EFFICIENT HANDLING OF MULTILINGUAL LANGUAGE MODELS

Christian Fügen, Sebastian Stüker, Hagen Soltau, Florian Metze

Interactive Systems Labs
University of Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany

Tanja Schultz

Interactive Systems Labs
Carnegie Mellon University
School of Computer Science
Pittsburgh, PA 15221, USA

ABSTRACT

In this paper we introduce techniques for building a multilingual speech recognizer. More specifically, we present a new language model method that allows for the combination of several monolingual into one multilingual language model. Furthermore, we extend our techniques to the concept of grammars. All linguistic knowledge sources share one common interface to the search engine. As a consequence, new language model types can be easily integrated into our Ibis decoder. Based on a multilingual acoustic model we compare multilingual statistical n-gram language models with multilingual grammars. Results are given in terms of recognition performance as well as resource requirements. They show that (a) n-gram LMs can be easily combined at the meta level without major loss in performance, (b) grammars are very suitable to model multilinguality, (c) language switches can be significantly reduced by using the introduced techniques, (d) the resource overhead for handling multiple languages in one language model is acceptable, and (e) language identification can be done implicitly during decoding.

1. INTRODUCTION

The increasing interest in multilingual mobile devices has generated a need for small footprint engines and speech recognition systems which can not only handle input from many languages, but also can switch seamlessly between those languages. The problems associated with building a multilingual speech recognizer are closely related to the multi-domain case. A solution to the first problem might also solve the second as shown by Hazen et. al [1] for a FST-based recognizer. As of today, the majority of speech recognizers only processes one language (and one domain) at a time. Recently, many groups have investigated the potential of multilingual acoustic models, but only little research has been done on handling language models for multiple languages [2, 3] or combining both multilingual acoustic and language models, into one single system [4, 5].

Over the years, the Interactive Systems Labs have been involved in various multilingual and cross-cultural projects and has gained significant experience in building human-human interfaces for speech translation and multilingual speech applications, as well as human-machine interfaces in many different languages. As a consequence, our work is motivated by the strong need for a truly multilingual speech recognizer that is flexible enough for use with various applications. The work presented here focuses on a speech engine that simultaneously serves such diverse applications as the human-robot interface of the SFB project[1], the interface to an intelligent room in the FAME project[2], and our mobile linguistic tourist assistant, LingWear.

In the SFB project, a very flexible human-machine interface is required that allows to seamlessly switch between several languages and domains. Our interface, combined with a dialogue manager, allows those switches, and, together with grammar based speech recognition, the dialogue manager even manipulates the grammar at the rule level to achieve a better recognition performance on user responses. In the FAME project the interface is required to handle verbal interactions in different languages and multiple domains but also to implicitly identify the language during decoding. Our mobile linguistic tourist assistant LingWear [6] not only requires a flexible multilingual speech interface, but has strict limitations on available computing and memory resources. As a consequence, we use semantic grammars to model the linguistic system knowledge. Furthermore, since LingWear also integrates speech-to-speech translation it is necessary to share as much knowledge between the different input languages as possible.

In order to meet the above described requirements, we introduce a new language modeling method, *multilingual Meta-LM*, that allows the user to easily combine several monolingual language models into one multilingual LM. Furthermore, we extend the research in multilingual lan-

[1] Sonderforschungsbereich SFB, No. 588, "Humanoide Roboter - Lernende und kooperierende multimodale Roboter"

[2] FAME (Facilitating Agent for Multicultural Exchange) and is funded by the European Commission as IST-2000-28323

guage modeling to the concept of grammars, which often prove to be superior to statistical language models on small domains. We apply the Meta-LM approach to monolingual n-gram language models and compare it to our multilingual grammars in terms of recognition performance and resource requirements.

Since we combine the multilingual language models with our multilingual acoustic model [7], we can build a truly multilingual speech recognizer that allows seamless switching between languages (and domains).

