice. The advice is usually offered by the speaker who is more mature or has the higher authority. Evaluations are rare in this category.

4.5.2.1.5 Consoling Consoling is a activity that described as one speaker giving emotional support to the listener in times of personal misfortune (a divorce, the loss of a family member, an accident). We decided to include also situations were one speaker is praising the other since this is similar on the surface of the conversation and hard to determine. There is little or no evaluation in this category.

4.5.2.1.6 Closing The function of closing is to end a discourse segment because the speaker wants to move on and talk to a third person or just to end the whole conversation. It usually includes all the greetings and farewell expressions. There are no well defined topics and the utterances are usually short, but the activity itself can be long. This is common in the CallHome Spanish database since ending conversations in Latin American countries are bound to a set of rules of courtesy. There is no dominant speaker in these discourse segment, but rather an interactive exchange of farewell expressions. Evaluations are rare in closings.

4.5.2.1.7 Interrogation An interrogations is characterized as obtaining information through the use of questions. There is one dominant speaker who initiates and dominates the conversation, the other speaker would usually not have volunteered the information in another situation. The questions are intended to get specific – usually personal – information from the other participant in the conversation. The responses to the questions are usually short and are limited to answer the questions. The passive speaker does not take the floor of the conversation through his/her answers. There are usually no appraisals and evaluations and the activity is rare in CallHome Spanish.

4.5.2.1.8 Recording The CallHome Spanish database contains segments that are directly and obviously generated by the recording environment. Typical topics are: The length of time that the speakers have available for talking, the purpose of the phone call, whether the phone call is free, whether or not the speakers are being recorded and finally how they found out about the free phone calls. In one extreme example the overhearer is explicitly addressed and educated about Spanish phonology. In the other databases people are sometimes still playing with the recording equipment or say some sentences for the speaker identification system.

4.5.2.1.9 Undetermined Discourse segments that are incomplete due to the segmentation of the transcript, especially when not enough material is available to make an activity decision. Some discourse segments are also labeled undetermined when the participants in the conversation purposely exchange information in codes or use language that can have multiple interpretations and make it, therefore, incomprehensible to the tagger. This activity is quite common at the beginning and at the end of conversations.

The tagging scheme might not capture all possible activity types. The taggers were encourage to tag segments that do not fit with special labels of their choosing. Although we tried hard we were not able to find more or different categories that would warrant a separate category. These cases were mapped on the "undetermined" category.

4.5.2.2 Orthogonal Activity Annotation

4.5.2.2.1 Evaluations Positive and negative evaluations can be used for people, relationships or behaviors. Evaluations of events, incidents, tangible things or social constructs however will not be marked. The original goal of marking evaluations is to discriminate between neutral stories that are being told and stories that have more gossip character. The evaluations that we are marking have to be *explicit* on the surface of the conversation. Other evaluations are usually very hard to decided as an overhearer so we left them unmarked. A dialogue can be

Orthogonal annotations (1353 segments in 120 dialogues)			
Evaluation	Count	Who or what	Count
positive	55	speaker A or speaker B only	342
negative	79	speaker A and speaker B only	56
divergent	10	other people (may include A or B)	562
neutral	1209	practical topics	228
		politics	17
		other	109
		phone call	39

Table 4.10: **Orthogonal Activity Annotation:** The topical segments in CallHome Spanish were not annotated with activities (see Tab. 4.9) but additionally with the main object/subject being discussed and the evaluation of the speakers. If the speakers disagreed in their evaluation it was labeled divergent (reproduced from Ries et al. (2000)).

marked as neutral (no tag), positive, negative evaluation by at least one speaker or divergent evaluation (Tab. 4.10).

4.5.2.2.2 Who or What The "who or what" category is orthogonal to the other activity features. It is meant to capture the main object/subject that is being discussed. In CallHome Spanish this may be a person in the storytelling activity, for the planning sections a trip to some place etc. The "who or what" category would have been called the atomic or discrete topic by Goutsos (1997). The annotators identified the actual object/subject and later on these have been classified (Tab. 4.10). This annotation, however, was not exploited further.

4.5.3 Meetings and Everyday Rejoinders

Meetings and the meeting like situations in the Santa Barbara database are data which is very similar to the target domain. The detection results for meetings over all activities are not extremely (Tab. 4.11) high which is no surprise when observing that the intercoder agreement is fairly low as well (Tab. 4.8). The results for the Santa Barbara corpus are much better and especially word level information and WordNet categories were important to use. The detection of interactive activities works fairly well using the dominance feature on SBC which is also natural since the relative dominance of speakers should describe what kind of interaction is exhibited. One important difference between our meeting database and the Santa Barbara database is that the participants in our meetings all know each other very well such that the meetings might be more informal in nature which might

				Detection accuracy in %			
Dialogue Act	Dominance	SBiber	WordNet	all		interactive	
				SBC	meet	SBC	meet
baseline (most likely activity)							
				32.7	41.1	50.5	54.6
no words in the input features							
○				28.1	37.6	47.7	56.7
●				28.0	36.2	46.7	65.3
	●			32.7	44.7	64.5	58.2
		●		24.3	35.5	53.3	58.9
			●	37.4	37.6	46.7	52.5
words per channel as a input feature							
				38.3	39.7	53.3	54.6
●				42.1	37.6	57.0	61.0
	●			41.1	41.1	52.3	58.9
		●		42.1	38.3	52.3	57.5
			●	49.5	39.0	53.3	57.5
	●	●		42.1	39.7	53.3	60.3
●		●		39.3	40.4	57.9	61.0
Annotations of the author							
				59.8	57.9	73.8	72.7

Table 4.11: **Activity Detection:** Activities are detected on the Santa Barbara Corpus (SBC) and the meeting database (meet) either without clustering the activities (all) or clustering them according to their interactivity (interactive) (see Sec. 4.5.3 for details). The ○ in the dialogue act detection column represents the dialogue act histogram per channel (where one channel is reserved for the most active speaker, the other for all the other speakers). Furthermore pause lengths, overlap, coverlap and segment length were tested with disappointing results (Sec. 4.2.3 describes the interactional features) (reproduced from Ries and Waibel (2001)).

explain these differences.

The dialogue act distribution on the other hand works fairly well on the more homogeneous meeting database were there is a better chance to see generalizations from more specific dialogue based information. Interactional features however or other features that pertain to segment lengths, pause lengths, overlap lengths, word lengths and so forth were completely ineffective. Overall the results show that the features that were suggested are reasonable and apply in intuitive ways to the corresponding detection problem. The differences in features that are useful in different tasks indicates that there are no golden bullets to solve the problem.

4.5.4 CallHome Spanish

The CallHome Spanish database allows for different types of results than the meeting and Santa Barbara databases in Sec. 4.5.3 since significantly more data is available (120 dialogues instead of less then 10) and dialogue acts and games have been manually annotated. On the other hand the conversations represent phone calls between relatives and as such may not represent a domain that is interesting for information access applications. As observed in Sec. 5, especially Tab. 5.5, the speaker initiative / dominance distribution in CallHome Spanish may be skewed in one direction such that dominance might not be so indicative. Similarly the storytelling activity is very frequent and is — as one might suspect — very hard to detect.

The manual annotation enabled the training of an automatic dialogue act and game detector for Spanish. The detector was trained and applied in a Round-Robin fashion on the whole CallHome Spanish database, annotating 40 dialogues in each turn while the model was trained on the remaining 80 dialogues. The detector used an integrated ngram language model classifier for dialogue act and game detection and the dialogue game discourse model was a unigram (Sec. 3.6.3, Tab. 3.7).

Tab. 4.12 shows the detection results using a neural network classifier without hidden units as used in Sec. 4.5.3. The results demonstrate that word level information is crucial in obtaining good results for detection as well as entropy. There is also clearly a performance hit when the detected dialogue acts and games are used but at the same time it is clear that using dialogue act information is just as crucial as word level information. Dialogue act and game information may have the advantage that it is more robust against changes in the scenario whereas word level information might be more easily fooled. The best feature combination seems to be word level, SBiber, dialogue act and dialogue game, whether dialogue acts and games were automatically detected or not.

The largest contributor to the error rate is the distinction between storytelling

Dialogue		Domi-	Biber	Accuracy in %		Entropy in Bit	
Act	Game	nance		manual	detected	manual	detected
no word features							
baseline				66.6		1.83	
○				67.5	67.5	1.51	1.57
●				68.4	67.6	1.82	1.54
	●			68.0	67.0	1.55	1.73
		●		66.6	66.6	1.82	1.82
			●	66.6		1.66	
○	●			68.3	67.4	1.44	1.56
●	●			68.6	67.6	1.42	1.53
	●	●		68.0	67.0	1.38	1.73
●	●	●	●	69.5	68.0	1.38	1.47
word features							
				67.5		1.50	
●				69.5	68.5	1.37	1.41
		●		67.5	67.3	1.50	1.50
			●	67.8		1.48	
		●	●	67.8	67.6	1.48	1.48
●			●	69.9	69.2	1.35	1.40
	●		●	69.3	67.7	1.37	1.46
●	●			70.1	68.8	1.32	1.41
●	●		●	70.4	69.9	1.32	1.40
	●	●		68.9	67.0	1.38	1.73
	●	●	●	69.3	67.8	1.37	1.46
●	●	●	●	70.4	69.0	1.32	1.39

Table 4.12: **Activity Detection on CallHome Spanish:** Using neural networks with no hidden units we achieved a reasonable detection accuracy. The detection results are given for *manual* as well as for automatically *detected* dialogue acts and games. The ○ in the dialogue act detection column represents the dialogue act histogram per channel (where one channel is reserved for the most active speaker, the other for all the other speakers). Furthermore pause lengths, overlap, coverlap and segment length were tested with disappointing results (Sec. 4.2.3 describes the interactional features). The entropy and accuracy results which are printed in the middle of the *manual* and *detected* columns are using feature sets that make no use of dialogue act or game information (reproduced from Ries et al. (2000) but using a larger database of 120 telephone-calls).

manual	automatic	
	storytelling	other
storytelling	300	50
other	87	84

Table 4.13: **Storytelling Confusion Matrix:** Discriminating between stories and non-stories can be done with a classifier at a 73.7% level while 67% is the baseline result just picking storytelling. This result seems to be the limiting factor for the activity detection results (reproduced from Ries et al. (2000)).

manual	automatic							
	advising		closing		discussion		planning	
		recording		consoling		interrogating		undeterm.
advising	12	7	·	·	·	1	4	7
recording	4	25	1	·	·	2	1	4
closing	1	2	4	·	·	·	1	1
consoling	·	1	·	·	·	·	1	·
discussing	1	·	·	·	·	·	2	1
interrogating	1	2	·	·	·	17	1	5
planning	6	4	·	·	·	1	11	6
undetermined	2	10	·	·	·	7	3	12

Table 4.14: **Activity Detection Confusion Matrix excluding Storytelling:** While this detection task is far from being solved (47.4% at a 21.6% baseline) it seems that the activity detection task excluding storytelling is significantly more tractable. The intercoder experiment shows a cross-human accuracy of 53.8% at a 27.7% baseline accuracy on 3 dialogues which seems to indicate that the classifier achieves a very good result (reproduced from Ries et al. (2000), only 80 dialogues used).

manual	automatic		
	negative	neutral	positive
negative	·	28	·
neutral	2	290	2
positive	·	24	2

Table 4.15: **Evaluation Detection Confusion Matrix:** The current detection results are just the baseline (84.4%). The detection of evaluations is considered a hard problem and may not be tractable in a fully spontaneous database like Call-Home (see also Bruce and Wiebe (1999)) (reproduced from Ries et al. (2000), only 80 dialogue used).

other activities while the discrimination between non-storytelling activities can be done with reasonable accuracy (Tab.4.13 and Tab. 4.14). This result gives hope for a more balanced database where storytelling is not that dominant. Finally we tried to detect the orthogonal evaluation attribute (Sec. 2.4.2.2) but the initial results are not very promising (Tab. 4.15). These orthogonal categories are therefore unlikely to be useful in information access. Overall the results are very promising, especially for databases where people are not as closely acquainted as they are here.

4.6 Conclusion

Database type detection is almost trivial according to the results and even the detection of subdatabases can be done at a good accuracy. The fact that gender was also easy to detect using stylistic features seems to suggest that social roles may also be detectable – however we didn't have enough other data to verify that claim by other means.

Within a rejoinder on a topical level one may want to distinguish activities such as storytelling, planning, discussing etc. These features also seem to be important for humans in information access as seen in the user study (Sec. 6.3). The disappointing part about activities is that they don't seem to be very easy to distinguish by humans themselves. However machine learning techniques seem to be able to extract a significant part of activity information from low level features. Fig. 4.1 summarizes the detection performance in comparison to machine performance in a graphical fashion. The data for meetings and the Santa Barbara corpus is fairly small such that the machine learning approach may not have seen enough examples. CallHome Spanish, presenting almost one order of magnitude more data than these two combined, features a lot of storytelling which is untypical for the other domains and hard to distinguish. If the storytelling category is eliminated the result of the classifier are quite good (Tab. 4.14) such that the classification approach appears to work in general although a significant database size is necessary. This also reveals that as long as the activities are expressed in a distinguishable form the classification approach can be successful.

It turns out that all features used in the study were to some extent important to achieve good performance but the best feature set depends on the application: There is no golden bullet. It is clear however that simple features, which are already fully sufficient for the database discrimination task, are not effective for the activity detection tasks whereas the more complex features seem to be a real waste of energy to apply to the simple task.

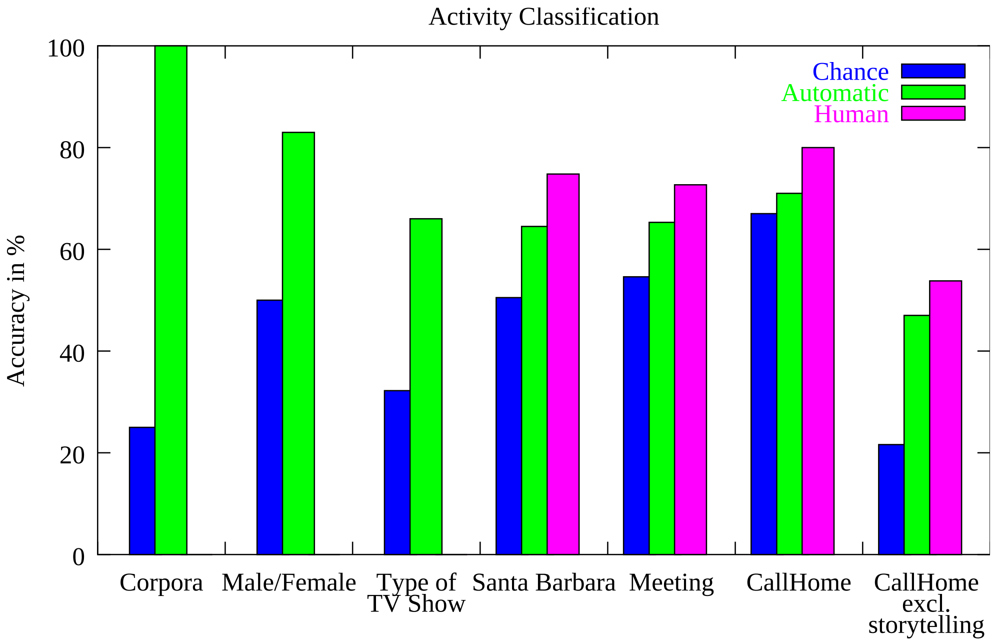


Figure 4.1: **Detection Accuracy Summary:** The automatic identification of different *corpora* can be done with high accuracy using simple features (Sec. 4.3). Similarly it was fairly easy to make the *Male/Female* speaker distinction with stylistic features alone. Discriminating between sub-databases such as *Types of TV Shows* (Sec. 4.4) can be done with reasonable accuracy. However it is a lot harder to discriminate between activities within one conversation for personal phone calls in *CallHome* (Sec. 4.5.4), for general rejoinders in the *Santa Barbara* corpus or in *Meetings* (Sec. 4.5.3). The activity detection result for CallHome Spanish excluding storytelling (*CallHome, no story*) is fairly good since the activities are reasonably different and the database is not too small. Chance performance is achieved by picking the most frequent database or activity, human performance measures the agreement of one human annotator with another. The automatic performance is achieved using neural network based classification.

Chapter 5

Topical Segmentation

5.1 Introduction

Segmenting a text or dialogue into meaningful units is a problem that has received considerable attention in the past and can be seen as a preprocessing stage to information retrieval (Mittendorf and Schäuble, 1994), summarization (Zechner and Waibel, 2000a), anaphora resolution, and text/dialogue understanding. This chapter introduces a probabilistic approach to dialogue segmentation using keyword repetition, speaker initiative and speaking style which underlines the importance of non-topical features in speech understanding. The chapter is in large a reproduction of work by the author (Ries, 2002). The results for prosody and pause features have only been mentioned in (Ries, 2002) and Tab. 5.3 and Tab. 5.4 have been added for completeness.

The basic algorithm is compared to other standard approaches for domain independent topic segmentation (Choi, 2000; Hearst, 1997) and shows excellent performance. The intended applications are navigation support for meetings and other everyday rejoinders and preprocessing for applications such as information retrieval (Arons, 1997; Choi et al., 1999; Stifelman, 1995; Stifelman et al., 2001; Whittaker et al., 1994).

A clean probabilistic framework is presented which allows to formulate “coherence” and “region” features. “Coherence features” are features which are coherent within one topical segment – examples are the keyword distribution or speaker initiative. Speaker initiative is encoded by the speaker identity for each turn and possibly the information whether the turn is long or short. (see Sec. 5.6 for a discussion and experiments on the encoding of speaker initiative). “Region features” on the other hand are designed to model properties of different regions of topical segments such as the boundary and the beginning, middle and end of a topic. Region features are a generalization of the more classic boundary classification approach and are used to model the change in the part of speech distribution over a topical segment. Region features could also be used to encode features such as prosody and pause lengths but the results on the databases tested were not as good as anticipated.

The chapter is first presenting the related work (Sec. 5.2), describes the evaluation measure (Sec. 5.3), continues with information about the coding instructions and intercoder agreement (Sec. 5.4) and the basic algorithm used (Sec. 5.5). Experimental results are presented (Sec. 5.6) and finally conclusions are offered (Sec. 5.7).

5.2 Related Work

5.2.1 Segmentation of Speech Data

The databases used in the experiments contains everyday rejoinders, meetings and (personal) telephone conversations. Additionally the publicly available database of a recent publication on topic segmentation (Choi, 2000) is used for comparison such that a more traditional written text comparison is available as well. However meetings and other everyday rejoinders are fairly different from broadcast news data which has been the focus of information access to speech documents in the recent TREC-SDR (information retrieval) (Garofolo et al., 1999) and TDT (topic detection and tracking) initiatives such that the results for those domains cannot be compared with. The following key properties are different:

speech recognition performance Typical best practice Large Vocabulary Speech Recognition (LVCSR) word error rates on Broadcast News have been around 20% (Allan et al., 1998) for fast decoders in 1998 whereas it is around 40% for systems without real-time considerations on meeting data in 2001 (Sec. 2.4.5.2).

domain knowledge Broadcast News seems to cover a large domain yet Yamron et al. (1998) were able to use only 100 topic models for segmentation. Such preconstructed topic models can't be assumed for everyday rejoinders such as meetings: In our meeting database "speech data transcription" was an important topic but the topic would be very rare in a general (speech) database. The number of topics of high national or international interest however is limited such that an approach which is based on a relatively small number of example topics may work fairly well.

Additionally keywords in everyday rejoinders may be highly idiosyncratic such that they may not be in the vocabulary of an LVCSR system. A lot of the information in the Broadcast News database is also available in print-media and in electronic form such that document expansion (Singhal and Pereira, 1999) and vocabulary adaptation (Geutner et al., 1998) can make use of these resources.

manual cuts and genre constraints Broadcasts and especially news shows are very specific genres which are designed for mass consumption: News shows for example are cut in short but very distinct stories which are introduced or ended with very specific phrases. A video of a newsshow would also likely have a "cut" at a topical boundary which could be detected easily (Sec. 2.4.7). Everyday conversations on the other hand don't exhibit such clear topical boundaries and topic-shifts may occur gradually.

5.2.2 Topic Segmentation Criteria

The definition of topic as such is difficult and is disputed. The only question that needs to be addressed here is how to annotate topic manually and later automatically in a reliable fashion. Human coders seem to be able to reach decent agreement about topical segmentations even with little instruction other than appealing to the natural notion of topic (Hearst, 1997).

Albeit this method does not directly support principled insight into the structure of topic it is a reasonable starting point and it has the advantage that a property that may be accessible to the user of an information retrieval system has been chosen. Given decent empirical data one may be able to design algorithms which are guided by presumed properties of topic and algorithms which are based on reliable criteria should perform well. Typically these studies involve the assessment of intercoder agreement to make sure that the manual annotation results in a valid empirical study (Passonneau and Litman, 1997).

Six basic approaches for automatic procedures and criteria can be distinguished: Halliday and Hasan (1976); Hearst (1997) and Yamron et al. (1998) suggest that segments are assumed to contain semantically distinct elements, usually presented by lexical cohesion; Beeferman et al. (1999); Passonneau and Litman (1997) suggest that local features indicate topic shifts; Marcu (1997) proposes an approach based on rhetorical structure to derive a hierarchical representation of the dialogue; Hirschberg and Nakatani (1998); Shriberg et al. (2000) show how to use automatically detected prosodic features for segmentation; Walker and Whittaker (1990) use initiative as a manual segmentation criterion and finally multimodal features such as gesture, movement and gaze are proposed by Quek et al. (2000).

Discourse theories such as Grosz and Sidner (1986); Mann and Thomson (1988) would also be attractive candidates for segmentation of human dialogue and indeed Marcu (1997) has shown success in parsing rhetorical structure in text domains using keywords and phrases (Sec. 2.4.3.4). It is however unclear if a tree structured theory is applicable to our databases and whether the keywords and phrases used would provide the necessary segmentation and structuring information (see Sec. 2.4.3.4 for a longer discussion).

