USING MORPHOLOGY TOWARDS BETTER LARGE-VOCABULARY SPEECH RECOGNITION SYSTEMS

P. Geutner

Interactive Systems Laboratories
Department of Computer Science,
University of Karlsruhe,
76128 Karlsruhe, Germany

ABSTRACT

To guarantee unrestricted natural language processing, state-of-the-art speech recognition systems require huge dictionaries that increase search space and result in performance degradations. This is especially true for languages where there do exist a large number of inflections and compound words such as German, Spanish, etc. One way to keep up decent recognition results with increasing vocabulary is the use of other base units than simply words. In this paper different decomposition methods originally based on morphological decomposition for the German language will be compared. Not only do they counteract the immense vocabulary growth with an increasing amount of training data, also the rate of out-of-vocabulary words, which worsens recognition performance significantly in German, is decreased. A smaller dictionary also leads to 30% speed improvement during the recognition process. Moreover even if the amount of available training data is quite huge it is often not enough to guarantee robust language model estimations, whereas morphem-based models are capable to do so.

1. INTRODUCTION

Continuous speech recognition systems suffer from various problems. First, when trying to recognize unrestricted speech utterances the acoustic dictionary of a system has to be very large. Huge dictionaries increase the search space, slow down recognition speed and also result in performance degradations. Second, even a huge dictionary will not be able to foresee all new words occurring in the test text. As a consequence there will always appear some words unknown to the recognizer that cannot be recognized properly and might lead to successive errors within the recognition process. Finally, in spite of large databases, there is still insufficient training material. This especially applies to the generation of statistical language models which need immense data to guarantee robust probability estimations. Hence a way has to be found to build robust language models even on the basis of insufficient training material. Decomposing the vocabulary into its morphem-based compounds is a way to solve at least some of these problems.

2. THE DATABASE

All data used for our experiments was taken from dialogues of the German Spontaneous Scheduling Task (GSST) and English Spontaneous Scheduling Task (ESST). This task consists of human-to-human dialogues recorded at different sites with various scenarios. Goal of every conversation is to arrange a meeting of two people within their given schedules. For comparing the vocabulary growth of English and German 146 English training dialogues with 1395 utterances and 14 test dialogues were available. In German a total of 250 dialogues from 4 different sites were available for training and testing. 225 of them were used to train an overall language model, the rest of 25 was used for testing. Recognition experiments were performed with the JANUS-2 system [1] trained on only 200 dialogues. Table 1 shows a detailed description of training and test material.

	Training	Testing
#dialogues	225	25
#utterances	5629	378
#words	117489	7803
vocabulary size	3821	735

Table 1. Used Training and Test Material

3. THE GERMAN LANGUAGE

Comparing various languages like English, Spanish and German, it can be easily seen that the German language differs from all other by an outstanding number of inflections. Consider the word "kommen" ("*to come*" in English). In German for almost every person in singular and plural there exists a different ending:

- ich komm-e[1] (*I come*)
- Du komm-st (*you come*)
- er/sie/es komm-t (*he/she/it comes*)
- wir komm-en (*we come*)
- ihr komm-t (*you come*)
- sie komm-en (*they come*)

So instead of 2 different endings as would be the case in English there are 4 of them in German.

Moreover several prefixes can be attached to every verb, every time creating a new word. Here is an example:

- hinein-gehen[1] (*to go in*)
- aus-gehen (*to go out*)
- weg-gehen (*to go away*)
- .....

[1]Hyphens are used for clarification purposes as decomposition markers only and do not appear in the actual German spelling.

0-7803-2431-5/95 $4.00

Also the German language has an uncountable number of compound words. Nouns can be concatenated to long noun chains, every chain creating a word with a new meaning, e.g. :

- Sprach-erkennungs-modul[2]
  (*speech recognition module*)
- Sprach-erkennungs-genauigkeit
  (*speech recognition accuracy*)

#Utterances	Vocabulary Size (Words)	Coverage	Vocabulary Size (Morphems)	Coverage
500	1301	65%	925	72%
1000	1696	70%	1151	76%
1500	2015	75%	1344	82%
2000	2271	78%	1485	84%
2500	2468	79%	1612	85%
3000	2793	81%	1814	87%
3500	3032	83%	1930	88%
4000	3331	85%	2087	89%
4500	3563	86%	2236	90%
5000	3658	87%	2293	90%
5500	3791	88%	2376	91%
5629	3821	88%	2391	91%