2. LANGUAGE MODELING IN THE IBIS DECODER

The Ibis decoder [8] was developed at the University of Karlsruhe as part of our Janus Recognition Toolkit (JRTk). Besides several other advantages such as smaller memory usage and higher recognition speed, Ibis allows us to decode along context free grammars in addition to the classical statistical n-gram language models. In the following sections, we describe our common language model interface to the single pass search engine and some features of our context free grammar implementation.

2.1. LM Interface and Vocabulary Mapper

The common language model interface abstracts from different linguistic knowledge sources, such as phrase language models, grammars or interpolated linguistic knowledge sources. It makes it easy to attach new language model types and consists mainly of three functions, which are usually called by the search engine:

lks.createLCT () creates the initial linguistic context, which consists usually only of the begin of a sentence,

lks.scoreArray (LCT) gives back the array of scores for all language model words given a specific linguistic context,

lks.extendLCT (LCT, LVX) extends a given linguistic context by a given language model word to form a new linguistic context,

In the case of a grammar, a linguistic context (LCT) is equal to a grammar state or, in the case of an n-gram language model, to the n-gram history of length $n-1$. The term LVX refers to the language model vocabulary index. A vocabulary mapper is used to map search vocabulary[3] words to language model words. In the standard case of an n-gram language model, each search vocabulary word is stripped of its variant information and directly mapped to the language model word with the same spelling. The mapping has to be surjective, which means that all search words unknown by the LM are mapped to a special unknown LM word.

[3]The search vocabulary is not necessarily identical to the pronunciation lexicon, e.g. two items in the search vocabulary can rely on the same pronunciation.

2.2. Context Free Grammars

Using grammars instead of n-gram language models is especially an advantage in small domains, where less domain dependent training data is available for n-gram language models. Rather than compiling one finite state graph out of all the terminals given by the grammars, we use a more dynamic approach, where several rule based finite state graphs consisting of terminals and non-terminals, are linked together by their non-terminal symbols. During decoding, a rule stack gives us the ability to enter and leave the linked finite state graphs. This kind of network organization has usually a smaller memory footprint and results in higher flexibility for using grammars in speech recognition in combination with a dialogue management system. Furthermore, it enables us to work with real context free grammars.

Several domain dependent sub-grammars can be activated/deactivated and loaded at run time. The activation/deactivation mechanism goes all the way to the rule level, giving the dialogue management system the full control over the speech recognizer. Furthermore, it is also allowed to penalize grammars or rules by giving them a penalty factor. Another feature is that grammars can be expanded on the fly by new rules or terminals without restarting the recognizer. Even new words can be added to the grammar and the search network on the fly. In most cases we work with non-statistical grammars, i.e. each transition to the next word has the same language model score.

To cope with spontaneous non-verbal speech events and non-human noises, we are using the mechanism of filler words in the decoder, which can potentially occur between any two terminals. Instead of asking the language model for their score, a predefined filler penalty is applied.

3. MULTILINGUAL LANGUAGE MODELING

The goal of multilingual language modeling is to be able to recognize utterances from multiple languages with the use of a single recognition engine. This not only requires a multilingual language model, but also a multilingual acoustic model. We studied the latter problem already extensively [7] and will apply our results in this paper.

Experiments in multilingual n-gram language modeling have been performed in [2, 3, 4, 5] and can be divided by the degree to which language switching is allowed. [4] merged the corpora of two languages to calculate one combined trigram language model, thus allowing language switching at any time. They introduced a penalty factor to control language switches, but could not regain full performance

compared to the monolingual case. [2] built a four-lingual language model that allows language switching through a common backoff-node which was implemented as a pause model, i.e. language switching is possible by inserting a pause. [3] calculated n-grams on two monolingual corpora and introduced one common sentence-begin and sentence-end symbol, therefore language switching is possible at the sentence level. With this setup, they could achieve good results if the amount of text was the same for both languages.