The author therefore decided to use the widely studied keyword repetition feature (Halliday and Hasan, 1976; Hearst, 1997; Yamron et al., 1998) and speaker initiative as a "coherence features". Speaker initiative was so far only used as a manual segmentation criterion (Walker and Whittaker, 1990) but is used as an automatic criterion here. Speaking style as encoded in the part-of-speech distribution is explored as a "region feature". Pitch and pause – which could also be encoded as "region features" – have been implemented but were not successful on our database. The suggested algorithm allows a direct integration of "coherence features" and "region features" in contrast to previous algorithm designs.

5.2.3 Keyword Repetition Algorithms

The part of the algorithm which handles coherence features is related to the approach of Reynar (1998); Yamron et al. (1998). Yamron et al. (1998) assumes that each segment of a conversation is generated by one out of a couple of hundred pretrained topics – the algorithm is therefore domain dependent. The algorithm presented here does not make that assumption and is therefore domain independent. The domain independence is achieved by training a model for each segment on the fly instead of relying on pretrained models. An advantage of Yamron et al. (1998) is that information about semantically related terms is implicitly exploited in their procedure. This may be achieved in our algorithmic framework using techniques such as Ponte and Croft (1997) – however the technique of Ponte and Croft (1997); Yamron et al. (1998) rely on the availability of adequate training material which may not be available for everyday discourse or meetings. A fair comparison to Yamron et al. (1998) is not possible since there is really no topic repetition across dialogues in our databases which would disfavor their approach while the TDT database would likely require to add synonym handling to the proposed algorithm to be competitive.

Reynar (1998) presents the domain independent probabilistic *wordfrequency algorithm* for topical segmentation. It estimates the probability of a boundary for every location in the text and uses a thresholding technique to derive the actual boundaries. The drawback is that the estimation assumes fixed sized windows around the boundary and the boundary placement is not optimized globally unlike the Viterbi search employed by Yamron et al. (1998) and the proposed algorithm.

Hearst (1997) is probably the most widely cited domain independent algorithm for topical segmentation and relies on cosine similarity measures combined with heuristic optimization criteria and optimization procedures. Similar algorithms, applying similar measures with different optimization criteria, are Choi (2000); Reynar (1998). Hearst (1997) and Choi (2000) were chosen to establish a comparison to existing domain independent algorithms: Hearst (1997) is known widely and Choi (2000) is the most recent publication in this area which compares to Hearst (1997); Kan et al. (1998); Kozima (1993); Reynar (1998).

5.2.4 Boundary Classification Algorithms

Many algorithmic approaches have used boundary classification: A classifier is trained which has the output "Boundary: Yes/No". Using "region features" the classifier can be extended to produce other outputs for larger regions such as "Begin of topic", "End of topic" and so forth. The UMass approach in Allan et al. (1998) seems to model word type information in different regions of topical segments using an HMM model. The model presented here can be trained using a discriminative classifier but imposes a fixed structure on the topical segment.

Since news shows are a highly organized genres following specific scripts very specific topic shift indicators (such as LIVE, C. N. N.) can work very well which was used by Beeferman et al. (1999); Hirschberg and Nakatani (1998). Other topic indicators studied as are keyphrases, pauses and prosodic features such as a preceding low boundary tone or a pitch range reset (Hirschberg and Nakatani, 1998; Passonneau and Litman, 1997; Shriberg et al., 2000). While prosodic features may be modeled easily as region features the author hasn't been able to establish good results on the dialogues although the prosody module has been tested successfully on an emotion detection task (Sec. 3.8). The reason may again be a difference in genre since topic changes may be marked explicitly in broadcasts.

Boundary classification algorithms may also integrate information about the change in the keyword distribution using features similar to most keyword repetition algorithms (Beeferman et al., 1999; Hirschberg and Nakatani, 1998). The critique of this technique is that it relies on local, window based changes of the keyword distribution and that the algorithms don't apply a global optimization over all possible sequences [1]. On the other hand the algorithm presented in this work as well as Choi (2000) and Yamron et al. (1998) integrate keyword information over the complete topical segment.

[1] One may argue that exponential segmentation models (Beeferman et al., 1999) may weigh the contribution of the keyword repetition feature with the other models in a principled way. On the other hand the parameterization of the exponential models used may also be interpreted as a different weighting scheme between the log probabilities "coherence features" and "region features". A pilot experiment using this weighting scheme and a couple of settings did not indicate any change of the segmentation accuracy.

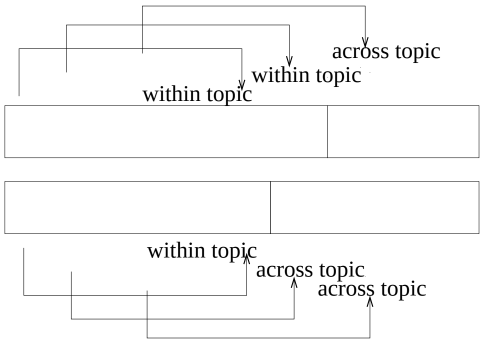


Figure 5.1: **Link Error Rate:** The link error rate is defined by looking at links of fixed length k (words at distance k) and determining whether they are within or across topic. If corresponding links in the reference and hypothesis disagree on this classification a link error occurs. In the figure only the middle link represents a link error. The link error rate is the average link error.

5.3 Evaluation Methods

A standard evaluation metric for text segmentation has been suggested by Allan et al. (1998); Beeferman et al. (1999). The metric is fairly intuitive and Beeferman et al. (1999) argues that it is fairly robust against simple cheating attempts. The intuition behind the metric is that a segmentation is good if two words that belong to the same topic in the reference belong to the same topic in the hypothesis.

More specifically if two words have distance k they form a *link*. A link is called *within topic* if the two words belong to the same topic, otherwise it is across topic. If the corresponding links in the hypothesis and reference are both *within topic* or both *across topic* the hypothesis is correct, otherwise it is an error. The reported metric is the average link error rate in percent. For each database k is half the average topic length of the database (Fig. 5.1).

All speech databases have been manually segmented based on the manual transcripts. The results for automatic transcripts of the meeting database have been obtained by transferring the manual topic segmentation to the results generated by the speech recognizer. The speech recognition system segments the input by pause length. Based on time stamps the next best utterance beginning is chosen as the segment boundary.

Choi (2000) calculates the link error rate differently and his technique is used when reporting results on the C99-database. The first step in his procedure is to calculate the average link error rate for every text in the database. The link length

k for every text is determined as half the average topic length of the respective text. The average link error rate of a database is the average of the average link error rate of all texts in the database.

As a baseline an "equal distance segmentation" is being used, similar to B_e in Choi (2000). The dialogue is segmented into utterances with equal sized topics of length d where d is the average length of a topic in a training set. The parameter d is estimated in a Round Robin procedure.

5.4 Coding Instructions

Currently the only reliable way to construct a database of topics is to annotate them by hand and check the agreement between coders (Sec. 5.2.2). A much simpler solution could be to compose artificial data by randomly picking initial segments from different documents as topics of an artificial document. This method is used by Choi (2000) in his C99-database. The problem with that approach is that the modeling of topic length may be artificial and the inner structure of topics may not be natural.

However this work is concerned with the segmentation of naturally occuring dialogue in meetings and everyday rejoinders where topic shifts are usually smooth and uninitiated in contrast to the construction of the C99 database. The topic definition applied in this work instructs the coders to place a boundary where the topic changes or the speakers engage in a different activity such as discussing, storytelling etc. The activities were annotated at the same time as the topic segmentation was produced (Ries et al., 2000; Ries and Waibel, 2001). The topic definition therefore contains the notion of activity additionally to the standard appeal for a "natural" topic definitions (see. Sec. 5.2.2). The primary annotation for all databases was done by semi-naive subjects. Some had considerable experience annotating the databases with dialogue features however no special linguistic training or criteria were provided for the topic segmentation task beyond the definition of activities.

The meeting database was also segmented by the author. The intercoder agreement was measured by (a) treating the second human similar to a machine using the standard evaluation metric (Sec. 5.3, Fig. 5.5), (b) measuring κ for the boundary/non-boundary distinction for each utterance ($\kappa = 0.36$) and (c) measuring κ for the classification of links [2] as *within topic* / *across topic* ($\kappa = 0.35$). The κ-statistics (Carletta et al., 1997) therefore indicates that the intercoder agreement is relatively low overall which is not surprising given the difficulty of the task. The result seems to be in the same range as other similar annotations (Passonneau

[2] Refer to Sec. 5.3 for the description of links.

and Litman, 1997).

5.5 Probabilistic Modeling

5.5.1 Introduction

The algorithm is based on a standard probabilistic modeling approach. If d is a dialogue and L is a possible segmentation the Viterbi algorithm is used to find the best segmentation L^*

$$L^* = \mathrm{argmax}_L \;\; p(L|d) = \mathrm{argmin}_L \;\; -\log p(\mathrm{s})$$

where $\mathrm{s} = \langle d_0, \ldots, d_n \rangle$ is the dialogue segmented into topical segments d_i. The model for $p(s)$ is assumed to be decomposable into models for the number of segments per dialogue $p(\#segments)$, the length of each segment $p(\mathrm{length}(d_i))$ and models for the content of each segment given the segment length $p(d_i|\mathrm{length}(d_i))$:

$$p(s) = p(\#segments) \prod_i p(\mathrm{length}(d_i)) p(d_i|\mathrm{length}(d_i))$$

The most crucial assumption of this model is that all segments are be independent of each other which is invalid in general, especially when a topic is resumed after a digression. If the exponential model is chosen as the model for the number of segments in a conversation one assumes that the probability that the conversation is over (p_{segs}) does not change depending on the number of topics in the conversation so far such that

$$p(\#\mathrm{segments}) = p_{segs} \cdot (1 - p_{\mathrm{segs}})^{\#\mathrm{segments}}$$

holds. While this assumption seems to be simplistic it simplifies the the derivation of the algorithm since $p(s)$ simplifies to:

$$p(s) = p_{\mathrm{segs}} \cdot \prod_i (1 - p_{\mathrm{segs}}) p(\mathrm{length}(d_i)) p(d_i|\mathrm{length}(d_i))$$

If the length for each segment is described by an exponential model as well

$$p(\mathrm{length}(d_i)) = p_{utts} \cdot (1 - p_{\mathrm{utts}})^{\mathrm{length}(d_i)}$$

we obtain [3]:

$$p(s) = p_{\mathrm{segs}} \cdot (1 - p_{\mathrm{utts}})^{\sum_i \mathrm{length}(d_i)} \cdot \prod_i (1 - p_{\mathrm{segs}}) \cdot p_{\mathrm{utts}} \cdot p(d_i|\mathrm{length}(d_i))$$

[3] The simplification is not necessary to obtain an effective algorithmic solution. Another option is to view the lengthmodel as a separate model which has the same properties as coherence and repetition features; the assumption of an exponential model leads to the subsumption of the model into a general parameter that needs to be tuned independently.

Since the number of utterances in a conversation ($\sum_i \mathrm{length}(d_i)$) is independent of the actual segmentation $p_{\mathrm{segs}} \cdot (1 - p_{\mathrm{utts}})^{\sum_i \mathrm{length}(d_i)}$ is a constant which does not depend on the actual segmentation. It can therefore be eliminated from the optimization criterion. The dialogue segmentation model is therefore

$$L^* = \mathrm{argmin}_L \sum_i P \; - \log p(d_i|\mathrm{length}(d_i))$$

where $P := -\log(1 - p_{\mathrm{segs}}) \cdot p_{\mathrm{utts}}$ is a constant which may be chosen to optimize the segmentation performance. Since P controls the number of segments it is also called "lengthmodel".

Since the dialogue d is known we may call $d[k : l]$ the segment ranging from k to l and define

$$M_{k,l-k-1} := -\log p(d[k : l]|\mathrm{length}(d[k : l]))$$

Finding the most likely sequence corresponds to finding the best sequence L of strictly ascending indices such that the sequence contains 0 as the first index and the size of M as the last:

$$L^* = \mathrm{argmin}_L \sum_{0<i\leq \mathrm{size}(M)} P + M_{L[i-1],L[i]-L[i-1]-1}$$

Note that both the repetition (Sec. 5.5.3) as well as the region model (Sec. 5.5.2) may be formulated in this framework and that the minimization can be carried out using a dynamic programming approach.

Since very long segments are extremely unlikely our implementation uses a maximum length constraint, which will be 300 turns for all the experiments presented (see Fig. 5.2). Complexity reduction is achieved in two ways: The search has to consider far fewer alternatives and the matrix M has to be constructed only for segments shorter than the maximal length constraint. It is easy to observe that the algorithmic complexity of the search algorithm without an upper bound is $O(n^2)$ while the complexity with an upper bound is $O(n)$ where n is the length of the dialogue. The same observations can be made about the (partial) construction of M as long as efficient construction procedures are used for M. Linear runtime is very important for segmenting oral communication since rejoinders are often fairly long. Four types of lengthmodels have been tested (Fig. 5.1):

cheating A genie tells the algorithm the number of segments from the manually annotated reference. A logarithmic search for the segment penalty is performed to achieve the desired number of segments

average length Given the number of turns use the average segment length to calculate the expected number of segments. The average segment length can be estimated easily by observing all segment lengths in a training corpus. This method is used for all dialogue experiments.

penalty The segmentation algorithm is applied to the training dataset where an optimal P is determined for each conversation. The average over these P is calculated and used for testing. During testing this penalty is added to each segment such that this method is more computational intensive in training but does not add time in testing. It delivers the best results for the C99 database and is used in all experiments on that database.

choif For comparison the number of segments as delivered by the procedure C99 of Choi (2000) can be used.

explicit After learning the probability distribution of segment lengths it can be added to the probabilistic model – indeed the derivation of the segmentation model does not depend on the assumption of an exponential model of segment length. Histogram based models, normal and gamma distributions have been used but results will only be presented for the normal distribution. This model may be combined with the penalty approach.

For all models the necessary parameters are trained in a Round-Robin procedure on all but some segments and tested on the held out segments. This procedure is repeated such that all elements of the data set are tested once.

5.5.2 Coherence Features

Keyword repetition and speaker initiative can both be modeled as coherence features by assuming that each segment follows its own language model. In the case of keyword repetition the language model describes the keywords, for speaker initiative it describes the speaker identity of an utterance and potentially an indication of the initiative such as utterance length (see Sec. 5.6 for details on the implementation of the features). The probabilistic model requires to define $\log p(d_i|\mathrm{length}(d_i))$ in an appropriate fashion. In speech recognition Kuhn and de Mori (1990) pioneered the use of so called cache models that adapt themselves over time. At the beginning a (dynamic) cache model is initialized with an a priori distribution but as each new word is predicted and becomes part of the context the model is adapted. However in the speech recognition community a static cache model has also been used frequently: A segment of speech is recognized, the language model is adapted to that segment and the segment is recognized with the adapted model. For the modeling of coherence features this idea might be used by training a (discounted) model on the topical segment and testing it on the same segment. While initial experiments used both models there are no differences in the experimental results. Since the static model is simpler to implement and faster in execution it is used for all experiments reported. The static model approach seems to be somewhat counterintuitive at first but it can also be explained in the minimum description length framework (Cook and Holder, 1994). The probabilistic framework can explain the static model by associating a language model with each segment boundary. The random variable L may be reinterpreted as the segmentation including the segment language model. If all language models are assumed to be equally likely it is modeled by another penalty that can be subsumed by P. To obtain better estimates and avoid "zero probabilities" the cache model was smoothed using absolute discounting with a fixed parameter $D = 0.5$ (Ney et al., 1994) [4]. If $count_w$ is the count for words w in a segment d_i the log-likelihood

$$-\log p(d_i) = -\sum_w \mathrm{count}_w \log \hat{p}(w)$$

is based on the probabilities of the individual words $\hat{p}(w)$ according to absolute discounting for a fixed and finite vocabulary V of size $|V|$:

$$\hat{p}(w) \quad := \quad \begin{cases} \frac{\mathrm{count}_w - D}{s} & \text{if } \mathrm{count}_w > 0 \\ \frac{(|V| - \mathrm{zero}) \cdot D}{\mathrm{zero} \cdot s} & \text{if } \mathrm{count}_w = 0 \end{cases}$$

where $s := \sum_w \mathrm{count}_w$ and $\mathrm{zero} := \sum_{w,\mathrm{count}_w = 0} 1$. The second case does not contribute to the overall likelihood and can therefore be omitted. Additionally the implementation chose V to be the words in the segment such that no word had a count of 0.

$$\sum_i \log p(\mathrm{d}_i|\mathrm{length}_i) \quad = \quad s \cdot \log s - \sum_{w, count_w > 0} \mathrm{count}_w \cdot \log(\mathrm{count}_w - D)$$

One may observe that $M_{i,j+1}$ can be calculated from $M_{i,j}$ such that M can be constructed efficiently.

5.5.3 Region Features

Region features are an extension of the common boundary modeling approach to discourse segmentation. A region mapping is a function f which maps an integer k onto an array of k region labels. It can therefore be naturally extended to a function f' which maps a segment d_i containing k utterances to k segmentation

[4] The discounting method and parameters used were fairly uncritical when compared to alternatives during prestudies. Absolute discounting was chosen for its simplicity and since it is in widespread use.

labels. The intuition is that if the length of the segment is known it has to follow a certain fixed pattern. The simplest example is the classic *boundary modeling* approach where

$$f(k)[j] \quad := \quad \begin{cases} \text{BOUNDARY} & \text{if } j = 0 \\ \text{NONBOUNDARY} & \text{otherwise} \end{cases}$$

Boundary modeling assumes that there are very specific phrase or intonational events at or near the boundary (key words and phrases). The *equal size regions* approach (3 regions for Begin, Middle and End) can easily model changes in general distributions such as the part of speech distribution: At the beginning new items are introduced explicitly whereas they are referred to anaphorically towards the end. They can be combined in the *equal size + boundary* approach which features one region for the boundary and Begin, Middle and End regions. Since f is a deterministic function of $\text{length}(d_i)$

$$p(d_i|\text{length}(d_i)) = p(d_i|f(\text{length}(d_i)), \text{length}(d_i))$$

In order to make this quantity tractable independence assumptions have to be made ($f'(d_i)_j$ is the jth region label of $f'(d_i)$, d_{i_j} is the jth utterance of the region d_i):

conditional independence all segments in a topic are independent given the segmentation labels
$p(d_i|f(\text{length}(d_i)), \text{length}(d_i)) = \prod_j p(d_{i_j}|f(\text{length}(d_i)), \text{length}(d_i))$

independence of context all segments depend only on their respective segmentation label
$p(d_{i_j}|f(\text{length}(d_i)), \text{length}(d_i)) = p(d_{i_j}|f(\text{length}(d_i))_j, \text{length}(d_i))$

independence from length given the utterance label the segment does not depend on the length of the topic
$p(d_{i_j}|f(\text{length}(d_i))_j, \text{length}(d_i)) = p(d_{i_j}|f'(d_i)_j)$

If those assumptions are applied the model can be rewritten as:

$$p(d_i|\text{length}(d_i)) = \prod_j \frac{p(f'(d_i)_j|d_{i_j})}{p(f'(d_i)_j)} \cdot \prod_j p(d_{i_j})$$

Since $p(d_{i_j})$ is independent of the segmentation L it can be ignored in the search procedure. The score of the model is therefore just the probability of the region label given the utterance divided by the prior of the region label.

The advantage of this approach is that it extends boundary classification to the classification of multiple regions. It is particularly useful if we assume that simple regions of topics have different properties which may provide a natural model of prosodic and stylistic differences across regions.

5.6 Experiments

5.6.1 Databases

The experiments were carried out on the CallHome Spanish, a corpus of meetings and the Santa Barbara corpus (see Sec. 1.4.3 for a more detailed description of the corpora). In Fig. 5.2 the length distribution among these corpora are shown and it can be seen that 300 turns seems to be a reasonable upper bound which is rarely reached. All experiments were carried out on manual transcripts unless noted otherwise – only for a subset of the meeting corpus speech recognition results have been available. Additionally to the dialogue databases a written text database (C99) was used which has been presented by Choi (2000) for the comparison of segmentation algorithms. He used the Brown corpus to generate an artificial topic segmentation problem by randomly grabbing initial portions of Brown corpus texts and concatenating them as the topics of an artificial text. The resulting corpus is publicly available (see Choi (2000)).

The database consists of four subparts: 100 texts with topics 3–5, 6–8 and 9–11 sentences in length and 400 texts with topics of 3–11 sentences in length. The "all" database is the concatenation of these four databases.

5.6.2 Lengthmodels

Tab. 5.1 shows dialogue act segmentation with different length models for the CallHome Spanish database using a word repetition model [5]. The difference between the cheating length model and the best non-cheating model is not very large but the baseline performance is obviously a lot lower. Since the average length model fixes the number of segments independent of the actual segmentation model the comparison is more neutral and the average length model is therefore used by default.

5.6.3 Algorithm Comparison

Besides a pleasing theory any algorithm should perform well on practical standard tasks and it was compared with keyword repetition and speaker initiative modeling as "coherence features" to other standard procedures. The stopwords were removed from all databases and the first four letters were retained from each word for all databases. To allow direct comparisons with Choi (2000) the C99-database Porter stemming (Porter, 1980) was used instead of the 4 letter stemming and the performance measure was adapted to Choi (2000) (Sec. 5.3) for the calculation of the link error rate was used Choi (2000). The following algorithms were available:

[5] The *4 letter* text normalization without stopwords is being used, see below in this section.

