# INTRODUCING LINGUISTIC CONSTRAINTS INTO STATISTICAL LANGUAGE MODELING

*Petra Geutner*

pgeutner@ira.uka.de
Interactive Systems Laboratories
University of Karlsruhe (Germany)
Carnegie Mellon University (USA)

## ABSTRACT

Building robust stochastic language models is a major issue in speech recognition systems. Conventional word-based n-gram models do not capture any linguistic constraints inherent in speech. In this paper the notion of function and content words (open/closed word classes) is used to provide linguistic knowledge that can be incorporated into language models. Function words are articles, prepositions, personal pronouns – content words are nouns, verbs, adjectives and adverbs. Based on this class definition resulting in function and content word markers, a new language model is defined. A combination of the word-based model with this new model will be introduced. The combined model shows modest improvements both in perplexity results and recognition performance.

## 1. INTRODUCTION

Conventional stochastic n-gram language models based on word units are being widely used. Due to the lack of training text material, especially in spontaneous speech, robust probability estimates are often not possible. Also, only considering word order of a given training text is not enough to capture linguistic constraints typical for a particular language.

One approach is to make use of syntactic and semantic knowledge that is inherent in the notion of function and content words. Many attempts have been made to incorporate more than local constraints into language modeling [2, 6]. Here, the prediction of the next word is extended not only to the (n-1) last words but to longer-term dependencies. Motivated by the work of [3, 5] a similar approach has been implemented: the next word is predicted through the last function/content word pair, wherever these have been found in the word history. Based on this idea a separate language model is trained. Combining the conventional n-gram model with this special function/content word model decreases perplexity on word basis by 4% and also leads to some improvement in word accuracy.

## 2. FUNCTION AND CONTENT WORDS

The notion of function and content words (sometimes also referred to as open and closed word classes) is well known. Isotani et al. already used this distinction for the Japanese language in order to differentiate between particles and content words [4]. The same way of class distinction is possible for the German language: function words can be thought of as articles, prepositions, personal pronouns; content words are nouns, verbs, adjectives and adverbs – briefly everything that cannot be captured or enumerated within a closed class.

Taking advantage of this idea, the usual local constraints embedded in conventional trigrams can be extended to longer-term dependencies. For the function and content words model this means: beside a normal trigram model the next word is also predicted through the last function/content word pair, resulting in the definition of two separate language models.

1. A conventional trigram model $LM_{word}$, where the prediction of word $w_i$ is based on the previous two words $w_{i-1}$ and $w_{i-2}$:

$$P_{word}(w_i|w_{i-2}, w_{i-1}) := P(w_i|w_0 w_1 ... w_{i-1})$$

2. A function/content word model $LM_{F/C}$, where the history used to predict the next word $w_i$ is based on the last word $w_{i-1}$ and (depending if this word was a content word) the last function word $f$ seen or vice versa:

$$P_{F/C}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} P(w_i|f, w_{i-1}) \\ \quad \text{if } w_{i-1} \text{ is a content word} \\ P(w_i|c, w_{i-1}) \\ \quad \text{if } w_{i-1} \text{ is a function word} \end{cases}$$

The advantage of such dependencies can be seen when looking at an example:

we will ride on the bus.

which is a sequence of

(function function content function function content )

words. Normally "on the" would predict "bus" but "ride the" also is a good predictor for the word "bus". Based on this idea all experiments described in the next section try to take advantage of the linguistic knowledge available.

# 3. EXPERIMENTS

## 3.1. Database

All experiments within this paper are performed on a German database called the **G**erman **S**pontaneous **S**cheduling **T**ask (GSST) which is collected as part of the VERBMOBIL project. In this task human-to-human dialogues are collected at four different sites within Germany. Two individuals are given different calendars with various appointments already scheduled. Only few time slots are available to schedule a meeting between the two of them. Goal of the conversation is to figure out a time that will suit both of them. A total of 616 dialogues were available for training and testing. 608 were used to train the two different language models $LM_{word}$ and $LM_{F/C}$, and 8 represented an independent test set. Table 1 shows a detailed description of all available data. All recognition results reported have been performed with the JANUS system [1].

| | **Training** | **Test** |
|---|---|---|
| #dialogues | 608 | 8 |
| #Utterances | 10735 | 110 |
| #Words | 281160 | 2346 |
| Vocabulary Size | 5442 | 543 |

**Table 1:** GSST Database

## 3.2. Perplexity Results

As baseline a word-based trigram language model was trained and the perplexity of 67.2 was used as reference for all further experiments.

The entire text training corpus has been tagged with function and content word markers. Out of this corpus two language models were built that differ slightly in the treatment of noise words. As spontaneous speech transcriptions include a great many of noise words, the adequate treatment of those words has an important effect on a reliable prediction of the next word.

For training our first function/content word language model an intuitive approach was taken. Transcribed noises that occurred in our training text like #AEH#, #AEHM#, #HM#, #PAUSE#, #LAUGHTER# and so on, were treated as function words as they could be enumerated within a closed class.

| | **Trigram PP** |
|---|---|
| Word Model ($LM_{word}$) | 62.7 |
| F/C Model with Noises ($LM_{F/C}$) | 60.6 |
| F/C Model without Noises ($LM_{F/C}$) | 60.3 |

**Table 2:** Perplexity Results w/o Noises

Treating noises as function words means that they are also used as part of a function/content word pair to predict the next word. The optimal interpolation factor was determined through tests on a cross validation set and best results could be achieved with an interpolation factor of 0.8 yielding a perplexity of 60.6.