3.1. Multilingual N-gram LMs

3.1.1. Previous Work

In [5], we compared three approaches to multilingual n-gram language modeling. In the first, we concatenated two language tagged training corpora, and calculated a language model on the new corpus as introduced by [4]. This approach fails when the sizes of the training corpora are highly unbalanced, like in our case, where the English corpus was about 218 times larger than the German one. On average, German n-grams were assigned a smaller probability than English n-grams, and, therefore, German words were often incorrectly recognized as English ones. In our second approach we interpolated between two monolingual language models using constant weights. Because of the unbalanced corpora, less n-grams for a specific n were seen for German compared to English. So most of the n-grams probabilities in the German language model were higher than most of the n-gram probabilities in the English model. Now English words were often incorrectly recognized as German ones because the German n-gram probabilities dominated the English. In order to overcome these problems, we balanced the probability distribution functions of the two languages by assigning similar probabilities to two n-grams obtained from different corpora if they had a similar frequency rank with respect to the rest of the n-grams obtained from the respective corpus. This third approach showed the best performance, but the multilingual recognition results were still relatively poor compared to the monolingual case [5].

3.1.2. Combining Monolingual LMs at a Meta Level

In our task, language switching takes place only at utterance boundaries. It is, therefore, desirable to suppress the language switching within an utterance. However, the methods for multilingual language models from [5] still allow language switching after every word. Furthermore, those methods require complex manipulations of the counts of the individual language models before merging them. Also, through this process of balancing the probability distributions, the relations of the n-grams within a language are changed. Though the ranking of the n-grams in terms of frequency remains the same, the distance between the n-grams changes. The information that is captured in the shape of the probability distribution over the n-grams of a language is possibly corrupted. We therefore implemented a new approach for the Ibis language model interface that allows to easily combine two monolingual language models and effectively prevents language switching within an utterance. To do so we introduced a meta layer between the monolingual language models and the decoder and call it *multilingual Meta-LM*.

The idea behind our Meta-LM is to prevent language switches by allowing expansions of a given LCT by words taken from the same language as the LCT only. So in this case, an LCT not only represents the n-gram history, but also information about the languages that the words in the history come from. The extension of an LCT by an unwanted word gets the worst possible score. If the language of the LCT is not yet decided, because it only consists of words from all languages, the score is interpolated from all monolingual language models.

Since the n-gram language models for different languages give probabilities that differ in magnitude for reasons described above, our multilingual language model allows the application of different weights to the scores from the individual monolingual language models. Though we applied our Meta-LM so far only to n-gram language models, it is possible to use any type of language model, e.g. interpolated language models or grammars.

3.2. Multilingual Grammars

To handle multilingual grammars we make use of the mechanism of having several different grammars in parallel, as described in section 2.1. When asking for the next scores given a linguistic context the grammar itself ensures, that language switching is not permissible by giving the worst possible score to unwanted words.

The multilingual grammar, together with its language dependent sub-grammars, appears as one linguistic knowledge source – a grammar set – to the decoder. Therefore, the language model vocabulary is built over all grammars in the set. When mapping the different multilingual search vocabularies to the language model indices, one has a choice of two mapping methods:

- Language dependent search vocabulary words can be mapped to the appropriate language dependent language model word. Therefore, all language model words with the same orthographic representation have to be tagged with a language id.
- Language model words with the same orthographic representation and semantic meaning can be shared across the languages, which means, in our case, that e.g. a street name shares the same LVX in all languages.

The second method has a major advantage over the first one, especially in the case that speakers of one language are familiar with the other language. The method allows the decoding of a word with the pronunciation of another language, e.g. a street name, without switching the language model and without destroying the semantic meaning of the word.