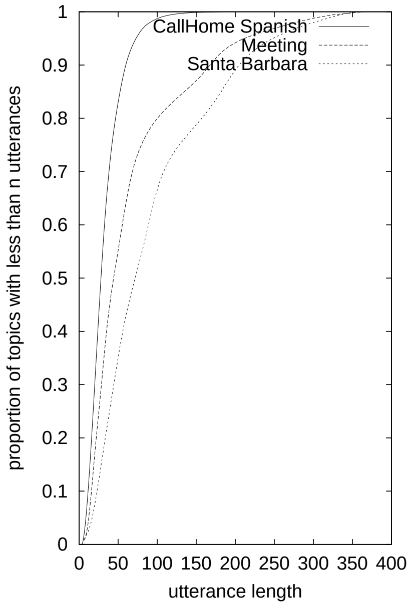


Figure 5.2: **Lengths of Dialogue Topics:** The length of topical segments in turns varies significantly between genres while the instructions for segmentation where fairly restricted and consistent. The length of the topics for CallHome Spanish – where the speakers know each other very well – is much shorter than in our meetings and in the Santa Barbara corpus. One could also assume that the use of the telephone might shorten the topics, that the 20min maximum on the phone called played a role or that close relatives can mention topics in a very brief manner.

Segmentation model	link-error rate in %
Cheating	35.5
Explicit length model	37.5
Average length	37.9
Penalty	39.4
Baseline	45.9

Table 5.1: **Segmentation with Different Length Models:** The dialogue segmentation performance is measured on the CallHome Spanish database using the *4 letter* word normalization and the word repetition model.

R01 the probabilistic algorithm proposed here.

C99 the algorithm by Choi (2000) who compared his algorithm to many others on the C99 database. His algorithm emerged as far the best in his comparison. Since results directly comparable to Choi (2000) have been provided the other algorithms in his study can also be compared with (Kan et al., 1998; Kozima, 1993; Reynar, 1998).

Tile the algorithm of Hearst (1997) has been implemented by Hearst (TileH), Choi (2000) (TileC) and the author (TileR).

R01, C99 and Tile were compared against each other on different datasets (Tab. 5.2). The native stopping criteria have been used for all algorithms; the native criterion for R01 is the average length criterion for the dialogue database and the penalty criterion for the C99 database task. To present Tile in the best light TileH was chosen on the dialogue segmentation tasks and TileC on the C99 database [6]. For speaker initiative each utterance was replaced by a single token representing the speaker identity and the information whether the speaker turn was long or short (a turn was defined as short if it contains three words or less).

The results show that the new algorithm delivers excellent results on the dialogue databases: The results are always better than the other algorithms, in some cases by large margins [7]. The only exception is the speaker initiative criterion for the CallHome database which may be a bad example since speaker initiative is likely a bad criterion for that database (see further discussion below). Surprisingly Tile and C99 seem to perform similar on the dialogue databases which may suggest that C99 has been tuned on its own database.

The results are available in more detail for the dialogue segmentation (Fig. 5.3) and C99 segmentation task (Fig. 5.4). Those tables also contain a comparison of the different stopping criteria and the results for the different implementations of Hearst (1997) that were available. The stopping criteria that are available are the native criterion for each algorithm as well as the criteria average length, cheating and choif discussed in Sec. 5.5.1.

[6] Note that the same stemming algorithms were used for all procedures – Choi (2000) didn't use the Porter stemming in the tiling implementation while it was used for his own algorithm. The author replaced the stopword removal and stemming from the external algorithms and replaced them with his own implementation, in this case the *4 letter* stemming. The native implementation of the Porter algorithm of C99 delivered identical results to the reimplementation used here. Since the stopping criterion of TileH could not be influenced only the "native" criterion was available for TileH.

[7] As can be seen below results for Tile and C99 improve when their native criterion for determining the number of segments is replaced by the average length criterion presented here – however R01 still performs better.

	Link error rate in %			
Database	R01	C99	Tile	Baseline
Dialogue segmentation, keyword repetition				
SantaBarbara	39.0	53.7	49.2	49.0
CallHome	37.9	40.4	44.1	45.9
Meetings	37.6	45.2	44.6	47.8
Dialogue segmentation, speaker initiative				
SantaBarbara	35.3	41.6	42.3	49.0
CallHome	45.5	43.8	43.0	45.9
Meetings	38.9	39.7	39.7	47.8
C99 database, keyword repetition				
All	13.8	12.8	30.4	42
3-11	13.6	13.0	29.9	45
3-5	17.2	17.7	36.7	38
6-8	8.9	9.6	26.8	39
9-11	16.1	10.0	29.7	36

Table 5.2: **Algorithm Comparison:** The proposed algorithm (R01) is compared to Choi (2000) (C99) and Hearst (1997) (Tile) for keyword coherence and speaker initiative based topical segmentation. The equal distance baseline (baseline) is listed for comparison. R01 delivers excellent results on all databases but the C99 database where is slightly worse than C99. The results on the C99 database have to be taken with a grain of salt due to the artificial construction of the database. It also appears that the stopping criterion plays a major role on that database (Tab. 5.4).

It can be observed that the average length criterion offers significantly better results for dialogue segmentation using the C99 procedure and the various implementations of Tile – however the probabilistic approach still typically performs better than the other algorithms, sometimes by a large margin.

The results for the C99 database are different, the C99 and R01 algorithms perform similar with the exception of the 9–11 part of the database where C99 performs a lot better. The situation changes if the algorithm for determining the number of segments is changed: If the number of segments for R01 is chosen to be the number of C99 the result of R01 is not much worse. If both algorithms are given the number of segments (cheat) from the reference R01 performs better. As noted above R01 worked a lot better on the C99-database using the *penalty criterion* unlike the *average length criterion* used on the dialogue databases. Given these results the author cautions the interpretation of the results on the C99-database since it has been artificially constructed. Specifically the length distribution of the segments seem to be unnatural and may place too much weight on the algorithm determining the number of segments. Overall R01 is slightly worse than C99 on this database yet much better than Tile and the other algorithms tested in Choi (2000).

5.6.4 Coherence Features

Tab. 5.5 compares coherence features. For word repetition the following choices can be made: (a) should stopwords be modeled as well and (b) should a word be mapped onto some baseform (stemming). The inclusion of stopwords may model the speaker identity implicitly or it may model general speaking style. The stemming algorithms tested were *No mapping* which doesn't perform any stemming, the *4-letter* stemming which maps a word onto its first 4 letters and the *trigram* method which maps each word onto the trigrams that occur in it. The *4 letter* stemming seems to be effective. An attempt to use Porter stemming (Porter, 1980) on the English database did not show improved results. The *trigram* stemming may capture endearments or other morphological features in Spanish which may explain its effectiveness on CallHome. The inclusion of stopwords is typically improving the performance if speaker initiative is not modeled.

For speaker initiative each utterance can either be replaced by the speaker identity itself (*Speaker*) or the speaker identity plus the information whether the utterance was long or short (*Speaker+LS*). An utterance is called short if it contains three words or less. This definition is designed to capture the information whether a speaker issued a dominant dialogue act or a non-dominant dialogue act. Short utterance tend to be non-dominant dialogue acts such as backchannels or answers. A strong correlation of dominance and the dialogue act type has been shown empirically by Linell et al. (1988) and the results indicate that the *Speaker+LS* feature performs significantly better than the *Speaker* feature by itself. The long/short criterion has the advantage that it may be implemented easily without having access to a speech recognition engine [8].

The speaker initiative approach doesn't seem to be very successful on CallHome Spanish. The reason for that fact may be seen in the familiarity of the speakers and their established (dominance) relationship as well as in the fact that one speaker is abroad whereas the other is “back home”. Both properties may lead to dialogues where the dominance is rarely shifting between speakers. For the multi-party dialogues in the Santa Barbara and meeting corpus however speaker initiative outperforms the keyword based approach. On meetings the combination of the two delivers the best results.

[8] Other encodings of speaker initiative did not improve the results. In prestudies speaker initiative indicators based the dialogue acts have been used. However they produced worse results than the long/short distinction.

	Link error rate in %				
Stopping criterion	R01	C99	TileR	TileC	TileH
Dialogue segmentation, 4 letter stemming					
Santa Barbara (baseline 49.0%)					
native	41.4	53.7	58.4	51.6	49.2
cheat	38.3	46.6	43.4	41.8	-
ave	39.0	48.0	44.7	42.3	-
CallHome Spanish (baseline 45.9%)					
native	39.4	40.4	40.7	44.3	44.1
cheat	35.5	36.7	38.6	43.4	-
ave	37.9	39.1	40.8	45.1	-
Meetings (baseline 47.8%)					
native	36.4	45.2	53.8	47.8	44.6
cheat	31.6	36.1	43.9	41.8	-
ave	37.6	37.5	46.9	42.5	-
Dialogue segmentation, Speaker+LS					
Santa Barbara (baseline 49.0%)					
native	40.8	41.6	56.3	42.4	42.3
cheat	33.6	41.9	42.2	44.4	-
ave	35.3	41.9	41.8	44.7	-
CallHome Spanish (baseline 45.9%)					
native	47.2	43.8	43.1	42.4	43.0
cheat	45.0	44.1	40.9	45.5	-
ave	45.5	44.4	42.9	47.5	-
Meetings (baseline 47.8%)					
native	41.1	39.7	52.8	44.7	39.7
cheat	39.1	39.2	40.6	44.6	-
ave	38.9	39.2	46.0	44.6	-

Table 5.3: **Dialogue Segmentation with Different Algorithms:** A detailed comparison of segmentation performance is presented and for each database all algorithms are tried in conjunction with several stopping criteria. The algorithms used are the newly introduced probabilistic algorithm (R01), Choi (2000) original implementation with a new stemming algorithm (C99) and implementations of Hearst (1997) by myself (TileR), Choi (TileC) and from Hearst's WWW site (TileH). The results are compared under different stopping criteria and the average length criterion seems to perform very well.

	Link error rate in %				
Stopping criterion	R01	C99	TileR	TileC	TileH
C99 database					
All					
native	13.8	12.8	43.9	30.4	45.2
cheat	10.2	11.2	43.8	47.9	-
choif	13.8	12.8	43.8	47.8	-
seg	15.8	15.7	43.8	47.8	-
3-11					
native	13.6	13.0	43.3	29.9	45.4
cheat	12.0	12.5	43.3	47.8	-
choif	13.7	13.0	43.3	47.8	-
seg	13.9	14.1	43.3	47.9	-
3-5					
native	17.2	17.7	46.1	36.7	46.2
cheat	10.1	10.4	46.1	47.0	-
choif	17.5	17.8	46.2	46.4	-
seg	26.5	28.5	46.1	46.0	-
6-8					
native	8.9	9.6	45.7	26.8	45.2
cheat	7.4	9.5	45.6	47.9	-
choif	10.6	9.6	45.6	47.8	-
seg	8.0	9.8	45.6	47.9	-
9-11					
native	16.1	10.0	42.1	29.7	43.7
cheat	5.7	8.6	41.7	49.3	-
choif	13.6	10.0	41.7	49.1	-
seg	20.3	15.2	41.7	48.7	-

Table 5.4: **C99 Segmentation with Different Algorithms:** Similar to Tab. 5.3 segmentation algorithms are compared on the written text database C99. Additionally to the stopping criteria presented there the number of segments created by the C99 algorithms (choif) was also used as a stopping criterion.

	Link Error Rate in %					
Features	None	4 letters		No mapping		Trigram
Stopwords		No	Yes	No	Yes	Yes
Santa Barbara (baseline 49.0%)						
		39.0	38.6	41.1	41.0	43.8
Speaker + LS	35.3	38.5	38.7	35.9	41.8	41.8
Speaker	39.1	36.4	39.4	36.2	40.5	41.7
CallHome Spanish (baseline 45.9%)						
		38.6	38.4	39.4	38.6	37.2
Speaker + LS	45.6	38.8	39.6	39.3	38.3	37.3
Speaker	45.3	38.4	38.3	39.1	38.8	37.2
Meetings,topic segmented database (8 meetings) manual transcript (baseline 47.8%) Second human 32.3%						
		37.6	33.1	37.6	34.3	35.3
Speaker + LS	38.9	35.6	33.1	36.9	32.9	33.7
Speaker	42.6	36.0	32.9	37.9	33.8	34.9

Table 5.5: **Speaker Initiative and Keyword Repetition for Topic Segmentation:** Topical segmentation was tested on the Santa Barbara corpus, CallHome Spanish and the meeting corpus, all corpora manually transcribed and annotated with speakers. Two types of features are being compared, keyword repetition and speaker repetition.

5.6.5 Speech recognition

The effect of speech recognition on the segmentation accuracy is demonstrated in Tab. 5.6. While the result for keywords is worse using speech recognition it is not as bad as one might assume. This result may also be due to consistent misrecognitions that might be produced by a speech recognizer due to keywords that are missing from the vocabulary. Using stopwords additionally to words resulted in a significant improvement in the link error rate with no degradation introduced by the speech recognizer. Speaker initiative can be used by itself and it can be combined successfully with word and stopword information. The results have to be taken with caution due to the small size of the database available and the manual annotation of speaker identity. It should be noted again that the determination of speaker identity is often reliable and inexpensive (Sec. 2.4.5.4).

Meetings, LVCSR database		
	Link error rate in %	
Features	manual	machine
baseline	42.4	42.1
words, no stopwords	34.3	38.7
words+stopwords	32.5	35.2
speaker+ls	36.9	36.5
speaker+ls and words, no stopwords	34.0	39.4
speaker+ls and words+stopwords	30.4	33.6

Table 5.6: **Effect of Speech Recognition Errors:** Two of the meetings have been fully decoded by an LVCSR system with a word error rate of approximately 40% (Waibel et al., 2001a). The *4 letter* word normalization has been used (see Tab. 5.5).

5.6.6 Region Features

Part of speech can encode the general style of a segment as was seen in the activity detection task (Sec. 4.2). Typically one would assume that the beginning of a topic introduces the new objects that are being talked about, they are elaborate in the middle segment and in a final segment we'll find conclusions, a lack of interest to continue that discussion and attempts to change the subject. If these are valid assumptions the part of speech distribution should be different in those regions, for example one would assume nouns in the beginning to refer to objects and pronouns in the middle and end. The results of segmentation based on part-of-speech features for region modeling is shown in Tab. 5.7. A neural network classifier was trained without hidden units, the softmax output function was used as the output function. The vocabulary for the neural network (NN) and language model (LM) classifier were the most frequent 500 word/part of speech pairs while the remaining words are mapped on their part of speech. The effects are clear especially for the equal size+boundary region model and the improvements can also be confirmed when combining the model with repetition modeling, especially on CallHome Spanish. It is therefore clear that there are changes in the word and part of speech distributions in different topical regions. However the combination of word based region modeling with the best repetition model didn't always yield better results for the other databases. Neural network performed significantly better than language models as region classifiers on some segmentation tasks but are slightly worse on some others.

	Link error rate in %						
Coherence	No	Segmentation					
feature	region	boundary		equal size		both	
		NN	LM	NN	LM	NN	LM
CallHome Spanish							
none	45.9	43.4	42.6	38.3	39.3	36.5	37.8
keyword	38.6	35.8	34.1	36.2	35.6	34.7	33.6
speaker+ls	45.6	42.8	42.3	43.2	42.2	41.9	41.4
both	38.8	36.5	34.3	37.4	35.7	35.6	34.9
Meeting							
none	47.8	38.9	37.4	42.1	45.1	41.5	46.0
keyword	37.6	36.2	37.9	37.9	36.9	37.1	39.5
speaker+ls	35.6	38.0	36.7	40.7	36.9	39.7	40.1
both	36.0	37.7	36.1	35.6	36.3	36.6	35.8
Santa Barbara							
none	49.0	40.1	41.0	43.4	48.9	44.1	48.2
keyword	39.0	40.0	38.1	39.7	40.4	39.5	40.8
speaker+ls	38.5	37.7	38.0	38.8	42.9	38.9	41.0
both	36.4	36.2	38.9	37.5	38.7	37.4	37.1

Table 5.7: **Segmentation Using Word Based Regions:** A neural network (NN) and language model classifier (LM) were trained to discriminate between different regions of topical segments, either just boundary vs. non-boundary, equal sized regions (begin/middle/end) or a combination of the two (both). The table shows the combination of these features with keyword repetition (keyword) and speaker initiative (speaker+LS) and the combination of the two (both).

	Link error rate in %			
Repetition	No	Boundary	Equal size	Both
pitch				
none	45.9	44.8	46.2	46.2
keyword	38.6	38.1	38.2	38.2
speaker+ls	45.6	43.9	45.2	45.2
both	38.8	38.2	38.2	38.1
pause				
none	45.9	48.5	48.9	47.0
keyword	38.6	37.9	37.4	37.6
speaker+ls	45.6	45.1	45.8	45.6
both	38.8	38.2	38.0	38.0
pitch+pause				
none	45.9	44.8	46.5	46.0
keyword	38.6	37.9	38.3	38.2
speaker+ls	45.6	44.2	44.9	44.8
both	38.8	38.0	38.4	38.1
pitch+pause+words				
none	45.9	44.3	41.2	40.3
keyword	38.6	35.0	36.1	34.9
speaker+ls	45.6	42.7	44.0	42.0
both	38.8	35.6	36.4	35.5

Table 5.8: **Segmentation Using Prosodic Regions:** The experiments were carried out on the CallHome Spanish database and the text normalization for the repetition model was the *4 letter* mapping without stopwords, the setup is similar to Tab. 5.7.

Alternative features for segmentation algorithms that are widely used according to the literature are (Shriberg et al., 2000, p. 4)

> [...] some combination of a long pause, a preceding final low boundary tone, and a pitch range reset, among other features

However there are few studies that actually present results that are based solely on automatic methods (Hirschberg and Nakatani, 1998; Shriberg et al., 2000). Two proven features are pause and pitch range reset. Pitch range reset can be modeled by the *equal size region* approach since the pitch range should be large in the beginning, smaller in the middle and really small in the end of the segment. Pauses are more likely in the beginning of a segment or at the end of a segment. The features used are a histogram of F_0 which has been obtained using the fundamental frequency derived using a pitch tracker (Schubert, 1999) and the histogram of pause lengths. Tab. 5.8 however does not demonstrate good results for these prosodic features in segmentation, although exactly the same implementation was used as for the emotion detection (Sec. 3.8). The reason might be genre difference since the speakers might not necessarily plan the end of a topic or may not mark it – the free flow is merely resulting in a topic change while an anchor speaker in a TV newsshow has a very good idea when the topic changes and needs to announce it as good as possible.

5.7 Conclusion

A probabilistic framework for dialogue segmentation is presented and applied. The algorithm proposed has a clean probabilistic interpretation and performs well compared to Choi (2000); Hearst (1997), especially on dialogue databases. There is still room for improvement, especially information about cooccurence of words could be included in the model as suggested by Beeferman et al. (1999); Ponte and Croft (1997); Yamron et al. (1998) and more work on prosodic features could be attempted. The algorithm was tested on a variety of spontaneous speech corpora and (stemmed) keywords, character n-gram and speaker initiative were used as features. Speaker initiative performs very well compared to keyword repetition: This finding confirms the intuition that topical change is correlated with the activity the speakers engage in and their speaking rights. The results however also show that speaker initiative may fail in certain situations such as CallHome Spanish where only one speaker is dominant while the topic may be changing. Determining speaker initiative according to the definition here should be tractable since speaker identity may be available trivially or it can be determined very effectively and reliably in meeting situations (Pan and Waibel, 2000). Modeling begin/middle/end as well as the boundary of a topical segment it was possible to exploit changes in the word and part of speech distribution within a topic in order to do topic segmentation.

Dialogue segmentation can therefore be done with a couple of features with similar performance. These features include keyword repetition, speaker initiative and changes in the part of speech profile. The results presented here therefore fit the general claim that dialogue style has to be an important feature in information access systems for spoken interactions. Speech recognition – even on hard corpora – didn't have a disastrous impact on the segmentation performance but resulted in significant degradation. Speaker initiative is a very powerful criterion which can likely be detected reliably without the need for expensive LVCSR. Speaking style might also be useful and the speech recognition problem that needs to be solved to find the part of speech distribution in a segment might not be as severe as the problem of finding keywords. Additionally the new probabilistic algorithm did not only provide a unique framework for comparing these experiments, it also proved to be a very effective approach, both regarding the results and computational efficiency.

Chapter 6

Information Access

6.1 Introduction

Information access to meetings is a very hard problem and requires a tremendous collection of technologies to be developed. While a final evaluation of this collection of technologies in the intelligent meeting room (Sec. 1.2) should be the ultimate goal the research needs to be evaluated by more immediate measures that can be calculated now and without confounding factors. This is done in this chapter using information theoretic measures, a user study and suggestions for the implementation of user interfaces for audio records.

A simple example motivates the information theoretic approach. We may assume that a classification of the rejoinders in X databases is given with high accuracy. We may additionally assume that for each query the user would know the right database and that all databases are equally likely to be requested. Taking the database information into account in this situation reduces the search space by a factor of X. One may be able to achieve the same reduction using different types of user interfaces such as a query based interface, a (hierarchical) browsing interface, a graphical skimming interface or an audio interface. The estimation of search space reduction using database membership information is therefore rather simple, especially since the automatic detection is also very easy.

To be able to quantify more complex and less obvious reductions an approach based on information theoretic measures is taken in Sec. 6.2. In the case cited above we would see the maximal reduction possible from X labels, namely a $-\log_2 X$ bit reduction in search space. The information theoretic approach allows to measure correlations between different types of factors and quantify more complex situations including indexing using multiple features. The information theoretic approach also allows us to draw conclusions about the relationship of keywords and speaker identity with non-keyword based features.

One of the hardest problem is to understand how users judge different non-keyword based features for indexing, which ones they prefer and how they integrate different sets of information. Users were asked to identify which one out of 4 similar rejoinders a set of short audio samples belongs to. Since some of the rejoinders are very similar the selection problem used may also be relevant for within-rejoinder navigation. It turns out that the presence of activity is crucial. Speaker identity, formality, and dialogue acts did not have a measurable effect on the results (Sec. 6.3).