Table 2. Vocabulary Coverage (German)

#Utterances	Vocabulary Size (Words)	Coverage
500	791	87%
1000	1013	91%
1395	1169	92%

Table 3. Vocabulary Coverage (English)

Naturally this fact leads to faster vocabulary growth when the amount of training data increases. As we are dealing with unrestricted spontaneous speech, an increasing number of training dialogues still results in a steady vocabulary growth with no saturation to be expected. Compare figure 1 for the vocabulary growth of the German database with figure 2 which shows the increase of the English vocabulary. In English 1395 utterances resulted in a vocabulary of 1169 words compared to 1971 words (168%!) in German after the same number of utterances.

While the number of words in the dictionary steadily grows, still not all out-of-vocabulary words that might appear in the recognition process can be foreseen. Tables 2 and 3 both show vocabulary coverage of the German test text and the English test text. The smaller English vocabulary already covers 92% of English words in the test dialogues whereas the fourfold amount of training data in German only covers 88%. As a logical consequence it is desirable to work on smaller base recognition units than words to be able to compose new unseen words out of several parts already known to the dictionary.

[2]Hyphens are used for clarification purposes as decomposition markers only and do not appear in the actual German spelling.

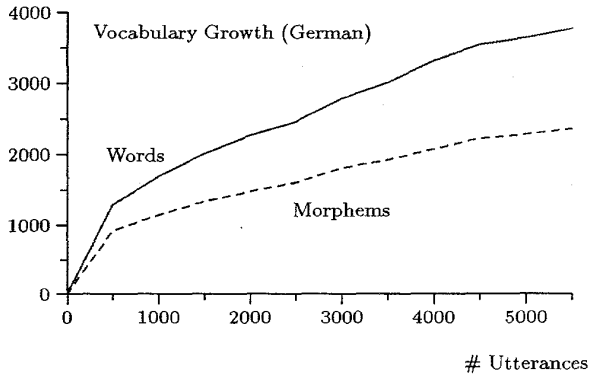

Figure 1. Vocabulary Growth of Words and Morphems in German

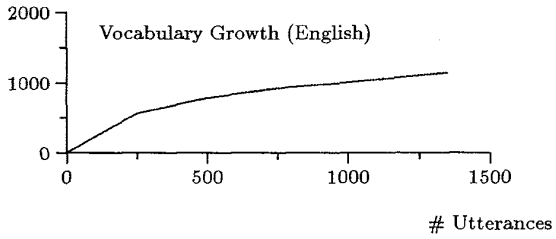

Figure 2. Vocabulary Growth in English

4. MORPHEM-BASED LANGUAGE MODELS

Morphem-based n-gram models allow more robust probability estimates for small training databases and also limit the large vocabulary growth with increasing training material.

Three different ways of decompositions can be performed:

1. strictly morphem-based decomposition, e.g. :
   - weggehen → weg-geh-en[2] (*to go away*)
   - Spracherkennung → Sprach-er-kenn-ung (*speech recognition*)
2. decomposition in root forms:
   - weggehen → weggeh@ (*to go away*)
   - Dialoge → Dialog@ (*dialogues*)
3. combination of strictly morphem-based decomposition and root forms

For the German Spontaneous Scheduling Task (GSST) the decomposition of training texts in strictly linguistically based morphems (**MORPH1**) results in a reduction of vocabulary size by 37% (see figure 1). Whereas the word dictionary contains 3821 words, the corresponding morphem dictionary consists of only 2391 entries (see table 4). This reduction will certainly get bigger the more data will be

	Words	Morphems
#tokens	117489	146990
vocabulary size	3821	2391

Table 4. Comparing Word and Morphem Vocabulary

	Bigram PP	Trigram PP
word (Baseline)	88	67
morphems (MORPH1)	46	33
morphems (MORPH2)	52	39
root forms (ROOT)	79	59
combination (COMB)	59	45