3.3. Multi-domain and context language modeling

Though our research in this work focuses on multilingual speech recognition, the techniques to combine monolingual grammars and language models can be easily transfered to multi-domain speech recognition. In this case one decodes different domains using one decoding engine and one common acoustic model but separate domain specific language models. This approach has several benefits. First, it is very easy to implement if the domain-dependent language models already exist. Second, the construction of several domains becomes decoupled from each other and is therefore easier to separately optimize and maintain. Third, by avoiding to combine the training material of different domains, there is no chance for contamination.

In the LingWear project we use this approach to perform speech translation in the medical domain in addition to speech recognition in the navigation domain. For SFB and FAME we use a multi-domain approach based on grammars using the activation/deactivation feature as mentioned in 2.2. We make use of this feature not only on the higher level of complete domains, but allow the dialogue manager to control the speech recognizer depending on single communication contexts by activation/deactivation or weighting of specific rules.

4. EXPERIMENTS AND RESULTS

In order to show the differences in recognition performance and resource requirements of the different approaches for multilingual language modeling, we applied them to the navigation domain of the LingWear mobile tourist assistant. Throughout the experiments, we use one multilingual acoustic model trained in the languages German and English. The performance of the multilingual acoustic in combination with monolingual grammars and n-gram language models serves as a baseline.

4.1. Multilingual Acoustic

For our experiments we used the multilingual acoustic described in [5]. 5000 context dependent tri-phone models were trained on 60 hours of spontaneous German (GSST) and 40 hours of spontaneous English (ESST) Verbmobil-II speech data, as well as 15h of English Broadcast News (BN). Every acoustic unit was modeled by a mixture of 30 Gaussians. The input vector is a combination of 13 Mel-frequency scaled cepstral coefficients (MFCC) at the time of the current frame and seven MFCC vectors to the left and right. A linear discriminant analysis (LDA) reduced the dimension of the input vector to 40.

Incorporated into this system are techniques such as incremental vocal tract length normalization, cepstral mean normalization, and feature space adaptation. The models were trained in a language independent way using a technique called *MLtag* [7]. In MLtag, the training material for phonemes is shared in a knowledge-based way with the help of the alphabet of the International Phonetic Association. Previous research has shown that it is best to preserve the language information of a phoneme so that the algorithm for clustering the context dependent models can ask for the language of a phoneme and decide whether it wants to share the parameters of poly-phones from different languages or not [7].

Language	Speakers	Utterances	Duration
German	11	457	17’
English	9	363	14’
Total	20	820	31’

Table 1. Overview of the test set

	German	English
vocabulary size	2.574	2.029
LM corpus size	9.167	260.914
grammar rules	132	198
grammar nodes	1.357	1.408
grammar arcs	2.061	1.357

Table 2. Size of the vocabularies, grammars and LM training corpora

4.2. Data

German and English test data were collected from native speakers using a laptop and a close-talking microphone. It contains spontaneous speech queries to LingWear while touring the city of Karlsruhe. The queries mainly included requests for directions and informations, whereby informations about streets, hotels, restaurants, sights and other places of interest were stored in the database. The users also asked for recommendations. Table 1 gives an overview.

Table 2 shows the sizes of the vocabularies used for recognizing the two languages and the size of the two language model training corpora, together with the grammar sizes. While the vocabularies and grammars are roughly equal in size, the amount of available language model training material differs substantially. The English training corpus is

approximately 28 times as big as the German one. Not all queries of the test set were covered by the n-gram LMs and grammars.

4.3. Baselines

Using the multilingual acoustic, we decoded the English and German test set separately with the help of monolingual n-gram language models as well as monolingual grammars. The results can be seen in the first two rows of table 3. For all experiments we used the same beam settings during decoding. When optimizing the language model parameters for the single languages, one can see, that there is only a small difference between the languages.

When comparing the monolingual numbers of the English grammar with the English n-gram LM an increase in WER by 14.7%, but a decrease in SER by 7% is seen. In German, there is a gain of 9.4% in WER and also a large gain of 9.4% in the SER. It is also noticed that the WERs and SERs for the grammars in English and German are lying in the same range, which means that both test sets are of equal difficulty. The difference between the results of English and German in the case of the n-gram LM can only be explained by the fact that the English LM training material outweighs the German material by a factor of 28.