Every information retrieval system has a user interface and the quality of the user interface may have a huge impact on the effectiveness the system. The non-keyword based features and audio records presented in this thesis haven't been studied extensively in the past such that user interfaces aren't fully developed yet. The access to large sets of rejoinders will require a query interface which may take the (sub-)database, restrictions on the time, participants, style and topic into account similar to standard techniques in IR (textual entry, buttons, menus). However, interactive search strategies such as browsing and skimming in rejoinders or small (sub-)databases might be different, especially since audio is a linear medium and it can take a long time to play back a rejoinder. Written information however can be skimmed visually and most people are used to textual access. A couple of alternatives are discussed for audio records: Excerpts, rapid playback, and graphical / textual representations (Sec. 6.4).

This chapter therefore discusses the information theoretic approach (Sec. 6.2). The information theoretic approach offers a variety of results on the search of rejoinders in databases, the correlation of style and topic, and more results from analyzing the available corpora. The user study (Sec.6.3) allows us to evaluate which types of features are useful for human in a retrieval task. A discussion of potential access interfaces (Sec. 6.4) investigates how non-keyword based features could be used in different types of user interfaces. Conclusions to this chapter are offered in Sec. 6.5.

6.2 Information Theoretic Approach

6.2.1 Introduction

This section introduces the information theoretic approach. Sec. 6.2.2 provides the background and describes the relationship of search space reduction with other measures. Sec. 6.2.3 describes how the prior of an information retrieval model may be changed to reflect user preferences. One of the most crucial confounding factors of this work might be the implicit correlation of speaker identity with speaking style and is discussed and quantified in Sec. 6.2.4. The search space reductions for rejoinders and segments in rejoinders are presented in Sec. 6.2.5 and conclusions are offered in Sec.6.2.6.

6.2.2 Background

This subsection introduces the necessary tools to quantify the reduction in search space using information theoretic measures; for a more general introduction to information theory see Cover and Thomas (1991). A probabilistic information retrieval model $q(D|R)$ is assumed where a query R selects a segment D of a conversation. The distribution may be used to represent the results effectively e.g. by ranking the segments in a list or highlighting them in a graphical representations. The user interface however is not a consideration of this section and it is assumed that there is an effective user interface to make use of q.

Independent of the retrieval model — which is a mechanical implementation of a retrieval engine — a distribution p of segments D and queries R has to be assumed. Obviously it is desirable to model q after p as much as possible. The quantity of interest is the reduction of the expected log-likelihood of the dialogue segment D if the query is given which lends itself to an application of Bayes' rule

$$\mathrm{E}_p\ \log_2 q(D|R) - \mathrm{E}_p\ \log_2 q(D) = \mathrm{E}_p\ \log_2 q(R|D) - \mathrm{E}_p\ \log_2 q(R)$$

If the query is a discrete random variable with a small range the probabilistic information retrieval model can adapt very well to the target distribution p such that $-\mathrm{E}_p \log_2 q(R) \approx \mathrm{H}(R)$. As a special case the reduction is $\mathrm{H}(R)$ if r can be read off the dialogue directly and the user would always issue the correct query (the query could be a function of the segment D such that $\mathrm{E}_p \log_2 q(R|D) = 0$). If on the other hand $q(R|D)$ needs to be estimated (for example using a neural network) the formula is used directly to determine the quantity. Another simple case is when q can be estimated very reliably since R and D are discrete random variables over a small set. We may then assume that $p = q$ such that the reduction in search space is exactly the mutual information between D and R.

If two variables can be used to find the segment D of a conversation one may ask how much effect that may have. Formally we want to know how much less the combined reduction

$$\mathrm{Red}_{\mathrm{comb}} = \mathrm{E}_p\ \log_2 q(D|R_1, R_2) - \log_2 q(D)$$

is than the sum of the individual reductions

$$\mathrm{Red}_{\mathrm{i}} = \mathrm{E}_p\ \log_2 q(D|R_{\mathrm{i}}) - \log_2 q(D)$$

The difference is

$$\mathrm{Red}_{\mathrm{comb}} - \mathrm{Red}_1 - \mathrm{Red}_2 = \mathrm{E}_p\ \log_2 \frac{q(R_1, R_2|D)}{q(R_1|D)q(R_2|D)} - \mathrm{E}_p\ \log_2 \frac{q(R_1)q(R_2)}{q(R_1, R_2)}$$

If we assume R_1 and R_2 to be conditionally independent given D [1] and we assume that q is almost the same as p for $q(R_{1,2})$ and $q(R_1, R_2)$ then the combined query results in a reduction given by the sum of the individual queries minus the mutual information of the two variable $\mathrm{MI}(R_1, R_2)$. The mutual information of the random variables does not only quantify their correlation but also the information access performance of the combined query.

6.2.3 Changing the Prior

Features don't have to be explicitly queried or selected by a user if it is obvious by contextual constraints that one database or one activity is more likely to be of interest to the user than others: If we make recordings in a meeting room that are continuous we do not want to retrieve the recordings where the meeting room was not used or a random conversation took place. Similarly we may prefer an in-depth report over a conversation on the street until we specify otherwise, even though there might be significantly more different random conversations on the streets on a given topic and only one in-depth report. Natural constraints that arise from detecting a certain desirable or undesirable style or that can be inferred from user preference can therefore be included into the model. If we have a desired a-priori distribution over the documents q' we can define a new retrieval model

$$q''(D,R) := q(D,R) \cdot \frac{q'(D)}{q(D)}$$

This model fulfills the desirable properties:

- $q''(D,R) = q(D,R) \cdot \frac{q'(D)}{q(D)} = q(R|D)q'(D)$
  $\leadsto$ q'' is a probability distribution
- $q''(R|D) = q(D,R) \cdot \frac{q'(D)}{q(D)} / q'(D) = q(R|D)$
- the marginal of q'' on D is q'

6.2.4 Speaker Identity and Speaking Style

Speaker identity is an important feature for information access since we may assume that it is remembered by a user who took part in a conversation or it may be a

[1] The conditional independence is $q(R_1, R_2|D) = q(R_1|D)q(R_2|D)$ or in other words the independence of two queries *given* the rejoinder. In many cases the queries are *functions* of the segment and even if they are not functions two different types of queries (keywords, activity) for a given document are likely independent. Queries R_i need to be very similar in nature in order to be conditionally dependent. The two queries however are not independent in general ($q(R_1, R_2) = q(R_1)q(R_2)$) since, for example, the topic of a conversation (R_1) might be correlated with the activity (R_2) in the conversation.

Features	Accuracy in %	Entropy in bit
Pick most likely speaker		
	31.1	3.52
Language model classifier		
unigram, words	46.9	3.14
bigram, words	52.5	5.06
dialogue act	31.3	3.53
Neural network		
dialogue act	32.3	3.35
words	41.9	4.10
WordNet	31.2	3.48
Biber	31.1	3.50
Biber, WordNet, dialogue acts	33.7	3.32
Biber, WordNet, dialogue acts, words	41.0	3.03

Table 6.1: **Speaker Detection Using Speaking Style:** Speaker detection was done on the meeting database where speakers participated in multiple meetings. The features used consisted of the most frequent 500 word / part of speech pairs and the parts of speech for the others, detected dialogue acts, WordNet classes and Biber features. Speaker identity is determined best using word level information.

good indicator which type of meeting took place. However speaker identity might also be correlated to speaking style such that unwanted correlations may be measured. If, for example, some microlevel feature is an indicator for variable R and for speaker identity but speaker identity is also a strong indicator for variable R as well we have to decide whether we want to attribute the effect of the microlevel feature to the speaker identity or to variable R directly. Sec. 4.2.2 mentions some of the safeguards applied. The purpose of this section is to quantify the strength of the speaker identity effect.

Indeed disciplines such as stylometry and authorship attribution build on the correlation of a writers identity and the stylistic features exhibited in a text. Authorship attribution is useful for example in literature studies when not enough historical evidence is present Holmes (1998). Spoken language is a realtime achievement such that deliberate shifts in style might be harder to achieve than in written language. It is therefore likely that microlevel features show patterns specific to speakers which might confound experiments building on microlevel features:

Tab. 6.1 shows that speaker identity can be detected fairly well using word and part of speech features alone, the bigram language model classifier delivers by far

Features	Accuracy in %	Entropy in bit
Pick most likely meeting		
	24.1	3.19
Neural Network Classifier		
Words	44.7	2.92
Speakers	73.0	2.31
Speakers, detected from words using neural network	19.1	3.62
Speakers, detected from words using bigram language model	35.5	3.37
Language Model Classifier		
Speakers	83.0	102.1

Table 6.2: **Rejoinder Identity Detection Using Speaker Identity:** The meeting database featured speakers that took part in multiple meetings. The meeting are indexed very well by speaker identity but word features also work. Indeed one can detect rejoinder identity by first detecting the speaker identities and based on that statistics detect the meeting identity. The word feature was a histogram of the most frequent 500 word / part of speech pairs and the parts of speech for the less frequent pairs.

the best performance in terms of detection accuracy. Additionally it was tested whether dialogue and style information would index speakers but it seems that dialogue acts, Biber and WordNet categories do not correlated much with speaker identity.

Tab. 6.2 shows that speaker identity indexes meeting identity very well. Word level features (which are also used to detect the speaker identity) are not quite that important. If we however detect speaker identity using word level features and detect the dialogue identity from those speaker identities similar results can be obtained as the detection accuracy from words themselves. This result indicates that the for the selection of rejoinders microlevel features such as word and part of speech distributions are in large influenced by speaker identity. Dialogue acts, WordNet and Biber features however do not index speaker identity the same way such that the results based on them can be interpreted more reliably. In the following Sec. 6.2.5 and especially Tab. 6.7 this will be important to notice.

Database	Entropy in bit	
	rejoinder	segment given rejoinder
CallHome Spanish	7.0	3.4
Meetings	3.2	3.9
Santa Barbara Corpus	2.9	3.8

Table 6.3: **Rejoinder Identity and Segment Identity:** The size of the database influences the entropy of the rejoinder identity while the segment given the rejoinder identity fluctuates between 3 and 4 bit (or approximately 8 and 16 segments). It should be noted that the CallHome Spanish database contains calls of maximally 20 minutes each while most of the other rejoinders where about an hour long. This explains why the entropy for CallHome is somewhat lower.

6.2.5 Finding Rejoinders and Topical Segments in Rejoinders

6.2.5.1 Introduction

The problem of finding a topical segment of a rejoinder in a database can be divided into two aspects, namely finding the rejoinder and given the rejoinder finding the actual segment. These two aspects may be solved best by different features such that it is interesting to ask which ones they are. Since the segments are assumed to be equiprobable their entropy in bit is simply $-\log_2 \#\mathrm{segments}$, the entropy of the rejoinder identity can be estimated from data and the entropy of the segment given the rejoinder is their difference (Cover and Thomas, 1991). The basic picture can be seen in Tab. 6.3. The search space aspect of finding the rejoinder depends on the database size while the number of topics per rejoinder seems to be rather similar, resulting in similar entropies.

The effort in Sec. 4 has been to detect features of dialogues (the (sub-)database detection problem) or to detect activities. To detect those, microlevel features have been consulted (Sec. 4.2) and neural networks have been trained as classifiers. If the features that are being used are effective to make that determination they can be turned into a retrieval model for topical segments and indeed the reduction in search space can be measured exactly in terms of the entropy reduction these classifiers achieve (Sec. 6.2.2). The retrieval model therefore selects documents with a certain activity label (the output of the classifier) given a set of microlevel features (the input of the classifier). Since activity is very hard to detect though there isn't a large reduction in entropy that could be measured, especially since exact probabilities are hard to estimate on small datasets such as Meeting and Santa Barbara.

However the same classification approach may also be used to find dialogues:

Given a segment in a dialogue we want to know which dialogue it belongs to. Why is that question important? As pointed out in Sec. 6.2.2 the expected reduction in search space for the rejoinder identity can be estimated by

mutual information If the query R and the the rejoinder identity D are discrete variables over a small range the reduction in search space can be estimated directly (Sec. 6.2.5.2).

neural network If the query R is complex but can be represented as a vector the reduction in search space can be estimated as the entropy of the rejoinder identity minus the negative log likelihood of the rejoinder identity output of a classifier (Sec. 6.2.5.3).

6.2.5.2 Direct mutual information estimation

Leaving-one-out estimates can be used to estimate probability distributions and information theoretic measures. These estimates tend to be fairly accurate in practice even if the database is relatively small (Ney et al., 1994). The measurement is achieved by measuring the average likelihood of a certain event if the training database is the whole database excluding the event that is currently being evaluated. It is therefore an n-fold cross-validation technique where n is the size of the database. In many situation (like the estimation problem here) the leaving-one-out estimate can be stated in a simple and efficient closed form.

This method may be applied to the estimation of the mutual information of rejoinder identity with activity and other similar discrete random variables over small ranges (Tab. 6.4). The mutual information is approximately $1\,\mathrm{bit}$ for all databases (Tab. 6.4) and may be interpreted as the reduction in search space for rejoinder identity if the activity is known.

The reduction in search space of a feature is the entropy of the feature if the user makes the same judgments about the feature as the retrieval system (Sec. 6.2.2), certainly an idealization for activity and topic annotation (Sec. 4.5.1, Tab. 4.8). However if that idealization is assumed the entropy of the activity given the rejoinder identity

$$\mathrm{H}(\mathrm{Activity}|\mathrm{RejoinderId}) = \mathrm{H}(\mathrm{Activity}) - \mathrm{MI}(\mathrm{Activity}, \mathrm{RejoinderId})$$

measures the search space reduction of the activity for a segment in a rejoinder if the rejoinder is known. This quantity is significant for meetings and the Santa Barbara Corpus (1.7 resp. $1.6\,\mathrm{bit}$) but much lower for CallHome Spanish ($0.6\,\mathrm{bit}$). This indicates that there is a significant activity variation with rejoinders which could be exploited for within-rejoinder search. CallHome Spanish however does not display such great changes in activities which may be due to the close family

	Entropy	Mutual Information		
	in bit	in bit		
		rejoinder id	activity	topic
CallHome Spanish				
rejoinder id	7.0		1.2	1.5
activity	1.8	1.2		0.4
topic	2.3	1.5	0.4	
CallHome Spanish, conditioned on rejoinder				
activity	0.6			0.0
topic	0.8		0.0	
Meeting database				
rejoinder id	3.2		1.0	
activity	2.7	1.0		
Santa Barbara corpus				
rejoinder id	2.9		0.8	
activity	2.4	0.8		

Table 6.4: **Correlation of Rejoinder, Topic and Activity:** The entropies were estimated using the leaving-one-out technique (Ney et al., 1994) and the mutual information is estimated via the entropy of the joined random variable. As can be seen there is some overlap of topic and activity indicated by a mutual information of ≈ 1 bit on all databases. The mutual information of topic and activity is $0\,\mathrm{bit}$ if the rejoinder is known.

ties of the callers. Similar observations about the low in-rejoinder variability of CallHome have also been made in the experiments which used speaker initiative for topic segmentation (Sec. 5.6.4).

On CallHome Spanish the topic distribution that was annotated as shown in Tab. 6.5. The annotation scheme was developed at the same time the activity annotation was developed. Furthermore Tab. 6.4 shows the correlation of annotated topic on CallHome Spanish with the rejoinder identity and the activity. Rejoinder identity and topic are highly correlated (mutual information $1.5\,\mathrm{bit}$) such that the topic given the rejoinder has only $0.8\,\mathrm{bit}$ of information. Topic and activity are also correlated with much lower mutual information, namely $0.4\,\mathrm{bit}$. There is no intercoder agreement available for this annotation, however, and the correlation of annotated topics and keywords seem to be low overall (Tab. 6.6). The results for topic given these manual annotations are therefore unclear but more results on the relationship of topical keywords and rejoinder identity will be obtained in the next section.

6.2.5.3 Neural Network Correlations

Finding the rejoinder is a task that can also be measured directly by building classifiers for rejoinder identity. This allows to quantify the contribution of that feature to rejoinder identity, both in detection accuracy as well as in the reduction of search space as measured by entropy (Sec. 6.2.2). The leaving-one-out estimates (Sec. 6.2.5.2) can be only applied to discrete variables over a small range while the neural network (resp. language model) based methods allows to use more features as indicators of rejoinder identity (Tab. 6.7). Both on meetings and on the Santa Barbara corpus the best results were achieved using word features and a neural network based classifier. In both cases a significant improvement was obtained by adding more words to the features, starting with 10 words moving to 1000 or 10000 words; most of the information is contained in the most frequent 50 words. Microlevel features based on words and parts of speech distributions however are difficult to interpret since they may encode speaker identity (Sec. 6.2.4). On the meeting corpus and the Santa Barbara Corpus only the most frequent words are effective and it has to be concluded that either speaker identity alone or speaking style is effective but not topic. On the CallHome corpus a reduction of $2\,\mathrm{bit}$ is obtained using the 100 most frequent word/part of speech pairs and parts of speech for all other words. However an additional $0.3\,\mathrm{bit}$ reduction could be obtained using many more words, probably due to topic information.

Topic	Count	Topic	Count
Well Being	753	Free telephone call	67
Travel	145	Education	58
Job	108	Foreign country	26
Health	93	Career planning	14
Money	80	Politics	9

Table 6.5: **Topic Annotation on CallHome Spanish:** The annotators were able to score each topic on a scale from 0 to 5 and the table shows which topics received the highest score.

6.2.6 Conclusion

This section uses information theoretic methods in order to estimate search space reductions that may be obtained different sets of features. Some of the assumptions made are optimistic but in the absence of a large user population instruments need to be developed that would be able to predict retrieval performance.

#words	Detection in %	Entropy in bit
10	52.0	2.45
100	51.1	2.55
1000	54.9	2.13
10000	57.2	2.05

Table 6.6: **Topic and Keywords:** The detection of the topic as annotated by human coders is fairly difficult. Picking the most likely topic (Well Being) results in a detection accuracy of 55.7, the entropy of the topic distribution is $2.27\,\mathrm{bit}$ Only a neural network classifier with a large number of keywords was capable of capturing some correlation between keywords and annotated topic on CallHome Spanish. Further experiments with stylistic features didn't show any improvement.

The first and most surprising result was that keyword based and topical information did not immediately come out as a good feature. While manually assigned "metatopics" in CallHome Spanish seem to make sense (only moderate correlation with activity but a significant potential search space reduction) these topics were hard to detect using keywords. Similarly the rejoinder identity was hard to assign using keywords and only a small search space reduction can be predicted.

Activities, on the other hand, may have – given assumptions made – a significant potential for search space reductions. The most optimistic estimate would be that the full entropy of the feature could be exploited which would result in a search space reduction of about 2.4 bit or approximately a factor of 5. A reduction of 1bit for finding the rejoinder was measured and a reduction of $\approx 1.6\,\mathrm{bit}$ for finding a segment given the rejoinder was measured on the meeting and Santa Barbara corpus. This number is a lot lower for the CallHome Spanish database, namely $0.6\mathrm{bit}$, since there doesn't seem to be much change in a rejoinder (see also Sec. 5.6.4). Additionally dialogue acts and games as well as WordNet features show promise for finding rejoinders in databases.

6.3 User Study

6.3.1 Introduction

At this point of the thesis we do know about a lot of different aspects of information access and on the detection of features which are non-keyword based. However there are a couple of aspects we haven't shed light on yet, especially we

Feature	Empirical entropy in bit			Classification accuracy in %		
	CHS	Meet	SBC	CHS	Meet	SBC
baseline	7.04	3.19	2.85	1.7	24.1	22.4
Words, ngram backoff model classifier						
10 words	7.21	50.0	3.27	4.7	29.1	53.3
100 words	7.21	111	3.37	4.7	29.8	52.3
1000 words	7.21	111	3.37	4.7	29.8	52.3
10000 words	7.21	111	3.37	4.7	29.8	52.3
Words, neural network classifier						
10 words	5.76	3.26	1.67	17.1	33.3	67.3
100 words	5.01	2.84	1.24	24.0	46.8	73.8
1000 words	4.75	2.93	1.18	29.0	44.0	79.4
10000 words	4.69	2.99	1.24	30.5	42.6	74.8
No words, neural network classifier						
Biber	6.35	3.59	2.80	5.8	26.2	21.5
WordNet	n/a	3.43	2.48	n/a	27.0	48.6
dialogue acts	5.96	3.49	2.69	10.6	27.7	42.1
dominance	6.95	3.49	2.75	1.3	27.0	39.3
dialogue games	6.64	n/a	n/a	3.9	n/a	n/a
dialogue acts + WordNet	n/a	12.3	2.00	n/a	36.2	57.0
50 words plus additional features, neural network classifier						
No additional features	5.26	2.92	1.29	21.3	46.1	76.6
Biber	5.18	3.37	2.75	21.9	33.3	26.2
WordNet	n/a	3.15	1.24	n/a	37.6	74.8
dialogue acts	4.95	2.92	1.37	25.7	42.6	75.7
dominance	5.26	2.95	1.27	21.4	39.7	78.5
dominance + Biber	5.18	3.22	2.75	22.0	32.6	26.2
dialogue acts + Biber	4.88	3.30	2.73	1.1	31.2	28.0
dialogue acts + WordNet	n/a	7.17	1.31	n/a	47.5	72.9

Table 6.7: **Empirical Entropy Reduction for Rejoinder Identity:** Meeting data and Santa Barbara corpus show that the most important information is in the most frequent 50 words, adding more words or other features is not making a (big) difference. This seems to indicate that these features are deeply confounded by speaker identity as discussed in Sec. 6.2.4. The results for the Santa Barbara corpus display a 0.85bit entropy reductions for dialogue act+WordNet features without consulting word level information. For the CallHome Spanish (CHS) corpus a reduction of 1bit can be obtained using dialogue act features alone and about 1 bit can be obtained from keywords (after discounting the reduction of the most frequent 10 keywords). The results for the meeting database have been published in part in Waibel et al. (2001a).

don’t know how users may

- integrate different types of knowledge
- use representations of non-keyword based knowledge
- use representations of microlevel features (formality/dialogue acts/speaker activity)

Additionally we don’t know how these features would improve an overall system and how appropriate some of the assumptions are that have been made. The most interesting type of study would obviously be to take the existing system and add non-keyword based indices and see how it performs. There are five reasons why this proposition would yield less reliable results:

effect of user interface The user interface issue for non-keyword based indices and audio records is far from being solved, this thesis only touches upon it in Sec. 6.4. Especially dynamic browsing interfaces featuring multiple types of features may have a strong impact depending on their sophistication.

lack of baseline system There is no evaluated baseline system (Sec. 1.2) and the current baseline system may also be lacking some core features such as rapid audio-playback and nice displays of features of a whole dialogue. Automatic speech recognition transcripts of the type of data we are interested in are still fairly hard to read which makes any comparison to text and keyword based methods difficult.

user time The study would need users to perform real life tasks, which would take hours if not days. Since we would like to test the effect of individual features we would need enough subjects for each group to make reliable judgments which would lead to a tremendous amount of work. This effort may be invested more wisely otherwise.

privacy concerns Privacy concerns often exist for recordings of meetings. Parts of the meeting database were therefore released only for our internal use and that material is being used in this study since only recently the recordings of publicly available material started. The study that was presented here was only conducted using subjects from our labs. This restricts the available subjects and their willingness to participate in long experiments. Besides the within-group constraint we found it to be very important that no embarrassing or highly personal information would be revealed. It would be difficult, however, to exclude such information from the meetings in our user study and users want to listen to the whole audio record randomly.

tainted subjects Once we asked one subject to perform a task on one rejoinder the subject has gathered additional knowledge of that rejoinder which can be used to answer questions about that rejoinder. If a subject has therefore used one conversation this conversation can't be reused for the same subject. A paired test design is therefore not possible.