A better approach than modeling noises as function words would be to introduce a third marker into the training text. For the second model, beside function and content markers, "noise"-markers were used. These "noise"-markers lead to a slightly different training of the second model: the function/content word language model did not include any noises in the history at all – noise words were not used to predict the next word. Instead the last function/content word pair was always taken to predict the next word, thereby ignoring noises that may have been uttered in between. Introducing the "noise"-markers improved perplexity slightly to 60.3 with a similar effect on word accuracy experiments.

Looking at an example makes this idea clear:

we will ride on the #AEHM# bus.

The word pair "ride #AEHM#" of course is a much worse predictor for the word "bus" than the function/content word pair "ride the" would be.

Whereas the function/content word model proposed by Isotani et al. was able to outperform conventional word bigram models, our newly defined model even improves trigram perplexity. The resulting perplexity of the pure function/content word model alone is, as expected, much higher than perplexity of the conventional word model. But interpolating a conventional trigram word model with the model that uses the last function/content word pair to predict the next word, improves test set perplexity by 4% from 62.7 (pure word model) to 60.3 (combined model) (see also table 3). As to be expected the function/content word model has no use on its own, but is still able to add new linguistic knowledge to the word model, so that perplexity results can

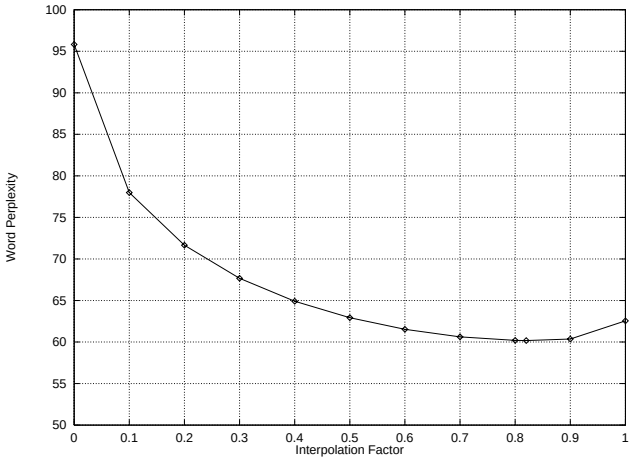


**Figure 1:** Interpolation Results of word-based ($LM_{word}$) and function/content word model ($LM_{F/C}$)

be improved. Figure 1 shows results of all linear combinations between word and function/content word model.

### 3.3. Recognition Performance

Acoustic training of our speech recognizer was done on the same amount of training data as the data available for language model training. A dictionary of 3439 words was used which did not include all words from the training text. The same cross validation set already used for language modeling experiments was used to adjust system parameters. Tests were then made on an independent test set.

Recognition experiments were performed on our JANUS recognizer using a conventional trigram word model to create word lattices. These lattices were rescored using a combination of the language model already applied to the search procedure ($LM_{word}$) and a language model based on function and content word tags ($LM_{F/C}$).

Baseline performance of our system with a word-based trigram language model was 70.6% word accuracy. Both models as described in section 3.2 (with and without noises) were tested. Perplexity improvements achieved on pure text data do not always hold for recognition experiments. Our first model, using noises as part of the word pair that is supposed to contain linguistic information, gave a very small recognition improvement only. For the second model, ignoring noises to predict the next word and using "real" function/content word pairs instead improved baseline recognition performance from 70.6% to 71.0% word accuracy.

## 4. CONCLUSIONS

Using linguistic constraints could be shown to yield into better language models than conventional word trigrams commonly used. Interpolation of a word-based n-gram model with a language model based on the notion of function and content words improves perplexity by 4% and also yields slightly better word accuracy. Better integration of both language models than linear interpolation is an approach to further improve perplexity and recognition performance. Also, another way to further utilize linguistics would be the usage of part-of-speech (POS) markers. The expectation is that integrating a third language model based on POS tags might further improve perplexity and recognition results.

| | **Trigram PP** | **Word Accuracy** |
|---|---|---|
| Word Model | 62.7 | 70.6% |
| Function/Content Word Model | 95.3 | – |
| Combination of Word Model and Function/Content Word Model | 60.3 | 71% |

**Table 3:** Perplexity and Recognition Results

## 5. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Federal Ministry of Education, Science, Research and Technology (BMBF) as a part of the VERBMOBIL project. The views and conclusions contained in this document are those of the author.

## 6. REFERENCES

1. P. Geutner, B. Suhm, F.-D. Buø, L. Mayfield, T. Kemp, A. E. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel. Integrating Different Learning Approaches into a Multilingual Spoken Language Translation System. In Stefan Wermter, Ellen Riloff and Gabriele Scheler, editors, *Connectionist, statistical, and symbolic approaches to learning for natural language processing*, pages 117–131. Springer, Berlin Heidelberg, March 1996. Lecture Notes in Artificial Intelligence.
2. X. Huang, F. Alleva, H.W. Hon, M.-Y. Hwang and R. Rosenfeld. The SPHINX-II Speech Recognition System: an Overview. *Computer, Speech and Language*, 7:137–148, 1993.
3. R. Isotani and S. Matsunaga. Speech Recognition using a stochastic language model integrating local and global constraints. *ARPA SLT Workshop*, pages 87–92, March 1994.
4. R. Isotani and S. Sagayama. Speech Recognition using particle N-grams and content-word N-grams. *Proceed-*

*ings of Eurospeech'93*, pages 1955–1958, September 1993. Berlin, Germany.

5. R. Isotani and S. Matsunaga. A Stochastic Language Model for Speech Recognition Integrating Local and Global Constraints. *Proceedings of the IEEE 1994 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5–8, May 1994. Adelaide, Australia.

6. R. Lau, R. Rosenfeld and S. Roukos. Trigger-based Language Models: A Maximum Entropy Approach. *Proceedings of the IEEE 1993 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 45–48, May 1993. Minneapolis, Minnesota.