Table 5. Word and Morphem Perplexity

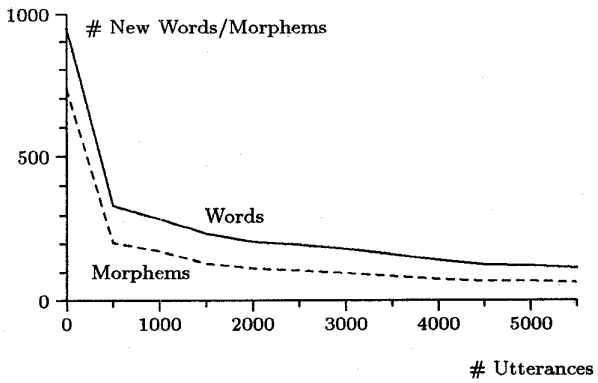

Figure 3. Reduction of New Words in German

available. Moreover the number of new words within the test set decreases much faster when using morphems instead of words (see figure 3). Finally, as it can be seen in table 2, the test set coverage of morphems almost approaches the 92% coverage found in the English language and exceeds German word coverage of 88% by 3% based on the same training data.

Comparing the number of tokens in table 4, we see that on the average one word becomes 1.25 tokens within the morphem-based framework. All available 225 training dialogues were used for building two overall language models: One based on words, the other on their morphem decompositions. Smoothing was done by absolute discounting [2] in both cases.

As to be expected the reduction in vocabulary growth leads to a significant perplexity reduction when comparing morphem-based language models with word models. Taking into account that only every fourth word has been decomposed the perplexity results are surprising: Morphem bigram perplexity is 48% lower than word bigram, for trigrams there is a 51% reduction (see table 5).

4.1. Morphem-based Decomposition

Even though perplexity reduction (and also the restriction of dictionary growth) is highest when using a strictly linguistic-based decomposition of words, recognition results (see table 6) are degrading compared to the word-based recognition process. Whereas the language model profits from a very small unit decomposition, the acoustic part of the JANUS-2 speech recognizer – as expected – suffers from these small components. Hence a more balanced way of decomposition has to be found which both yields a lower perplexity than the word-based bigram model and also guarantees improved recognition performance. As a result a second – not strictly linguistically oriented – decomposition was created (**MORPH2**). The resulting vocabulary reduction is smaller and also the perplexity reduction is not as high as before but recognition results are slightly better than in word-based recognition assuming an open-vocabulary scenario. Perplexity results for the second morphem decomposition are 52 for the bigram and 39 for the trigram case.

4.2. Root Form Decomposition

Another way of reducing the number of different vocabulary words and building a stronger language model is decomposition in root forms (**ROOT**). In this experiment all words of the same root stem but different suffixes are reduced to their root form resulting in a vocabulary of 3205 words instead of the original 3821 words[3]. The vocabulary reduction thus is only 16%. As a consequence perplexity reduction for bi- and trigrams is much lower than in both experiments of the MORPH case: 79 and 59 respectively. Recognition results of this method are given below in section 5.

4.3. Combination

For our last experiment methods MORPH2 and ROOT were combined (**COMB**) yielding a lower perplexity than word models but higher than morphem decomposition only. Perplexity results of all four newly created language models are summarized in table 5.

	Dictionary Size	Word Accuracy	Speed Acceleration
Word Bigram Model (closed-vocabulary)	3085	66.9%	-
Word Bigram Model (open-vocabulary)	3062	64.7%	-
Morphem Bigram Model (open-vocabulary)	2204	65.4%	$\approx$ 30%
Morphem Bigram Model (trigram rescoring)	2204	65.8%	-

Table 6. Recognition Results

5. RECOGNITION RESULTS

Recognition performance was tested on the conventional word-based speech recognizer as well as on the four decomposition methods described above. Acoustic training of the speech recognizer was done on less dialogues than training for language models, resulting in a smaller acoustic dictionary of 3085 words. Recognition results are 64.7% with conventional word bigram models. In this experiment the test set contained 9% new words regarding training vocabulary. The average percentage of unknown words in an

[3]Note that of course this only applies to the language model vocabulary, the acoustic dictionary still has to contain full form words.

unseen test text was determined through cross validation on all available text data. As a desirable baseline, word accuracy was also tested on a closed-vocabulary scenario yielding a performance of 66.9%.