Instead a study was designed which simulates a real retrieval situation: A user may have some vague recollection of a meeting and wants to find it in a database. We make the following approximations of that situation:

vague recollection Of course we can not have 20 subjects with the same vague recollection of a meeting. Instead we play 4 short randomly selected snippets of a meeting (6 for Santa Barbara) of 15sec length each. The samples are separated by some distinctive sounds (beeping). The samples create an overall impression of the meeting which is typically 15-70 min long.

database The database is short, 4 rejoinders from Santa Barbara and 4 from the meeting database are selected. The users know which database a rejoinder belongs to and has to select among those rejoinders. The databases are short for three reasons:

- Only 4 rejoinders of the meeting database could be used for privacy considerations (see footnote 4).
- Searching over large databases would raise user interface issues.
- Our estimate for the maximal search space is reduction was a factor of 2 such that a larger database would result in random picking of similar rejoinders which complicates a proper evaluation.

retrieval The user has to select one meeting in the database as the one which belongs to the audio excerpt. The meetings in the database are represented using graphical representations of non-keyword based indices such as dialogue acts, activities, speakers and formality.

Using this design we get a somewhat indirect assessment of retrieval performance by randomly assigning different feature sets to the users and measuring their detection accuracy. This design minimizes the effect of the user interface especially since there are no dynamic components. The design is also hopefully fairly easy and fast to learn such that subject would perform similarly. The study was designed to take about half an hour to ensure that many subjects would take part in it which was important since the pool of subjects was restricted to our lab. The details of the design are presented in Sec. 6.3.2 and the study was conducted using 20 subjects (Sec. 6.3.3).

6.3.2 Experimental Setup

The user study was implemented using WWW technology such that users were able to conduct the study at their convenience. Two designated Windows PCs have been set up, one in our lab in Karlsruhe and one in Pittsburgh. Users were encourage to use those PCs since they had been tested and the (audio) data was available locally. Access for users desktop machines was provided via local NFS filesystems for UNIX machines and via our standard WWW servers for Windows machines. The user study was restricted to active members of our lab since the data contained recordings of our own meetings which we did not want to give to outsiders [2]. The setup required the users to have audio-playback capabilities as well as email [3] to submit the results. Each user was given a URL which redirected the user to a randomly selected experiment. All experiments were identical besides the feature set that was being presented. The user study is therefore able to create results about information access properties of those features. The users were presented a page which contained the following information:

General explanation How and why the study is conducted is introduced. The users were asked to agree/acknowledge those conditions.

Audio setup Instructions were provided in order to make sure UNIX users are accommodated.

Demographic data Users were asked to add information about their name and email (to be separated during the experiments), their age, English proficiency, etc.

Main experiment Users were asked to identify which one of four rejoinders represented as a graphical representation has been presented an audio excerpt.

Comments / Rationale Users were asked how they made their judgments (rationale) and if they had other comments about the study.

The main experiment consists of two parts, one consisting of 4 audio excerpts for the Santa Barbara corpus and 8 for the meeting corpus. The users were (correctly) informed that each one of the 4 rejoinders of the Santa Barbara and meeting corpus

[2] Technical protection of the files was provided by making them either accessible over PCs in our lab, filesystems that are only accessible in our group or password and IP restricted access to our WWW servers. The files were erased immediately after the last user completed the study and the users signed that they would not be allowed to use or save the data for any other purpose than this study.

[3] As trivial as this may sound the email setup on many of our Windows PCs was not very consistent and prone to failures. This problem resulted in frustration by participants and in a small number of uncollected results and was the major technical problem.

were represented by at least one audio excerpt but that they may occur any number of times among the audio excerpts. Each audio excerpt contains randomly picked samples [4] of 15 sec lengths, 4 (resp. 6) samples were selected for each meeting (resp. Santa Barbara) excerpt and the samples were concatenated with an audio beep between them. The original rejoinders were about an hour long and were represented graphically on a time scale using the following features (Fig. 6.1) which are also available in the meeting browser interface (Sec. 1.2):

Formality Formality as described by the formula in Sec. 2.4.2.3 has been used on a topical segment bases. Users were instructed that in formal speech the participants "express" themselves in a formal manner in order to encourage a more narrow interpretation of the term.

Channel For each speaker a channel is presented which shows when the speaker is producing words. For the meeting database the speaker identity is revealed since many participants would be able to guess which channel represents which speaker. Additionally audio samples of the speakers are provided for the meeting database.

Dialogue acts A selection of dialogue acts was provided which is basically a classification of the segments of the channel feature. The classification was done using a detector trained on Switchboard. The resulting representation is nothing more than a colored version of the channel feature.

Activities Manually annotated activities have been provided.

5 different experiments (Tab.6.8) were constructed that contained different featuresets. The basic information was created from transcripts which contain utterances, manually transcribed word, time-stamps and channel information. Activity has been annotated manually as well. Dialogue act information was added automatically and formality was calculated. The data was stored in the meeting browser format and a script was used to create the WWW based interface including the graphical representation directly from the meeting browser format.

[4] The samples for the meeting corpus have been evaluated for suitability in the study. Both the author and two members of the data collection team that were also heard on the records listened to the recordings. The main criteria for removal were the presence of private and embarrassing audio as well as identifiable references to outside people or institutions. The meeting database is already filtered but this additional step was necessary; although no record was kept only about a quarter of the segments were retained which may also illustrate how much data may be considered sensitive in other meetings. The data was fairly neutral after the selection process but the author still decided to error on the safe side and restrict the study to active members of our lab. Originally another set of meetings from our own collection was intended to be part of the study however there were too many references to outside people or institutions such that the filtering process was time intensive and resulted in unrepresentative excerpts.

Exp.	Formality	Channel	Activity
1	●		
2	●	●	
3	●	Dialogue Acts	
4	●	Dialogue Acts	●
5	●		●

Table 6.8: **Experiments in User Study:** All 5 experiments made use of the formality feature. In experiment 2 the channel activity is added which is extended by the dialogue act coloring in experiment 3 and 4. Experiment 4 adds activities. For comparison experiment 5 removes the dialogue act annotation from the feature set.

Dialogue Acts	Activities	Formality
statement	storytelling&informing	highly informal
backchannel (yeah right!,OK)	discussion	fairly informal
question	interrogating	somewhat informal
disbelieve	giving advise	somewhat formal
undetermined	planning	fairly formal
	other	

Figure 6.1: **Color Encoding:** Formality, dialogue acts and activity labels have been color coded for the graphical representation and this table is identical to the table that was available to users. Users were only shown the features relevant for their experiment.

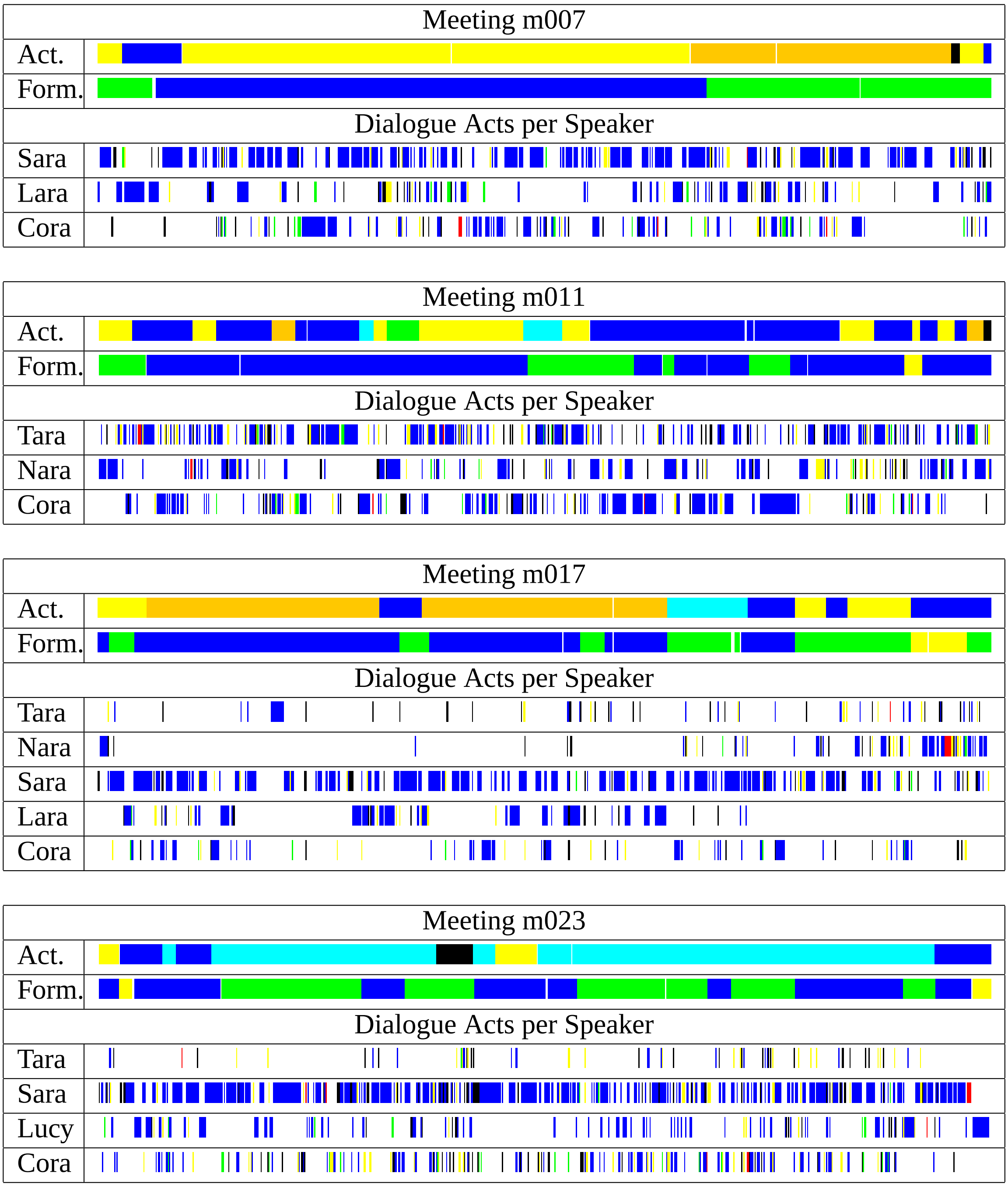


Figure 6.2: **Graphical Meeting Representation:** The representation of 4 meetings for experiment 4 (Tab. 6.8) is shown and the colors can be interpreted from the codes in Fig. 6.1. This figure contains exactly what the users had available, namely activities, dialogue acts and formality. The original meetings are 15-70min long and the users are asked which rejoinder they heard short excerpts of. The representation is generated from a file format that is used by the meeting browser interface for the integration of dialogue information. The original recognizable names of the participants have been replaced with neutral names.

	Available			# meetings identified correctly												
	Features			Santa Barbara				Meeting								
#	Form.	Ch.	Act.	Avg.	0	1	2	Avg	0	1	2	3	4	5	6	7
1	•			1.7		1	2	2.3		1		2				
2	•	•		0.3	2	1		2.0		1	1	1				
3	•	DA		0.7	1	2		3.0			1	1	1			
4	•	DA	•	0.8	3	1	2	3.7			2	1	2			1
5	•		•	0.6	3	1	1	3.8			1	2		1	1	

Table 6.9: **Basic Results User Study:** The experiments on the Santa Barbara Corpus don't show many interesting results. However the results on the meeting database suggest that there is an effect from activities and may be from dialogue acts. The results are obtained from 4 Santa Barbara and 8 Meeting Corpus excerpts, the users were able to choose among 4 rejoinders from each corpus respectively. The available features are formality (Form.), activity (Act.), channel information and dialogue acts (• resp. DA in column Ch).

6.3.3 Results

The user study was conducted over a period of 1 month. Users were invited by email initially and asked again for participation. 20 individuals participated in the study and the experience was that they took the study seriously. However the demographic data may not be considered too accurate and is therefore largely ignored. Most participant were PhD candidates in our lab and most were non-native speakers and the average time to complete the study was 37min. A lot of users reported in the comments that they felt insecure about their judgments, even if they had fairly good results in their experiment.

The basic result is the number of times a user has identified a meeting in a database correctly (Tab. 6.9). Picking one out of 4 meetings from the respective database at random results in an expected baseline performance of 1 on the Santa Barbara Corpus (4 excerpts) and 2 on the meeting database (8 excerpts). While it is immediately clear that effects can be observed for the meeting database it is not quite that clear for the Santa Barbara Corpus. The author found this surprising since he was able to use the information himself in pre-studies. The reasons for that can only be guessed but there are some difference between the two databases which are worth noting

one-to-one assignment The Santa Barbara Corpus design required to use only one audio excerpt per rejoinder instead of many as in the meeting database. The reason is that Santa Barbara audios excerpts can easily be identified as belonging to the same recording by speaker identity (no speaker overlap),

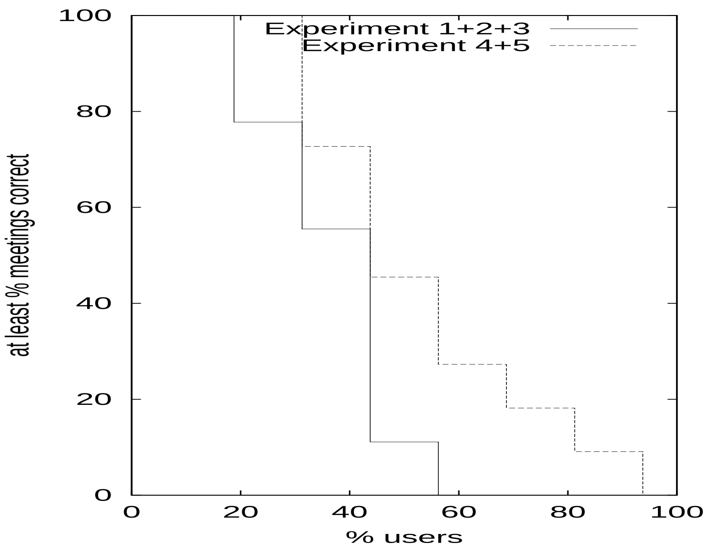


Figure 6.3: **Cummulative Results User Study:** The plot shows how many users got at least y% of meetings correct combining the experiments 1, 2 and 3 and the experiment 4 and 5 (on the meeting corpus, not excluding unusual users). The figure illustrates that users from experiments 4 and 5 recognized significantly more meetings correct.

background noise and other acoustic room conditions. If there were multiple excerpts users would likely be able to use that information which is unintended. However this also means that users try to make use of the constraint that each rejoinder is represented by exactly one excerpt which may or may not be a beneficial strategy.

small sample size Since the number of excerpts is so small in the Santa Barbara Corpus it is hard to draw conclusions from it.

familiarity The meeting corpus contains speakers that most of the users know while there were unfamiliar situations in the Santa Barbara Corpus. The excerpts might also have been to short for the users to identify them. The Santa Barbara Corpus situations are vastly different from each other but it may be too short to figure out the activity. This may change if users have more experience with the feature representations and the excerpts are longer.

practice effect The Santa Barbara Corpus experiment was earlier on the WWW page layout than the meeting experiment. Users would therefore likely first solve the Santa Barbara Corpus problems and gain experience with the general setup. It could therefore also be seen as a warm-up session before the main experiment.

large variation Another explanation of the results could be that due to the large variation in dialogue style the users may have only been able to use the simplest feature, formality.

In the sequel we will therefore only consider the meeting corpus since there might have been some fundamental problems with the Santa Barbara setup which cannot be resolved after the study had been started. Given the limited pool of participants, their time constraints and the "tainted subject" problem only very limited prestudies were possible and we will therefore ignore the results on the Santa Barbara Corpus. A close inspection of the individual user performance on the meeting database revealed some important insights:

familiarity helps Two members of our group were also on the recordings and they appeared there a lot. Indeed they achieved the highest number of identified excerpts: 7 resp. 6 excerpts in experiment 4 resp. 5. The meetings where conducted more than a year ago such that an immediate recall was impossible.

failed strategies One user described a search strategy in the comments that was elaborate but made invalid assumptions [5]. The user scored the lowest (2 excerpts correct) in experiment 4.

feature usage Users referred heavily to activity when available according to their comments. However there were also users which used formality or dialogue acts even when other features were available. Feature usage may therefore be a matter of personal preference.

Given the observation of failed strategies and familiarity the results will always be reported both without any correction and with a correction eliminating the user with the "failed strategy" and the highly familiar users. The basic results as well as results of pooled experiments are shown in Tab. 6.10. The table also features a Student t-test to show whether the mean is significantly larger than 2 which is the performance of a system picking randomly. Indeed it can be observed that this is the case for many experiments, especially when they are pooled. This is especially true if dialogue acts or activities are in the feature set. However it is hard

[5] The assumption was that the graphical representation corresponds *exactly* to the excerpt. Using that assumption the user searched for the most differentiating position in the representation and listened to the audio at that position to make the judgment.

to determine whether one feature set performed better than another in a reliable way: Tab. 6.11 shows the results of various (pooled) feature-set comparisons and assigns significance values with three different statistical tests. Among statistical practitioners the Student t-test is often applied even if the application conditions are violated or untested (especially the normality of the variable). Making a rank-transformation on the input data is enhancing the robustness of the test such that the results are often considered more trustworthy (rank-transformed t-test). The non-parametric Whitney-Mann U-test is less sensitive, especially on our data set where we have many ties (identical values, see Tab. 6.9). Judging that there is significance even if that test fails is therefore not unreasonable, especially if the significance value of the t-test and the ranked transformed t-test are good. One may therefore draw the following conclusions:

formality Formality couldn't be confirmed to be usable by the subjects on the meeting database and it didn't perform better than a random baseline. However there may be an effect on the Santa Barbara database.

channel Channel information couldn't be confirmed to be an effective feature, even in combination with formality. It seems unlikely to prove effective even in larger studies of similar data.

dialogue acts Dialogue acts have a higher average value over the baseline but the result could not be proven to be significant since the dataset is too small.

activity There is a strong and significant effect for activity over the baseline. Both the mean is significantly higher than random guessing and the feature is significantly better than other features.

6.4 Interactive Access

6.4.1 Introduction

The insights gained from dialogue detection, dialogue segmentation, user study and information access assessment work can also be applied to gauge the theoretic value of features for interactive information access. To bring these features to bear however on a concrete search task a user interface has to be designed. The user interface design to search in very large document sets would probably be similar to traditional search engines, possibly with some additional selectors for database and activity types, participants, timing information and so forth. When it comes to browsing and skimming smaller document sets however the properties of the different modalities come into play such that the user interface design is critical.

	Available			# meetings identified correctly			
	Features			All users		No unusual users	
#	Form.	Ch.	Act.	Avg.	S	Avg	S
1	●			2.3		2.3	
2	●	●		2.0		2.0	
3	●	DA		3.0		3.0	
4	●	DA	●	3.7	0.05	3.3	0.05
5	●		●	3.8	0.05	3.3	
6	1+3			2.7		2.7	
7	1+2+3			2.4		2.4	
8	4+5			3.7	0.01	3.3	0.01
9	3+4+5			3.6	0.01	3.2	0.01
10	2+3+4+5			3.3	0.01	2.9	0.01
11	1+2+3+4+5			3.2	0.01	2.8	0.01
12	1+2			2.2		2.2	
13	3+4			3.4	0.025	3.1	0.01

Table 6.10: **Meeting Database Average Performance:** The results for all users and without unusual users (people that occur on the record and a user with a clear misunderstanding of the study) on the meeting database are shown. The all-user part corresponds to Tab. 6.9 and the averages (Avg.) are shown, unusual users resulted in 2 and 7 correct excerpts in experiment 4 and 6 correct excerpts in experiment 5. To overcome the small number of experiments we may pool experiments which are similar. For example experiment 9 (containing all the subjects from experiments 3,4 and 5) clearly results in an above random #meetings correctly identified. Only for some of the original experiments the result passes a t-test with the Nullhypothesis that the mean is 2, however pooled experiments show enough significance (S) is the achieved significance level).