When comparing the real time factors between the n-gram LMs and the grammars, it can be seen that the grammar is roughly twice as fast as the n-gram LM. The n-gram LM loses more in the case of German due to the poor coverage of the LM. Thus, decoding along context free grammars in such restricted domains gives a large advantage over the standard n-gram approach.

4.4. Simple Merging of N-gram LM Training Corpora

In analogy to the first experiment in [5] mentioned above, we simply concatenated the two language tagged LM training corpora. We then trained a new n-gram model on this corpus and call it the *multilingual Mix-LM*. The results in table 3 show that the performance on English stays the same. However, on the German test set, the WER increases severely by 17%, the SER increases by 3.3%. As mentioned above, and also observed by [5], the difference between the WERs can only be explained by the large difference of the LM training corpora sizes. This explanation is also validated by the LID, which is in the case of German the worst observed error rate, while on English it is almost as good as the Meta-LM.

The run time behavior of the system is, as expected, slower than the monolingual systems, due to the larger search space. The system is faster than the Meta-LM system, however, because of the missing meta layer, which produces some additional overhead.

4.5. Multilingual Grammar and N-gram LM

4.5.1. Multilingual n-gram Meta-LM

In German, both the WER and the SER increases slightly by 1.8% compared to the monolingual case. Therefore, the Meta-LM outperforms the mixed one. However, in English the WER falls short of the monolingual and mixed LM results. It increases by 14%. However, the SER only increases by 1.7%.

The increase in WER is due to the high amount of language switches, the highest in English among all presented methods. Even though the number of switches roughly equals that of the mixed LM, a language switch in the Meta-LM causes more word errors than for the mixed LM, since the Meta-LM always recognizes a whole utterance in the wrong language in the case of a language switch. One reason for the switches seems to be, that German noises at the beginning of a sentence get a higher LM score than the English ones due to the imbalance of the LM training corpora. Since all noises share the same acoustic independent of their language but are seen as language dependent words in the language model, German noises are preferred over English, therefore pruning away English hypotheses at the beginning of the search.

The Meta-LM shows the slowest runtime behavior of all examined methods, because of the extra meta layer introduced above the monolingual LMs. But the loss in speed is still acceptable when compared to the ease of use of the Meta-LM.

4.5.2. Multilingual Grammar

The results of the experiments of the multilingual grammar show that measured over both languages, the grammar approach outperforms the n-gram LM approaches. Compared to the monolingual grammar baseline, the overall WER increases only by 4% and the SERs are nearly equal. In the case of English the decrease in the SER is a result of the search vocabulary to language model vocabulary mapping. Mainly street names, like ‘Kaiserstrasse’, and other points of interest with a German orthography are replaced by their German pronunciation variant. This was not counted as a language switch, because the semantic meaning of the sentence was not changed. Examples other than street names include ‘okay’, ‘in’ or the noise models. It can also be seen, that the balance in WERs and SERs between both languages is not changed by using the multilingual approach, because our non-statistical grammars do not have to cope with unbalanced training corpora sizes. Due to the larger search space when using the multilingual grammars, the RTFs increase by nearly 28%, but are still twice as small as compared to the multilingual n-gram LM approaches.

As mentioned already in the multilingual n-gram Meta-

		English				German			
Language	Type	WER	SER	LID	RTF	WER	SER	LID	RTF
monolingual	n-gram	21.7%	47.4%	–	0.70	26.5%	54.4%	–	1.29
monolingual	grammar	24.9%	44.1%	–	0.48	24.0%	42.2%	–	0.48
multilingual	n-gram Mix-LM	21.6%	46.8%	5.51%	1.00	31.0%	56.2%	10.28%	1.38
multilingual	n-gram Meta-LM	24.8%	48.2%	6.06%	1.24	27.0%	55.4%	0.22%	1.51
multilingual	grammar	25.9%	43.2%	3.58%	0.67	25.0%	42.7%	1.75%	0.61

Table 3. Performance and resource usage of both systems. SER is the sentence error rate. LID (language identification error rate) is the percentage of sentences with language switches. The real-time factor (RTF) was computed on a Laptop with a 1.13GHz Pentium III Mobile Processor.