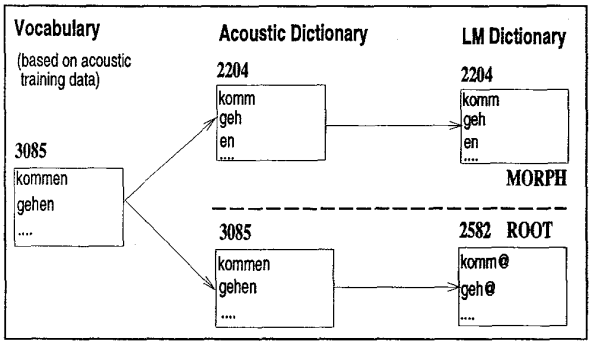

Figure 4. Mapping of Acoustic and Language Modeling Dictionaries during Recognition Process

5.1. Morphem-based Decomposition

Pure morphem-based recognition (as described in MORPH2) measured on word basis slightly outperforms the result achieved with word bigram models by 0.7% (table 6). As the vocabulary size of the acoustic dictionary used within the recognition process is much smaller than on word basis, recognition speed is accelerated by one third.

5.2. Root Form Decomposition

Using root forms only reduces the original language model dictionary from 3821 words to 3205 root forms. This means a 16% reduction in the vocabulary used as basis for language modeling. The relatively small decrease results in a 10% perplexity improvement and thus a slightly stronger bigram language model. However, root forms cannot be used as acoustic dictionary for the recognizer since the suffixes of all inflections also have to be recognized acoustically. During the recognition process these full forms are mapped to their root forms and this information is passed on to the language model module, as it is shown in figure 4. Therefore the acoustic dictionary for the recognizer has to consist of all 3085 full forms. When determining the achieved performance of this experiment, suffixes of recognized words are ignored, thereby measuring root form accuracy instead of word accuracy. As the recognizer output is supposed to be input into a semantic-based parser, good recognition of root forms would be sufficient for the following parsing process [3, 4] leaving good translation accuracy untouched. For comparison the root form accuracy of the open-vocabulary word bigram model case has been taken: 66.2%. The corresponding experiment for root form decomposition results in 63.5%. Obviously a full form word-based language model better supports the recognition process than a root form based one.

5.3. Combination

Regarding the combination of morphem decomposition and root form reduction the same fact applies. The achieved performance outperforms the root form only language model, but with 65.1% accuracy still stays below our assumed baseline of 66.2%.

	Dictionary Size	Root Form Accuracy
Simulation of Root Form Decomposition (Words) (open-vocabulary)	3062	66.2%
Root Form Bigram Model (closed-vocabulary)	3085	63.5%
Combined Bigram Model (open-vocabulary)	2998	65.1%

Table 7. Recognition Results (Root Forms)

In several preliminary experiments trigram rescoring has been applied to the so far best performing morphem-based speech recognition systems. Table 6 shows that even though trigram perplexity is much lower, surprisingly only a small improvement could be achieved, resulting in an overall performance of 65.8%.

6. CONCLUSIONS

As can be seen morphem-based language models yield much better bigram and trigram perplexity results than conventional word n-gram models. Still, with the so far used semi-automatic decomposition methods only little improvement in the recognition performance has been achieved. Even higher-order n-gram models could not improve performance significantly. For future work further evaluations are in progress to find more efficient, less acoustic confusable decompositions automatically.

7. ACKNOWLEDGEMENTS

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

REFERENCES

[1] M. Woszczyna, N. Aoki-Waibel, F. D. Buo, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel. *Janus93: Towards Spontaneous Speech Translation.* Proceedings of the IEEE 1994 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Adelaide, South Australia, pp. 345-348, April 1994.

[2] Hermann Ney and Ute Essen. *Estimating small probabilities by leaving-one-out.* Eurospeech'93, Berlin, Germany, pp. 2239-2242.

[3] Wayne Ward. *Understanding Spontaneous Speech.* Proceedings of the IEEE 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, pp. 365-367, May 1991.

[4] Laura Mayfield, Marsal Gavalda, Wayne Ward and Alex Waibel. *Concept-Based Parsing.* Proceedings of the IEEE 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Detroit, Michigan, May 1995.