Visualization and playback capabilities have been addressed by the construction of the meeting browser such that the work in this thesis abstained from building specific interfaces. The user study uses graphical representations of non-keyword based features and evaluates them for use in the meeting browser (Fig. 2.1). Indeed the graphical representations are created based on an intermediate format that is compatible with the meeting browser dialogue file-format. Three general options seem to be available for generating descriptions of a dialogue:

excerpts Extracting characteristic element of the discourse that can be played, seen or read easily since they are short may be a powerful technique. Textual summarizations and thumbnails of keyshots are examples of excerpts.

Pool		All users			No unusual		
Better pool	Worse pool	t	r-t	U	t	r-t	U
5	1	0.05	0.05				0.314
4	2	0.025	0.025	0.131		0.05	0.196
5	2	0.025	0.025	0.125		0.05	0.2
3+4	1+2	0.001	0.001		0.025	0.001	
4+5	1+3	0.01	0.001			0.05	
3+4+5	1+2	0.001	0.001	0.005	0.01	0.001	

Table 6.11: **Interfeature Comparisons:** (Pooled) experiments are compared against each other. While there are a good number of comparisons which result in successful one-sided t-tests (t) or rank-transformed t-tests (r-t) the non-parametric Whitney-Mann U-test (U) only rarely yields significance.

Rejoinder Name	Correctly identified	Description of meeting
007	5 6 6	New group member, introduction, a lot of advising
011	3 6 9	Usual group leader missing, a lot of idle talk and storytelling
017	2	Advising and organization of accommodation
023	7	Planning of an event

Table 6.12: **Recognition by Excerpt:** Given the rejoinder identity the table states how many times an excerpt of a meeting was identified correctly. A short description of the meetings is added.

rapid playback Many features may be preserved even if the playback is fairly rapid. Rapid video playback would still allow to see the scene and probably also who is attending the rejoinder. Rapid audio playback might show who is talking and what kind of interaction is taking place, even if the word level information is not available anymore. Arons (1994, 1997); Roy (1995) present work on rapid playback and structured access in "audio only" interfaces but restrict their attention to intelligible speech and are not explicitly concerned with dialogues. It is unclear at this point whether activity can be maintained in rapid playback but Sec. 6.4.3 will make suggestion how this may be done for rapid audio playback.

graphical / textual representation Dialogue acts for example can be visualized by color bars that stretch a track for an individual speaker for the extent of the act. The dominance distribution of speakers may be shown, the speakers may be color-coded, keywords could be displayed et cetera. Graphical and textual representations are explored in the meeting browser interface. The non-keyword based features are presented as color coded time bars as in the user study (Fig. 6.2) whereas textual information is presented in the form of plain text and summaries,

Since excerpts (Sec. 6.4.2) and rapid playback (Sec. 6.4.3) are not explored in this work this section provides ideas how to use them, especially how to exploit non-keyword based features in excerpts and rapid playback.

6.4.2 Excerpts

Excerpts are a simple yet effective method that has been widely used to summarize topic. In that application parts of the rejoinder are weighted heuristically to represent most topical information and the result is usually displayed as a text. However excerpts could also be chosen to represent the stylistic and situational aspect of the rejoinder or segment in a rejoinder: To represent the style a longer continuous segment in the middle of the rejoinder should be chosen. A continuous segment needs to be chosen to represent the dialogue act sequences properly and the middle is preferred since it is not tainted by specific effects at the beginning or end of the rejoinder. The general genre of the conversation should also be easy to understand from a short excerpt since it was shown that the genre differ along many elementary feature dimensions and the user study showed that human can differentiate between activities from short excerpts.. This kind of excerpt may also be played back rapidly since the dialogue act information is probably retained even at fast playback speeds. An audio excerpt may additionally provide situation information about the participants, the location (background noise) and other aspects of the situation (is pizza being eaten, is the rejoinder in a large or small room or outside). Video information however would add more information about the situation and may carry body language which relates to the style of the conversation (see also Sec. 2.4.7 and Eldridge et al. (1991)).

A variation on excerpts are icons, for example one could select pictures from the rejoinder, pictures of the speakers, the room, or activities. These icons may be placed in a time-continuous display of the rejoinder to represent certain dialogue information.

6.4.3 Rapid Playback

Rapid playback for audio documents was a technique that was first studied extensively by Arons (1994, 1997) (Sec. 2.2). Rapid playback may be supported by the results of this thesis by adding information where to find segment boundaries (Sec. 5) as well as by the description of features that are indicative of the content of the rejoinder and that might be remembered. But there is more information available now what is important in order to find information: First of all there is no need to use just audio – a textual or graphical display could be used at the same time. If for example the playback is too fast for individual keywords to be intelligible the playback mechanism may display the most important keywords. Additionally an interesting research problem is how the intelligibility of selected keywords could be maintained although the playback speed is very high. The speed of the playback could either be reduced entirely while producing the keyword or – on a second track – the keywords could be produced.

If the playback would be very fast it might even be difficult to process who was speaking at a given point in time and what dialogue act was produced. This problem could be addressed by using surround audio to place the different speakers as detected by a speaker detection system at different locations in audio space and encode the dialogue acts using specific sounds or sound manipulations of the original. Since dialogue acts are indicators of activity and speaker identity is useful for finding topical changes the rapid playback may maintain crucial dialogue information. This playback mode could be supported by a display that correlates to this information in the form of a map of the audio space which is showing the dialogue acts at the same time they are being produced on the audio system. Keywords may be overlayed similarly and could be originating from a certain point in audio space, possibly in the voice of the original speaker, and be displayed in textual form as well.

Video lends itself much more naturally to rapid playback as the common experience of the fast forward function of video recorders illustrates. But there might also be a limit where the movements are becoming too fast and the body language becomes unintelligible. At that point one may consider to fall back to a mode where keyframes for individual segments are selected which are then displayed instead of the actual movements. Overall the potential for rapid playback is still largely untapped although it may be a crucial tool and it certainly should be part of any comprehensive system that features information access to spoken communication. This thesis has provided insight in the nature of the features that need to be represented in rapid playback which should enable the targeted development of such systems which are easy and natural to understand for the general public.

6.5 Conclusion

This section measures the success of features in information access. The first set of results is obtained using an information theoretic approach which allows to produce many results but carries a significant number of implicit assumptions. Additionally the technique is restricted since unwanted correlations between speaker identity and other features need to be estimated and eliminated. The second approach is a user study which measures the rejoinder retrieval performance and also evaluates a certain feature representation. The advantage of the user study is that it may simulate a real life experience closely but the disadvantage is that only a small amount of reliable results can be obtained.

The information theoretic approach (Sec. 6.2) reveals that topical information as encoded by keywords is not necessarily a very good retrieval method. However it can be shown that activity may be important to find a rejoinder. Word level information was hard to judge since it may and probably does encode speaker identity as well. Indeed most of the discrimination capability is achieved using a few highly frequent words as well as the parts-of-speech. On CallHome Spanish some search space reduction was possible using keywords.

The user study (Sec. 6.3) was also able to confirm that activity annotation is useful when finding rejoinders using a graphical representation. However expressed formality and channel information are very unlikely to be good features according to this study. Dialogue acts could not be proven to be good features due to the size of the database but the results suggest that they may prove good features if more data could be gathered to confirm the results.

Finally applications of this work to user interfaces for audio records are suggested (Sec. 6.4). Specifically the creation of excerpts and fast audio playback to maintain non-keyword based features is described.

Chapter 7

Conclusion

7.1 Introduction

Our daily environment is rich in audio and video that could be recorded for documentation purposes. This data could be highly valuable but it may be expensive to document and therefore important pieces of information may be lost. In order to gain access to these potentially large and unstructured ressources we need indexing techniques. Keyword based indices are likely of limited use since the performance of speech recognizers is low on rejoinders such as meetings, the number of (unique) keywords is lower than in written language and the keywords being used may be idiosyncratic (Sec. 1.2). Research into human memory has proven that the speaking situation is often recollected such that it may be a used as an index by an information seeker which took part in the conversation (Sec. 2.4.4). The speaking situation may also be used by an information seeker who did not take part in the conversation if it is known a-priori which kinds of conversations are good candidates for the desired information.

This thesis shows that a number of non-keyword based indices reflecting the speaking situation can be detected automatically (Sec. 7.2) and that conversations can be segmented into topical units using non-keyword based methods (Sec. 7.3). Some non-keyword based methods don't require (full-scale) speech recognition and may be implemented even on small mobile devices. The results of information theoretic methods and a user study show that non-keyword based methods are effective (Sec. 7.4). These results have an impact on practical systems (Sec. 7.5) and can be used as a basis of future work (Sec. 7.6).

7.2 Automatic Style Detection

7.2.1 Textstyle Classification

This thesis shows that “databases” and “sub-databases” of rejoinders can be detected automatically with high accuracy which allows to search for a specific type of rejoinder (personal and non-personal phone calls, broadcasts, etc.). The detection is done using neural networks and relatively simple features (Tab. 7.1) and the resulting reduction in search space is approximately a factor of 4-10.

At the topical segment level “activities” can be distinguished (storytelling, discussion, . . .). Those activities are sometimes even hard to distinguish for humans and are hard to detect using machine learning techniques (Fig. 4.1, Tab.7.1, Sec. 4). However, the detection results presented here are still fairly good, given the difficulty of the task. If the activities are reasonably distinct and the training database is reasonably sized one can achieve better results (CallHome Spanish without storytelling). Additionally a user study has shown that activities have a high potential for information access. A maximum search space reduction factor of 6 is theoretically possible on the databases used.

7.2.2 Dialog Act and Game Detection

At the utterance level detectors for dialogue acts (questions, statements, backchannels, . . .) and dialogue games have been built (Sec. 3). Through an embedding of traditional language model classifiers into exponential models an efficient discriminative training technique was developed. The algorithm results in improved detection accuracy for dialogue acts using simpler features and less context dependency. The search procedure is able to segment dialogue acts on multiple channels. This work also presented the first computational dialogue game detector (e.g., question/answer pairs) using a multi-level HMM / NN framework. Dialog acts and games are useful for building classifiers for speaking style. Similarily a user may remember a certain dialog act sequence and may search it in a graphical representation – the results of the user study seem to indicate that but statistical significance could not be established due to the small data set.

7.3 Topical Segmentation

A new topical segmentation framework has been proposed to break up rejoinders into more manageable chunks. Topic segmentation is an important task and may be used in audio skimming and summarization, among others. Efficiency may be important since topic segmentation might be one of the first algorithms

Example Class Labels	Search Space Reduction	Accuracy in % (a)/ (b) /(c)	Comments	Section
Document Level Classification				
CallHome English, Spanish, Broadcast News, Switchboard	4.0	25/100/	simple features like word,segment and speaker overlap length sufficient	4.3
Talk-shows, Movies, Newscasts, . . .	9.9	32/ 67 /	large TV show corpus, good results (>60%) using only stylistic features	4.4
cross-product earn, acq, money-fx, grain, crude, trade, interest, wheat, ship, . . .	18.2	/ 87 /	Reuters database, microavg. F-score of SVM classifier (Yang and Liu, 1999) times 100 for number in classifier accuracy, can be compared to accuracy (Sec. 2.4.6.2)	2.4.6.2
Sub-Document Level Classification				
Storytelling, discussion, planning, . . .	6.4	55/ 65 /73	Meeting corpus, see also Fig. 4.1	4.5.3
Well Being, Travel, Job, Money, Health, . . .	4.9	55/ 57 /	Handannotated topic classes, CallHome Spanish (Tab. 6.5)	6.2.5.2

Table 7.1: **Document and Sub-Document Classification Results:** The *Search Space Reduction* is a factor measured using perplexity. The perplexity can be calculated from the entropy x in bits as 2^{-x} (Sec. 6.2.2). In the first document classification example 4 classes are distributed evenly and the resulting perplexity of 4 corresponds intuitively to the search space reduction. The *Accuracy in %* show the the chance of picking the right category by (a) picking the most frequent category, (b) using a machine learning classifier and (c) by using a second human annotator. It is easy to see that the document level style detection results in significant search space reductions and that automatic classification is easy to obtain. Topical classification on the Reuters database – which lends itself much better to a topical classificaton scheme than speech data – results in search space reduction which are larger but are in the same order of magnitude. On the document level activities seem to offer a reasonable search space reduction, however detection is difficult. Topic annotation appears to offer some search space reduction however it is extremely hard to detect. See also Fig. 4.1.

which would be deployed, e.g. for audio skimming on mobile devices (Sec. 7.5.3, Sec. 7.5.4). The new algorithm is very interesting since it performs very well using traditional keyword repetition features on a number of corpora. Rejoinders can be fairly long such that the linear runtime of the procedure (if a maximum topic-length constraint is assumed) can be important. Most importantly however new features have been investigated which are used successfully for segmentation:

- Speaker initiative requires only speaker identity and turn length information.

- Stopwords are effective and might be easier to detect reliably with a (small) speech recognizer.

- Speaking style differences in the beginning, middle and end of the rejoinder described by the most frequent words and parts of speech.

7.4 Information Access Assessment

Indices for spoken documents are features that can be used for information access. Features make indices for spoken documents if they are typically known to an information seeker, they can be effectively extracted from a collection of spoken documents and they can effectively used and processed by the information access system) (see also Tab. 1.1).

Keywords may be, depending on the application, known to the information seeker. However it should be recalled from Sec. 1.2 and Tab. 1.2 that keywords are unlikely to perform as well in spoken documents from environments such as meetings since keywords are repeated (instead of synonyms used) and rare which makes them bad indices. Furthermore the precise keyword is rarely remembered of a conversation but rather its meaning (Sec. 2.4.4). Keywords are difficult to use for an audio-only information access device since they need to be played back accurately while they can be used effectively in conjunction with summarization techniques. They may also be assumed to be largely independent of dialog or stylistic features as observed in Sec. 6.2.5. Keywords may therefore be useful however they are likely much less .powerful than in other scenarios (Sec. 2.4.5.2).

The assessment of information access is trivial for the *database* features: Since the database can be detected with very high accuracy from a rejoinder and we may assume that a user remembers them correctly the reduction can be maximal, e.g., if there are 5 databases which are equally likely the search space is reduced by a factor of 5. The author assumes that this information is highly correlated with topical information which may be represented by keywords. However the content of a database may change more rapidly than its general style (e.g. an investigative newspaper will change topics, not style).

Information theoretic measures allow to refine this kind of argument (Sec. 6.1). For *sub-databases*, for example, the following models may be assumed: If a human knows the exact sub-database and it has been annotated correctly in the database, a reduction by more than 3 bit is possible on our TV-show task (Sec. 4.4). If a machine however annotates the database and the human knows the exact sub-database a reduction of approximately 1bit is possible, still an average search space reduction by a factor of two (Tab. 4.7).

While this argument is still straightforward it is much more difficult to assess the search in a database or within a rejoinder. Some microlevel features, especially very frequent words and parts of speech, index speaker identity and speaker identity indexes the rejoinder, such those features can't be evaluated independently. It can be shown that microlevel features such *semantic word categories* and *dialogue acts* hold some potential for finding rejoinders in a database. *Keywords* have only be shown to be effective for the selection of a rejoinder in one of our databases.

Activities have a potential for search space reductions for picking a rejoinder and a segment in a rejoinder of about 2.4 bit, however under quite optimistic assumptions (Sec. 6.2.5.2). About 1 bit may be attributed to the search space reduction within a rejoinder. Activities are mostly independent of topical features in our database as shown by manual topic annotation (Tab. 6.4) and the fact that adding more than the most frequent 50 words did not impact activity detection.

A study was conducted to investigate whether users can search effectively using *expressed formality* [1], speaker identity, dialogue acts, and activity (Sec. 6.3). The results were obtained on a small database of meetings from the same group. Formality has no effect and neither does speaker identity. Dialogue acts may perform well but the data is too sparse to make judgments. A positive retrieval effect however can be measured for activities. The results therefore suggest that high level features are important since they can be associated more easily than lower level features. Microlevel features are also much less likely known to an information seeker and if they are available the information seeker may also know good high level features and precise keywords [2].

[1] Expressed formality is calculated from the parts of speech distributions and is empirically established by Heylighen and Dewaele (1999)

[2] It should be noted that this thesis only evaluated features which are available on the audio channel. Microlevel features on the video channel may be easier to remember since they may transport higher meanings and may be more emoting (e.g. "smiling at someone", "hitting someone", ...). Video recordings that accurately represent such features are more difficult to fabricate and video recordings are unavailable in audio-only media such as telephone conversations or radio broadcasts.

7.5 Impact

7.5.1 Introduction

Non-keyword based indices and their investigation affect the design of audio documentation systems in three major ways which will be illustrated by applications: Improved indices are obtained (intelligent meeting room, Sec. 7.5.2), the computational requirements are reduced (mobile platforms, Sec. 7.5.3) and new applications can be designed given the analysis (audio skimming, Sec. 7.5.4). These examples present excellent opportunities for future work (Sec. 7.6), both in a scientific as well as in a commercial environment.

7.5.2 Meeting Rooms

The intelligent meeting room scenario is extremely challenging since the recording conditions are hard to control and the speech recognition accuracy tends to be low. However, the language being used is also likely to be harder to index by standard keyword based methods than written language (Sec. 1.2). The major goal is therefore to arrive at better indices for the audio records. Assuming that we have a database of similar meetings there might be two important tasks: Finding a meeting and finding a piece of information in a meeting.

Using graphical representations of activities a user can pick the correct meeting out of four meetings from the same group in 43% of the time. The users were able to listen only to four randomly excerpts of 15 sec length such that this result is extremely pessimistic. Obviously this statement is fairly limited as it stands but there are a number of important remarks: The meetings that have been tested were fairly similar with almost the same set of speakers such that there might be stronger constraints in larger or more varied databases. The selection between fairly similar rejoinders is likely similar to the navigation within a rejoinder such that users would be able to navigate in a rejoinder using dialogue act distributions and activities. Information theoretic measure estimate that activity has a much larger potential (the entropy of entropy is 2.7 bit) and that topical information is likely orthogonal to it.

The meeting browser is currently best suited for playback and skimming of relatively short rejoinders. Non-keyword based features are likely most effective to navigate across larger regions of a rejoinder. Local information such as the dialogue act of an utterance may as well be inferred by the user from a textual display: Showing the “statement”-class for “The grass is always greener.” seems to be rather pointless if the user has the sentence printed on the screen or it is played back.

A visually attractive and effective combination of text, (noun-phrase) summaries, acoustic scene analysis, activities, dialogue acts, formality, etc., for accessing a small number of (long) meetings is certainly a research project of its own and would allow to judge the individual merits of the various feature classes. Similarly a much large database of meetings would allow to judge the effectiveness and relative merit of features for query based document access in larger collections. Both tasks require a significant amount of infrastructure and funding in order to obtain any system-oriented results (Sec. 7.6).

7.5.3 Mobile Platforms

Some non-keyword based methods are computationally a lot more effective such that they could be implemented on mobile devices with small footprint. Specifically the detection of the database and possibly even the sub-database can be done with very simple features which require little processing (Sec. 4.3) and the detection of topic change is greatly simplified with speaker initiative features (Sec. 5.5.2).

In conjunction with the work in our group on auditory scene analysis detecting very drastic changes in the audio environment (Sec. 1.2) and the work on speaker detection (Sec. 2.4.5.4) – both techniques low in footprint – the audio channel could provide a lot of indexing information. The following information could be available on a mobile device such as a personal digital assistant (PDA):

- the general situation: room, outdoors, discussion
  auditory scene analysis and database detection
- who is there
  speaker detection
- when did the topic likely change
  topical segmentation
- when did the events happen
- what was scheduled in the calendar
- which notes, appointments or action items did the user make
- which material / book / slides did the user view on the PDA

The example shows that mobile platforms such as PDAs could benefit tremendously from the availability of additional indices. Full speech recognition is likely too demanding for this platform and the robustness necessary may also be too demanding.

7.5.4 Audio Skimming

Audio is a linear medium and it is hard to replicate the visual skimming that can be done in texts. Additionally rejoinders can be very long, complicating the situation. Arons (1994) presents an audio only interface to speech data. This thesis supports audio skimming since one of the important operations — topic segmentation — has been successfully addressed. However Arons (1994) doesn't address the problem of rapid playback beyond the point where the audio is intelligible (2-3 times faster than realtime). Given the analysis presented in this work speaker identity (e.g. for topical segmentation), activities and dialogue acts seem to be important to maintain. Sec. 6.4.3 presents a number of suggestions how to maintain those features to break this speed-barrier.

7.6 Future Work

This thesis is still just at the beginning of a complete audio documentation system. The most pressing issue is the construction of a complete system that can be fielded and exposed to large user populations. This step would — while extremely labor intensive — yield a lot of user feedback and would also show different types of features in an integrated setup. One long-term strategy to simplify the construction of robust and complete systems would be the participation in industry standardization committees such as MPEG-7 since many features investigated in this work are not yet reflected in evolving industry work. The meeting scenario (Sec. 7.5.2) may currently offer only a limited amount of users since the technology is still very complex and brittle and it might be a little too challenging for a product ready for customers, however the PDA scenario (Sec. 7.5.3) or audio skimming (Sec. 7.5.4) might be interesting applications. Another scenario that would be interesting to investigate is "audio in the office" for the documentation of impromptu meetings and tutorials. The meeting scenario however is still interesting to pursue since it potentially has a high impact – a lot more meetings would need to be available and infrastructure to manually code significant amounts of data in order to make significant progress.

While this thesis shows results on dialogue act and game detection there is a lot of work to do to make this more practical: Testing on unrestricted databases is an interesting issue, lowering the footprint another one, the discriminative training not only of the emission probabilities but also the transition probabilities yet another one. Better detection algorithms with lower footprint which are more robust could bring dialogue act detection in an interesting space for the application on mobile devices.