LM, the grammar also had to cope with language switches, which leads to completely incorrect sentences, and, therefore, has a high influence on the WER. One third of the failed language identifications in English was produced by one speaker in the test set.

Again, without losing too much in word accuracy, decoding is nearly twice as fast compared to the n-gram LM approaches. Also the SER is much lower which makes it easier for the dialogue manager in LingWear to understand the user queries. Furthermore, when using grammars it is not necessary to use a separate parser for the language processing because the parse tree of the recognizer can be given directly to the dialogue manager.

5. SUMMARY AND FUTURE WORK

In this paper, we investigated techniques for building a multilingual speech interface. We extended the research in multilingual language modeling to the concept of grammars and also introduced a new language modeling method that allows to combine several monolingual into one multilingual LM. Our results proved that multilingual grammars can be used to efficiently decode the two languages English and German within a single system. On our LingWear task, we improved the runtime behavior while facing only a minor loss in performance compared to the monolingual systems. We also showed that n-gram LMs can be combined at a meta level thereby preserving language specific information captured in the individual LMs. The resulting system is easier to maintain, allows decoupled optimization, and implicitly identifies the spoken language during decoding.

In the future we plan to improve the handling of spontaneous speech effects in the multilingual LM and to integrate the mapping between search vocabulary and language model words currently used by the multilingual grammars into the n-gram Meta-LMs. Furthermore we will investigate the combination of n-gram LMs and grammars to benefit from the advantages of both approaches.

6. ACKNOWLEDGMENT

We would like to thank Victoria MacLaren for her help in collecting, transcribing the data, and editing the paper.

This work was partly carried out within the FAME (Facilitating Agent for Multicultural Exchange) projet that is funded by the European Commission as IST-2000-28323, and also within the Sonderforschungsbereich (SFB), No. 588 (Humanoide Roboter - Lernende und kooperierende multimodale Roboter) which is supported by the German research foundation (DFG).

7. REFERENCES

[1] Timothy J. Hazen, I.Lee Hetherington, and Alex Park, "FST-Based Recognition Techniques for Multi-Lingual and Multi-Domain Spontaneous Speech," in *Proceedings of the Eurospeech 2001*, Aalborg, Denmark, September 2001.

[2] Stefan Harbeck, Elmar Nöth, and Heinrich Niemann, "Multilingual Speech Recognition," in *Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, Pilzen, Czech Republic, 1997.

[3] Fulian Weng, H. Bratt, L. Neumeyer, and A. Stolke, "A Study of Multilingual Speech Recognition," in *Proceedings of the Eurospeech*, Rhodes, Greece, September 1997.

[4] Thomas Ward, S. Roukos, C. Neti, M. Epstein, and S. Dharanipragada, "Towards Speech Understanding across Multiple Languages," in *Proceedings of the ICSLP*, Sydney, Australia, November 1998.

[5] Zhirong Wang, Umut Topkara, Tanja Schultz, and Alex Waibel, "Towards Universal Speech Recognition," in *Proceedings of the ICMI*, Pittsburgh, 2002.

[6] Christian Fügen, Martin Westphal, Mike Schneider, Tanja Schultz, and Alex Waibel, "LingWear: A Mobile Tourist Information System," in *Proceedings of the Human Language Technology Meeting (HLT-2000)*, San Diego, USA, March 2000.

[7] Tanja Schultz and Alex Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, August 2001.

[8] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *Proceedings of the ASRU*, Madonna di Campiglio Trento, Italy, December 2001.