A crucial question for future work is the design of user interfaces to small collections of (long) audio documents. Audio is a linear medium and we can't let our ears wander as we can our eyes. This issue is important since small collections of written documentation can be skimmed visually and this access step is usually part of any information access system: After the query based restrictions have been applied users need to select which one of the retrieved documents would fulfill their information need. The problem can either be addressed in the context of the multimodal meeting browser interface (Sec. 1.2) or in an audio-only user interface (Sec. 7.5.4). Another scenario could be the construction of high-quality minutes by a designated minute taker (similar to Moran et al. (1997)).

Related to the problem of user interfaces is the question of support for reinterpretation (Sec. 2.3.1.3). Written documentation is constantly reinterpreted by citation, annotation on the document or the place a piece of paper is located at. The research question is how to achieve the same or better if we want to make notes to audio documents, send them to other people or want to publish them. One may rephrase this question as the problem of what a document is and how a multimedia documentation should look like, especially as the material is being reused, commented and agreed upon and altered over time.

If the author had unlimited funding the most interesting research project would be the analysis of social and thematic relationships and their impact on meetings. This research would be very expensive since a huge database of meetings would be required. This question is interesting since one could link meetings, their topics and the opinions by different people like a hyperlinked Web environment where the relationships are build dynamically.

Besides these concrete scientific projects there are also long term technology developments that inspired the author. Multimedia documents will allow documents about spoken communication to be grounded in the actual speech and give us back what the sterility of written documentation has deprived us from: The original, personal, emotionally charged, qualified oral account of a person that made a decision or transmitted information. Information access to oral communication is therefore an important step from oral culture over written culture back to oral culture. Oral culture is ultimatively more personable, holds the individuals and organizations accountable for their decisions and errors rather than hiding behind multiple levels of reinterpretation. Indexing spoken language will deepen our personal and organizational memory far beyond the level we have reached so far. Indexing of audiovisual documents may also be used for legitimate and appropriate surveillance applications ensuring the safety of individuals, groups or societies. On the other hand the ability to index oral communications – available in the hands of governments, private organizations, or even individuals – may also be used to monitor individuals at unprecedented levels disenfranchising them of civil liberties and may lead to unfair and non-market conformant business practices. The publication of this work is therefore important since the development of these capabilities is likely pursued by intelligence organizations independently and allows the public to participate in the political control of intelligence gathering operations. Understanding the potential surveillance capabilities also reinforces the need of effective access control to sensor devices especially for new kinds of mobile devices which might be underestimated otherwise.

List of Figures

List of Tables

Glossary

Activity The way people interact could be described as the activity they engage in. It is typically constant in one topical unit and can be characterized by action verbs such as storytelling, discussing, etc. The activity is not only an index but it also determines how something has to be taken: In an interrogation, for example, the answers by the party interrogated may be correct but they are usually not forthcoming. Activities are widely assumed to depend on sequences of dialogue acts. (page 46)

Audioskimming Arons (1997) presents a good review on the work on audio-only interfaces for skimming speech. They are based on rapid playback and jumping in audios. Audio can be played back at a factor of 2-3 without loosing intelligibility. Research in faster playback should include the judgment of activities and speaker initiative since these are an important index. (page 184)

C99 Database Fred Choi has used this database for experiments on topic segmentation (Choi, 2000). It is artificially constructed from Brown corpus texts by concatenating initial text segments. The results on topic segmentation seem to indicate that the artificial nature of this database may affect the results. (page 145)

CallHome Spanish (CHS) CallHome Spanish (CHS) is a corpus of personal telephone calls between family members which is available via the linguistic data consortium (LDC96S35). 120 calls have been recorded and 5-10 mins of each calls are transcribed by the LDC. The data has been annotated with topical segmentations, activities, dialogue acts and games in our group. The annotations were published recently (Waibel et al., 2001b). (page 15)

Dialogue Acts Dialogue acts are the actions that speakers take in a conversation and the granularity is approximately that of a turn. Typical dialogue act types are statements, backchannels and questions. (page 96)

Dialogue Games Dialogue games are short sequence of dialogue acts, often characterized by the first dialogue act in the sequence (example: Question-Answer game). (page 96)

Genre A genre is – similar to an activity – a pattern of communication. However genres are assumed to be an "institutionalized response to a recurrent situation" which gives them a stable form that is highly influenced by the social environment. Genres are often characterized by generic progression, a relatively stable form of stages a communication undergoes. (page 52)

Information Access Hierarchy Audio documents can be divided into (sub-)databases which are fairly different, into spoken communications which can be seen as recording units and topical segments within a spoken communication. Information access to an audio document typically requires to search for the information along that hierarchy (see also Fig. 1.1). (page 3)

Intelligent Meetingroom Rooms metaphors have been used by authors of pervasive computing applications. The intelligent meeting room is adds a documentation function to the room. The scenario in our lab currently consists of the speech recognition engine, the summarization compontent, the integation and visualization platform meeting browser, the dialogue module (this work) and finally auditory scene analysis. (page 4)

Interactional Features Turn lengths, the length of overlaps, pauses and words are very simple measures and may be easy to obtain from the audio signal. They are useful in the discrimination of databases. (page 111)

Keyword-based Approach Traditional information retrieval relies in large on the use of keywords. However, keywords based retrieval is unlikely to perform as well for audio documents from spoken interactions such as meetings as for other databases: People use idiosyncratic keywords, they use fewer of them and they repeat them. Additionally the performance of speech recognizers is rather poor on data such as meetings. (page 8)

Meeting Browser The meeting browser is an interface to meeting records and an integration platform for the various sensors and detectors in the intelligent meeting room framework. The meeting browser has been developed by Michael Bett in our working group at Carnegie Mellon University (Fig. 2.1). (page 23)

Meeting Database Meetings of our own research group have been transcribed and annotated with topical segments and activities. (page 15)

Reinterpretation Just saving a rejoinder on tape, even if we don't forget where it is, is not the same as the production of a written document. The result of the rejoinder need to be interpreted, summarized, retarget beyond the attendants. Reinterpretation is expensive and written documentation combines memorization with reinterpretation. Audio recordings however allow to separate memorization and reinterpretation. (page 26)

Santa Barbara Corpus (SBC) 12 interactions out of a large set of recorded interactions have been released via the LDC. A subcorpus consisting of 7 meeting-like situations has been annotated with topic boundaries and activities. (page 15)

Semantic Word Classes Semantic fields have been used to characterize social distance. In this work WordNet lexicographers classes are used as semantic classes. (page 112)

Speech Recognition Large Vocabulary Continuous Speech Recognition (LVCSR) is still an unsolved problem for meeting corpora since the error rates are around 40%. This word error rate regime makes it unlikely that humans would find the transcripts very useful as compared to the audio itself (Stark et al., 2000). A common application for LVCSR is the detection of keywords for information retrieval and works fairly good for the broadcast news database. Given the high error rate and the fact that there are far fewer unique keywords in rejoinders such as meetings compared to broadcasts the retrieval problem for meetings might be a lot harder. (page 61)

Summarization Automatic summarization is often based on measurements of the "topicality" of segments and redundancy reduction is attempted. Phrases are ranked by the combined score. Klaus Zechner in our working group has adapted automatic summarization to a speech environment. (page 6)

Switchboard A large conversational speech corpus which has been used for speech recognition evaluations as well as dialogue act annotation. (page 17)

Topic The definition of topic in this work is based on the topic intuition of naive subjects. While this definition might be seen as simplistic it has been followed by many authors. Additionally to appealing to their intuition the coders were required to annotate activities on these topical segments at the same time such that they may have been fairly sensitive to changes in the activity. (page 135)

TV Genre Corpus 1067 TV shows have been collected using close captioning and the type of TV program was available from online resources. (page 16)

Words and Parts-Of-Speech Frequent word / parts-of-speech (POS) pairs are usually used directly if the type is frequent enough while the POS information alone is used otherwise. The feature is presented as a histogram, often the most frequent 50 word / POS pairs are represented. (page 111)

Bibliography

International network for social network analysis. Web published in 2000: http://www.heinz.cmu.edu/project/INSNA/. Ressources for social network analysis.

G. D. Abowd. Classroom 2000: An experiment with the instrumentation of a living educational environment. *IBM Systems Journal, Special issue on Pervasive Computing*, 38(4):508–530, October 1999. URL http://www.research.ibm.com/journal/sj/384/abowd.html.

J. Allan, J. Carbonell, G. Doddington, J. P. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, February 1998. URL http://www.itl.nist.gov/iaui/894.01/publications/darpa98/.

J. Allen and M. Core. Draft of damsl: Dialog act markup in several layers. URL http://www.cs.rochester.edu/research/cisd/resources/damsl/. May 21st 1997.

J. Allwood. An activity based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics 76, Dept. of linguistics, University of Göteborg, 1995. URL http://www.ling.gu.se/~jens/publications/index.html. Forthcoming in Bunt & Black (eds.) Approaches to Pragmatics.

J. Allwood and J. Hagman. Some simple automatic measures of spoken interaction. In *Proceedings of the XIV Scandinavian Conference for Linguistics and 8th conference for Nordic and General Linguistics*, volume 72. University of Göteborg, 1994. URL http://www.ling.gu.se/~jens/publications/index.html.

S. Anderson and M. Conway. *COGNITIVE MODELS OF MEMORY*, chapter Representations of Autobiographical Memory, pages 218–246. Psychology Press, Philadelphia, PA, 1997.

B. Arons. *Interactively Skimming Recorded Speech*. PhD thesis, MIT, 1994. URL http://www.media.mit.edu/speech/papers/1994/arons_thesis94.pdf.

B. Arons. Speechskimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer Human Interaction*, 4(1):3–28, March 1997. URL http://www.media.mit.edu/speech/papers/1997/arons_ToCHI97_speechskimmer.pdf.

J. L. Austin. *How to Do Things with Words*. Oxford: Oxford University Press, 1962.

M. Bacchiani, D. Hindle, J. Hirschberg, P. Isenhour, M. Jones, A. Rosenberg, L. Stark, S. Whittaker, and G. Zamchick. Scanmail: Audio navigation in the voicemail domain. In *Human Language Technology Conference (HLT)*, Sand Diego, CA, USA, March 2001.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Books / Addison Wesley, New York, 1999.

M. M. Bahktin. *Speech Genres and other late Essays*, chapter The Problem of Speech Genres. University of Texas Press, Austin, 1986.

B. Baldwin, T. S. Morton, and A. Bagga. Overview of the University of Pennsylvania's Tipster Report. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 151–162, 1999. Omnipress, Inc.

F. Bargiela-Chiappini. *Managing language, The discourse of corporate meetings*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1997.

D. Beeferman, A. Berger, and J. Lafferty. A model of lexical attraction and repulsion. In *Proceedings of the ACL-EACL Joint Conference*, Madrid, Spain, 1997. URL http://www.dougb.com/research.html.

D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177–210, 1999. URL http://www.cs.cmu.edu/~lafferty. Special Issue on Natural Language Learning (C. Cardie and R. Mooney, eds).

V. Belotti and A. Sellen. Design for privacy in ubiquitous computing environments. In *Proceedings of the Third European Conference on Computer-Supported Cooperative Work - ECSCW'93*, pages 77–92, Milan, Italy, 1993. Kluwer.

A. Berger. Convexity, maximum likelihood and all that, 1998. URL http://www.cs.cmu.edu/~aberger.

A. Berger and H. Printz. A comparison of criteria for maximum entropy/minimum divergence feature selection. In *Third Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 96–107, Granada, Spain, 1998.

M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. In *Proceedings of RIAO2000*, Paris, France, April 2000.

D. Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

D. Biber. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(219-241), 1993.

D. Biber, S. Conrad, and R. Reppen. *Corpus Lingusitics: Investigating Language Structure and Use*. Cambridge University Press, 1998.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. *The Longman grammar of spoken and written English*. Longman, London, 1999.

S. Bird, D. Day, J. Garofolo, J. Henderson, C. Laprun, and M. Liberman. Atlas: A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, pages 1699–1706, Athens, 2000. European Language Resources Association (ELRA). URL http://arXiv.org/abs/cs/0007022.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

T. Bray, J. Paol, C. M. Sperberg-McQueen, and E. Maler. Extensible markup language (xml) 1.0 (second edition). W3c recommendation, W3C, 6 October 2000. URL http://www.w3.org/TR/REC-xml.

L. Breure. Development of the genre concep. Technical report, Information and Computing Sciences University of Utrecht, The Netherlands, August 2001. URL http://www.cs.ruu.nl/people/leen/GenreDev/GenreDevelopment.htm. internal report.

W. F. Brewer. *Remembering reconsidered: Ecological and traditional approaches to the study of memory*, chapter Memory for randomly sampled autobiographical events, pages 21–90. Cambridge: Cambridge University Press, 1988.

W. F. Brewer. *Autobiographical memory and the validity of retrospective reports*, chapter Autobiographic memory and survey research, pages 11–20. Springer, 1993.

D. Brickley and R. Guha. Resource description framework (rdf) schema specification 1.0. W3c candididate recommendation, W3C, 27 March 2000. URL `http://www.w3.org/TR/2000/CR-rdf-schema-20000327/`.

J. Bridle. *Neurocomputing: Algorithms, Architectures and Applications*, chapter Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. Springer-Verlag, Berlin, 1990.

E. Brill. A report on recent progress in transformation based error-driven learning. In *DARPA Workshop*, 1994a.

E. Brill. Some advances in transformation-based part of speech tagging. In *Proceeedings of AAAI-94*, 1994b.

R. F. Bruce and J. M. Wiebe. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(2), 1999.

F. D. Buø. *FeasPar - A Feature Structure PARser learning to parse spontaneous speech*. PhD thesis, University of Karlsruhe, 1996.

F. D. Buø and A. Waibel. Feaspar : A feature structure parser learning to parse spoken language. In *COLING*, 1996.

S. Carberry, J. Chu, N. Green, and L. Lambert. Rhetorical relations: Necessary but not sufficient. In *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*, pages 1–4, 1993. URL `http://www.bell-labs.com/user/jencc/papers/all.html`.

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, March 1997.

L. Carlson. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel, 1983.

S. F. Chen and J. Goodman. An empirical study of smoothign techniques for lamguage modeling. In *Proc. of the 34th Annual Meeting of the ACL*, June 1996. also available as `cmp-lg/9606011`.

S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-108, Computer Science Department, Carnegie Mellon University, 1999. URL http://www.cs.cmu.edu/~sfc.

F. Choi. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, Seattle, USA, 2000. Available with software at: http://www.cs.man.ac.uk/~choif/ http://xxx.lanl.gov/abs/cs.CL/0003083.

J. Choi, D. Hindle, F. Pereira, A. Singhal, and S. Whittaker. Spoken content-based audio navigation (SCAN). In *Proceedings of the ICPhS-99*, 1999. URL http://www.research.att.com/~stevew/.

H. H. Clark. *Using Language*. Cambridge University Press, 1996.

B. Clarkson and A. Pentland. Unsupervised clustering of ambulatory audio and video. In *ICASSP*, 1999.

D. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.

M. G. Core and J. Allen. Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley-Interscience, 1991.

L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle. Platform for privacy preferences 1.0 (p3p1.0) specification. Working Draft 15, W3C, September 2000. Also available http://www.w3.org/TR/2000/WD-P3P-20000915.

I. Csiszar. *Maximum entropy and bayesian methods*, chapter Maxent, mathematics and information theory. Kluwer Academic Publishers, 1996.

P. Delacourt and C. Wellekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1-2):111–126, 2000. URL http://www.elsevier.nl/inca/publications/store/5/0/5/5/9/7/. Special Issue on Accessing Information in Spoken Audio.

S. Deligne and F. Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In ICASSP-95 ICASSP-95.

S. Eggins and D. Slade. *Analysing Casual Conversation*. Cassell, 1998.

S. Eickeler and S. Mueller. Content-based video indexing of tv broadcast news using hidden markov models. In *ICASSP*, 1999.

Eldridge, Lamming, and Flynn. Does a video diary help recall. Technical Report EPC-1991-124, Xerox Research Center Europea, 1991. URL `http://www.xrce.xerox.com/publis/cam-trs/html/lamming.htm`.

M. Federico. A system for the retrieval of italian broadcast news. *Speech Communication*, 32(1-2):37–47, 2000. URL `http://www.elsevier.nl/inca/publications/store/5/0/5/5/9/7/`. Special Issue on Accessing Information in Spoken Audio.

C. Fellbaum, editor. *WordNet – An Electronic Lexical Database*. MIT press, 1998.

M. Finke, M. Lapata, A. Lavie, L. Levin, L. M. Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. Clarity: Automatic Discourse and Dialogue Analysis for a Speech and Natural Language Processing System. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

M. Finke, T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty, and A. Waibel. Switchboard evaluation report. In *Proceedings of LVCSR Hub 5 Workshop*, April 1996.

R. M. Ford, C. Robson, D. Temple, and M. Gerlach. Metrics for shot boundary detection in digital video sequences. *Springer: Multimedia Systems*, 8(1):37–46, 2000.

W. Franke. *Elementare Dialogstrukturen*. Tübingen, 1990.

L. Friedman. Web published in 2000: `http://www.heinz.cmu.edu/project/INSNA/na_inf.html`.

G. Fritz and F. Hundschnur. *Handbuch der Dialoganalyse*. Niemeyer, Tuebingen, 1994.

J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track : A success story. In E. Voorhees, editor, *Text Retrieval Conference (TREC) 8*, Gaithersburg, Maryland, USA, 1999. November 16-19. URL `http://trec.nist.gov/pubs/trec8/t8_proceedings.html`.

M. Gavaldà. Soup: A parser for real-world spontaneous speech. In *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy, February 2000.

J. P. Gee. Units in production of narrative discourse. *Discourse Processes*, 9: 391–422, 1986.

Georgia Institute of Technology. eclass lectures. URL http://eclass.cc.gatech.edu/zenpad/db/.

P. Geutner, M. Finke, and P. Scheytt. Adaptive vocabularies for transcribing multilingual broadcast news. In *ICASSP*, 1998.

J. Goldstein and J. Carbonell. Summarization: Using MMR for Diversity-Based Reranking and Evaluating Summaries. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 181–196, 1999. Omnipress, Inc.

A. Gorin. On automated language acquistition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

D. Goutsos. *Modeling Discourse Topic: Sequential Relations and Strategies in Expository Text*. Number LIX in Advances in Discourse Processes. Ablex Publishing, 1997.

R. Gross, J. Yang, and A. Waibel. Growing gaussian mixture model for pose invariant face recognition. In *International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, September 2000.

B. Grosz and C. Sidner. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):172–204, 1986.

M. Halliday. *An introduction to functional grammar*. Oxford University Press, 1994.

M. Halliday and R. Hasan. *Cohesion in English*. Longman Group, 1976.

W. Hanks. Discourse genre in a theory of practice. *American Ethnologist*, 14(4): 688–692, 1988.

P. Hart, N. Nielsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on SCC*, 4, 1968.

M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March 1997.

D. J. Herrmann. *Autobiographical memory and the validity of retrospective reports*, chapter The validity of retrospective reports as a function of the directness of retrieval processes, pages 21–31. Springer, 1993.

F. Heylighen and J.-M. Dewaele. Formality of language: definition, measurement and behavioural determinants. internal report, Center “Leo Apostel”, Free University of Brussels, 1999. URL `http://pespmc1.vub.ac.be/papers/PapersFH.html`.

J. Hirschberg and C. Nakatani. Acoustic indicators of topic segmentation. In *ICSLP*, Sidney, Australia, 1998.

D. I. Holmes. The evolution of stylometrie in humanities scholarship. *Litrary and Linguistic Computing*, 13(3):111–117, 1998.

W. Hürst, R. Müller, and C. Mayer. Multimedia information retrieval from recorded presentations. In *Proceedings of ACM SIGIR 2000*, Athens, Greece, July 2000.

ICASSP-95. *ICASSP*, 1995. IEEE.

P. Isokoski. A minimal device-independent text input method. Technical Report A-1999-14, Department of Computer Science, University of Tampere, 1999. URL `http://www.cs.uta.fi/~poika/g/g.html`.

S. Jaeger. Npen++: An on-line handwriting recognition system. In *7th International Workshop on Frontiers in Handwriting Recognition*, pages 249–260, Amsterdam, 2000.

F. Jelinek. *Readings in Speech Recognition*, chapter Self-Organized Language Modeling for Speech Recognition. Morgan Kaufmann, 1989. edited by K.-F. Lee and A. Waibel.

D. Jensen. Prospective assessment of ai technologies for fraud detection: A case study. In *Working papers of the AAAI-97 Workshop on Artificial Intelligence Approaches to Fraud Detection and Risk Management*, 1997. URL `http://eksl-www.cs.umass.edu/~jensen/papers/aaaiws97a.html`.

T. Joachims. *The Maximum-Margin Approach to Learning Text Classifiers: Methods, Theory, and Algorithms*. PhD thesis, Dortmund, 2000. URL `http://ais.gmd.de/~thorsten/svm_light/`.

Journal of Social Structure. Journal of social structure. Web published in 2000.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. V. Ess-Dykema. Automatic Detection of Discourse Structure for Speech Recognition and Understanding. In *IEEE Workshop on Speech Recognition and Understanding*, September 1997a.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. V. Ess-Dykema. SWBD Discourse Language Modeling Project, Final Report. Technical report, Johns Hopkins LVCSR Workshop-97, 1997b.

D. Jurafsky and L. Shriberg. Switchboard discourse type labeling project "switchboard damsl" (swbd-damsl) labeling system. CODER'S MANUAL, Draft 5, April 29th 1997.

M.-Y. Kan, J. Klavans, and K. R. McKeown. Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop on Very Large Corpora (WVLC-6)*, pages 197–205, Montreal, Canada, August 1998.

J. Karlgren. *Stylistic Experiments for Information Retrieval*. PhD thesis, University of Stockholm, February 2000. URL `http://www.sics.se/~jussi/Artiklar/2000_PhD/`.

J. Kattán-Ibarra. *Spanish grammar: a functional guide*. Cox & Wyman Ltd., Reading, Berkshire, 1991.

D. Kaufer. The docuscope workbook. A programmed learning guide through the Functional Categories of Written English Prose.

D. S. Kaufer and B. S. Butler. *Designing Interactive Worlds with Words, Principles of Writing as representational composition*. Lawrence Erlbaum Associates, Publishers, London, 2000.

J. M. Keenan and S. D. Baillet. Memory of personally and socially significant events. In R. Ncikerson, editor, *Attention and Performance VIII*. Lawrence Erlbaum Associates, Publishers, 1982.

J. M. Keenan, B. McWhinney, and D. Mayhew. *Memory observed*, chapter Pragmatice in Memory: A Study of Natural Conversation. W.H. Freeman Company, New York, 1982. reprinted from *Journal of Verbal Learning and Verbal Behaviour*, 1977, 16, 549-560.

B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association* *for Computational Linguistics*, pages 32–38. Morgan Kaufmann Publishers, San Francisco CA, 1997. URL `http://xxx.lanl.gov/abs/cmp-lg/9707002`.

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In ICASSP-95 ICASSP-95.

J. L. Kolodner. Reconstructive memory: A computer model. *Cognitive Science*, 7:281–328, 1983.

H. Kozima. Text segmentation based on similarity between words. In *Proceedings of the ACL*, pages 286–288, Ohio, USA, 1993.

T. Kristjansson, T. Huang, P. Ramesh, and B. Juang. A unified structure-based framework for indexing and gisting of meetings. In *IEEE International Conference on Multimedia Computing and Systems*, 1999. URL `http://newgist.uwaterloo.ca/~trausti/Papers/Papers.html`.

F. Kubala, S. Colbath, D. Liu, and J. Makhoul. Rough'n'Ready: a meeting recorder and browser. *ACM Computing Surveys*, 31(7), September 1999. URL `http://www.acm.org/pubs/citations/journals/surveys/1999-31-2es/a7-kubala/`. Article No. 7.

R. Kuhn and R. de Mori. A cache-base natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and machince Intelligence*, 12 (6):570–583, June 1990.

W. Labov and J. Waletzky. Narrative analysis: oral versions of personal experience. In J. Helm, editor, *Essays of the Verbal and Visual Arts*, pages 12–14, Washington DC, 1967. American Ethnological Society, proceedings of Spring Meeting 1967, University of Washington Press.

M. Lampert and S. Ervin-Tripp. *Talking Data: Transcription and coding in discourse research*, chapter Structured coding for the study of language and social interaction, pages 169–206. Lawrence Erlbaum Associates, Hillsdale NJ, 1993.

J. A. Landay and R. C. Davis. Making sharing pervasive: Ubiquitous computing for shared note taking. *IBM Systems Journal*, 38(4), 1999. URL `http://www.research.ibm.com/journal/sj/384/landay.html`.

LDC2000S85. Santa barbara corpus of spoken american english part-i, 2000. URL `http://www.ldc.upenn.edu/Catalog/LDC2000S85.html`.

LDC93S7. Switchboard, 1993. URL `http://www.ldc.upenn.edu/Catalog/LDC93S7.html`.

LDC96S35. Callhome spanish, lexicon, speech and transcripts, 1996. URL http://www.ldc.upenn.edu/Catalog/LDC96T17.html. catalogue number LDC96L16, LDC96S35 catalogue number LDC96L16, LDC96S35 .

D. Lee. *Modelling Variation in Spoken And Written English: the Multi-Dimensional Approach Revisited*. PhD thesis, Lancaster University, 2002. URL http://www.davidlee00.freeserve.co.uk/dave.htm. To be published by Routledge.

J. A. Levin and J. A. Moore. Dialogue games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420, 1977.

L. Levin, K. Ries, A. Thymé-Gobbel, and A. Lavie. Tagging of Speech Acts and Dialogue Games in Spanish Call Home. In *Proceedings of the ACL workshop on Discourse Tagging*, 1999.

S. C. Levinson. Activity types and languagen. *Linguistics*, 17(5):365–399, 1979.

P. Linell. *The written language bias in linguistics*. Tema Kommunikation, Linköping University, 1982. URL http://eserver.org/langs/linell/.

P. Linell. *The Dynamics of Dialogue*, chapter The power of dialogue dynamics. Harvester Wheatsheaf, 1990.

P. Linell. *Approaching Dialogue: talk, interaction and contexts in dialogical perspectives*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1994.

P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, 26:415–442, 1988.

Linguistic Data Consortium (LDC). Catalogue. http://www.ldc.upenn.edu/.

A. C. Long, J. A. Landay, and L. Rowe. Pda and gesture use in practice: Insights for designers of pen-based user interfaces. Technical report, Berkley Multimedia Research Center, 1997. URL http://bmrc.berkeley.edu/frame/research/publications/.

R. E. Longacre. *The grammar of discourse*. Plenum Press, 2nd edition, 1996.

R. Malkin. Experiments in environment detection and scene acquisition. draft technical report, 2002.

I. Mani and E. Bloedern. Multi-document summarization by graph search and merging. In *Proceedings of AAAI-97*, pages 622–628. AAAI, 1997.

W. C. Mann and S. Thomson. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281, 1988.

D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto, December 1997. Also published as Technical Report CSRG-371, Computer Systems Research Group, University of Toronto.

B. Martinovski. Speech and activity style. comparative study of two activities - an interview and a discussion. Gothenburg Papers in Theoretical Linguistics 79, Dept. of linguistics, University of Göteborg, 1996.

B. Martinovsky. *The Role of Repetitions and Reformulations in Court Proceedings - A Comparision of Sweden and Bulgaria*. PhD thesis, Göteburg University, Department of Linguistics, Sweden, 2000.

J. M. Martínez. Mpeg-7 overview (version 3.0). Technical Report W3752, Moving Picture Experts Group (MPEG) a working group of ISO/IEC, La Baulle, October 2000. URL `http://www.cselt.it/mpeg/`.

M. McKenna and E. Liddy. Multiple and Single Document Summarization Using DR-LINK. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 215–222, 1999. Omnipress, Inc.

McTavish, Litkowski, and Schrader. A computer content analysis approach to measuring social distance in residential organizations for older people. In *Conference of the Society for Conceptual and Content Analysis by Computer*, Mannheim, Germany, September 1995. `http://www.clres.com/index.html#papers`.

I. Mikic, K. Huang, and M. Trivedi. Activity monitoring and summarization for intelligent environments. In *Workshop on Human Motion*, Austin, Texas, December 2000. URL `http://swiftlet.ucsd.edu/research/papers/aviary/HUMOWorkshop.pdf`.

E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994.

T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle, and P. Zellweger. “i’ll get that off the audio”: A case study of salvaging multimedia meeting records. In *CHI 97*, 1997. URL http://www.acm.org/sigchi/chi97/proceedings/paper/tpm.htm.

Multimedia, Teleteaching and Electronic publishing group, Department of Applied Science, University of Freiburg. Local electronic library. URL http://ad.informatik.uni-freiburg.de/mmgroup.lib.

M. Munk. Shallow statistical parsing for machine translation. Diplomarbeit, Universität Kalrsruhe, 1999.

S. K. Murthy, S. Kasif, S. Salzberg, and R. Beigel. OC1: Randomized induction of oblique decision trees. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 322–327, Washington, D.C., 1993.

M. Nagata and T. Morimoto. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15: 193–203, 1994.

H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8:1–35, 1994.

C. Ng, R. Wilkinson, and J. Zobel. Experiments in spoken document retrieval using phoneme n-grams. *Speech Communication*, 32(1-2):61–77, 2000. URL http://www.elsevier.nl/inca/publications/store/5/0/5/5/9/7/. Special Issue on Accessing Information in Spoken Audio.

K. NG. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, MIT, February 2000. URL http://www.sls.lcs.mit.edu/˜kng/.

K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999. URL http://www.cs.cmu.edu/˜lafferty/.

U. Ohler, S. Harbeck, and H. Niemann. Discriminative training of language model classifiers. In *Proceedings of the Eurospeech*, volume 4, pages 1607–1610, Budapest, Hungary, September 1999. URL http://faui58f.informatik.uni-erlangen.de/HTML/English/Literatur/Alles/Alles/Alles.html.

W. J. Orlikowski and J. Yates. Genre repertoire: Norms and forms for work and interaction. Technical report, MIT Sloan School, Center for Coordination Science, March 1994. http://ccs.mit.edu/papers/CCSWP166.html.

Y. Pan and A. Waibel. The effects of room acoustics on MFCC speech parameters. In *Proceedings of the ICSLP*, Beijing, China, 2000.

R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103, March 1997. 139.

S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features in random fields. *IEEE Transactions on Pattern Analysis and machince Intelligence*, 19(3):1–13, March 1997.

G. A. Plum. *Text and contextual conditioning in spoken English: A genre-based approach*. PhD thesis, University of Sidney, 1988.

T. Polzin. *Detecting Verbal and Non-Verbal Cues in the Communication of Emotion*. PhD thesis, Carnegie Mellon University, November 1999.

T. S. Polzin and A. Waibel. Detecting emotions in speech. In *Proceedings of the CMC*, 1998.

J. M. Ponte and B. W. Croft. Text segmentation by topic. In *Proceedings of the first European Conference on research and advanced technology for digital libraries*, 1997. U.Mass. Computer Science Technical Report TR97-18.

M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

H. Printz. Fast computation of maximum entropy / minimum divergence feature gain. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, December 1998.

M. Przybocki and A. Martin. The 1999 nist speaker recognition evaluation speaker detection and speaker tracking. In *Proceedings of the Eurospeech*, Budapest, Hungary, September 1999. URL http://www.nist.gov/speech/tests/spk/1999/euro99_v2/sld001.htm.

F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. E. McCullough, and N. Furuyama. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of the Computer Vision and Pattern Recognition CVPR*, 2000. URL http://vislab.cs.wright.edu/Publications/Queetal00.html.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1992.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A comprehensive grammar of the English language*. Longman, 1985.

D. Radev and K. McKeown. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 24(3):469–501, September 1998.

D. R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *ANLP/NAACL 2000 Workshop*, pages 21–29. ACL, April 2000.

G. Radvansky and R. Zacks. *COGNITIVE MODELS OF MEMORY*, chapter Retrieval of Situation-specific Information, pages 218–246. Psychology Press, Philadelphia, PA, 1997.

N. Reithinger, R. Engel, M. Kipp, and M. Klesen. Predicting dialogue acts for a speech-to-speech translation system. In *ICSLP*, 1996.

S. Renals and T. Robinson. Accessing information in spoken audio. *Speech Communication*, 32(1-2):1–3, 2000. URL http://www.elsevier.nl/inca/publications/store/5/0/5/5/9/7/. Special Issue on Accessing Information in Spoken Audio.

J. C. Reynar. *Topic segmentation: Algorithms and applications*. PhD thesis, Computer and Information Science, University of Pennsylvania, 1998. URL http://www.cis.upenn.edu/˜ircs/reports/trs/abstracts98.html. Institute for Research in Cognitive Science (IRCS), University of Pennsylvania, Technical report: IRCS-98-21.

M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. of the IEEE Int. Conf. on Neural Networks*, pages 586–591, 1993.

K. Ries. HMM and Neural Network Based Speech Act Classification. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 497–500, Phoenix, AZ, March 1999a.

K. Ries. Towards the Detection and Description of Textual Meaning Indicators in Spontaneous Conversations. In *Proceedings of the Eurospeech*, volume 3, pages 1415–1418, Budapest, Hungary, September 1999b.

K. Ries. Segmenting Conversations by Topic, Initiative and Style. In A. Coden, E. W. Brown, and S. Srinivasan, editors, *SIGIR Workshop: Information Retrieval Techniques for Speech Applications [based on the workshop "Information Retrieval Techniques for Speech Applications", held as part of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in New Orleans, USA, in September 2001]*, volume 2273 of *Lecture Notes in Computer Science*. Springer, 2002. ISBN 3-540-43156-X.

K. Ries, F. D. Buø, and A. Waibel. Class Phrase Models for Language Modeling. In *Proceedings of the ICSLP*, Philadelphia, USA, 1996.

K. Ries, L. Levin, L. Valle, A. Lavie, and A. Waibel. Shallow Discourse Genre Annotation in CallHome Spanish. In *Proceecings of the International Conference on Language Ressources and Evaluation (LREC-2000)*, Athens, Greece, May 2000.

K. Ries and A. Waibel. Activity Detection for Information Access to Oral Communication. In *Human Language Technology Conference (HLT)*, Sand Diego, CA, USA, March 2001.

D. Roy. Newscomm: A hand-held device for interactive access to structured audio. Master's thesis, MIT Media Laboratory, 1995. URL `http://dkroy.www.media.mit.edu/people/dkroy/publications.html`.

D. Roy and C. Malamud. Integration of a large text and audio corpus using speaker identification. In *Proceedings of the AAAI Spring Symposium on the Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, Palo Alto, 1997.

K. Samuel, S. Carberry, and K. Vijay-Shanker. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics*, Waterloo, Ontario, Canada, 1999.

C. Saraceno. Video context extraction and representation using a joint audio and video processing. In *ICASSP*, 1999.

R. Schank. *Scripts, plans, goals and understanding*. Hilldale, N.J.:Erlbaum, 1977.

R. Schank. *Dynamic Memory*. New York, Cambridge Univesity Press, 1982.

R. E. Schapire and Y. Singer. Boostexter: A system for multiclass multi-label text categorization. submitted, 1999.

D. Schiffrin, editor. *Approaches to discourse*. Cambridge: Blackwell, 1994. P302 S334.

K. Schubert. Grundfrequenzverfolgung und deren anwendung in der spracherkennung. Master's thesis, Universität Karlsruhe, 1999.

J. Searle. The classification of illocutionary acts. *Language in Society*, 5:1–24, 1976.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M. Meteer, K. Ries, A. Stolcke, P. Taylor, and C. V. Ess-Dykema. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):439–487, 1998. URL `http://xxx.lanl.gov/abs/cs.CL/0006024`.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody modeling for automatic sentence and topic segmentation from speech. *Speech Communication*, 32(1-2):127–154, 2000. Special Issue on Accessing Information in Spoken Audio.

A. Singhal and F. Pereira. Document expansion for speech retrieval. In *In Proceedings of SIGIR*, 1999. URL `http://www.research.att.com/~singhal/`.

S. Slembrouck. What is meant by discourse analysis? working document online, 2001. URL `http://bank.rug.ac.be/da/da.htm`.

L. Stark, S. Whittaker, and J. Hirschberg. Asr satisficing: The effects of asr accuracy on speech retrieval. In *ICSLP*, Beijing, China, October 2000. URL `http://www.research.att.com/~stevew/icslp00-asr.pdf`.

G. C. Stein, T. Strzalkowsi, and G. B. Wise. Summarizing multiple documents using text extraction and interactive clustering. In *Proceedings of the Conference Pacific Association for Computat ion Linguistics*, pages 200–208. Pacling, August 1999.

R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *Proceedings of the ACM Multimedia*, 1999.

R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, September 2000.

L. Stifelman. A discourse analysis approach to structured speech. volume Empirical methods in Discourse Interpretation and Generation of *AAAI 1995 Spring Symposium Series*, Stanford University, March 27-29 1995.

L. Stifelman, B. Arons, and C. Schmandt. The audio notebook: paper and pen interaction with structured speech. In *Proceedings of the SIG-CHI on Human factors in computing systems*, pages 182–189, 2001. URL http://www.media.mit.edu/speech/papers/2001/stifelman_CHI01_audio_notebook.pdf.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3), September 2000.

K. Takahashi. Remarks on a real-time 3d human body posture estimation method using trinocular images. In *International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, September 2000.

D. Tannen. *Converational Style*. Language and Learning for Human Service Professions. Ablex Publishing Corporation, 1984.

D. Tannen, editor. *Framing in Discourse*. Oxford University Press, 1993.

P. Taylor, S. King, S. Isard, H. Wright, and J. Kowtko. Using intonation to constrain language models in speech recognition. In *EUROSPEECH*, Rhodes, Greece, 1997.

Textware Solutions. Fitaly keyboard, 2000. URL http://www.twsolutions.com/fitaly/fitaly.htm. Product information.

A. Thymé-Gobbel, L. Levin, K. Ries, and L. Valle. Dialogue act, dialogue game, and activity tagging manual for spanish conversational speech. Technical report, Carnegie Mellon University, 2001. in preperation, direct requests to mailto:ries@cs.cmu.edu.

van Bretan, J. Dewe, A. Hallberg, J. Karlgren, and N. Wolkert. Genres defined for a purpose, fast clustering, and an iterative information retrieval interface. In *Eighth DELOS Workshop on User Interfaces in Digital Libraries Långholmen*, pages 60–66, October 1998.

H. D. Wactlar. Auto summarization and visualization across multiple video documents and libraries. http://www.informedia.cs.cmu.edu/dli2/, 2000.

H. D. Wactlar, M. G. Christel, A. G. Hauptmann, and Y. Gong. Experience on demand. http://www.informedia.cs.cmu.edu/eod/, 2000.

H. D. Wactlar, A. G. Hauptmann, and M. J.Witbrock. Informedia news-on demand: Using speech recognition to create a digital video library. Technical Report 109, Carnegie Mellon University, Computer Science Department, march 1998. URL http://reports-archive.adm.cs.cmu.edu/anon/1998/abstracts/98-109.html.

A. Waibel, M. Bett, and M. Finke. Meeting browser: Tracking and summarising meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in Automatic Meeting Record Creation and Access. In *ICASSP*, Salt Lake City, Utah, USA, 2001a.

A. Waibel, A. Lavie, L. Levin, K. Ries, and L. Valle-Argueta. CallHome Spanish Dialogue Act Annotation, 2001b. URL http://www.ldc.upenn.edu/Catalog/LDC2001T61.html. Catalogue Number LDC2001T61.

A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze. Multilingual speech recognition. In *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer-Verlag, 2000.

M. A. Walker and S. Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *In Proc. 28th Annual Meeting of the ACL*, 1990. URL http://xxx.lanl.gov/abs/cmp-lg/9504007.

H. Wang. Experiments in syllable-based retrieval of broadcast news speech in mandarin chinese. *Speech Communication*, 32(1-2):49–60, 2000. URL http://www.elsevier.nl/inca/publications/store/5/0/5/5/9/7/. Special Issue on Accessing Information in Spoken Audio.

V. Warnke, S. Harbeck, E. Nöth, H. Niemann, and M. Levit. Discriminative estimation of interpolation parameters for language model classifiers. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 525–528, Phoenix, AZ, March 1999. URL http://faui58f.informatik.uni-erlangen.de/HTML/English/Literatur/Alles/Alles.html.

V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated dialog act segmentation and classification using prosodic features and language models. In *Eurospeech*, pages 207–210, 1997.

S. Wermter and V. Weber. Screen: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *JAIR*, 6:35–85, 1997. http://www.jair.org/abstracts/wermter97a.html.

S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of SIGIR99 Conference on Research and Development in Information Retrieval*, pages 26–33, 1999. URL http://www.research.att.com/~stevew/.

S. Whittaker, P. Hyland, and M. Wiley. Filochat: handwritten notes provide access to recorded conversations. In *In Proceedings of CHI94 Conference on Computer Human Interaction*, pages 271–277, 1994. URL http://www.research.att.com/~stevew/filochat_chi94.pdf.

E. Wiener, J. O. Pederson, and A. Weigend. A neural network based approach to topic spotting. In *Proceedings of the Fourth Anual Symposium on Aocument Analysis and Information Retrieval (SIDAIR'95)*, 1995.

L. D. Wilcox, B. N. Schilit, and N. Sawhney. Dynomite: A dynamically organized ink and audio notebook. In *CHI 97 Conference Proceedings*, pages 186–193, 1997. URL http://www.fxpal.xerox.com/PapersAndAbstracts/abstracts/wil97.htm.

M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, and A. Waibel. A modular approach to spoken language translation for large domains. In *AMTA-98*, 1998.

H. Wright. Automatic utterance type detection using suprasegmental features. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 1403–1406, Sydney, Australia, December 1998.

W. Xiong, C.-M. Lee, and R.-H. Ma. Automatic video data structuring through shot partitioning and key-frame computing. *Springer: Machine Vision and Applications*, 10(2):51–65, 1997.

XNS. P3p, xns, and internet privacy. Whitepaper, XNS Public Trust Organization, September 2000. Also available at http://www.xns.org/xns/whitepapers/privacy/.

Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden markov model approach to text segmentation and event tracking. In *Proceedings of ICASSP*, volume 1, pages 333–336, Seattle, WA, May 1998.

Y. Yang. An evaluation of statistical approaches to text categorizatio. *Journal of Information Retrieval*, 1(1/2):67–88, 1999.

Y. Yang. A study on thresholding strategies for text categorization. In *Proceedings of SIGIR*, pages 137–145, New Orleans, 2001.

Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of SIGIR*, pages 42–49, 1999.

H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel. New developments in automatic meeting transcription. In *Proceedings of the ICSLP*, Beijing, China, October 2000.

K. Zechner. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, November 2001. CMU-LTI-01-168.

K. Zechner and A. Waibel. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING*, Saarbrücken, Germany, 2000a.

K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-2000, Seattle, WA, April/May*, pages 186–193, 2000b.

A. Zell. Snns user manual, version 3.0. Technical Report 3, University of Stuttgart, Institute for Parallel and Distributed High Perfomance Systems, 1993.

T. Zeppenfeld, M. Finke, K. Ries, and A. Waibel. Recognition of Conversational Telephone Speech using the Janus Speech Engine. In *Proceedings of the ICASSP'97*, München, Germany, 1997.

G. Zipf. *The Psycho-Biology of Language*. Houghton Millin, 1935.