ADAPTIVE VOCABULARIES FOR TRANSCRIBING MULTILINGUAL BROADCAST NEWS

P. Geutner

pgeutner@ira.uka.de
Interactive Systems Laboratories
Department of Computer Science,
University of Karlsruhe,
76128 Karlsruhe, Germany

M. Finke and P. Scheytt

finkem@cs.cmu.edu, scheytt@cs.cmu.edu
Interactive Systems Laboratories
Carnegie Mellon University,
Pittsburgh, PA 15213, USA

ABSTRACT

One of the most prevailing problems of large-vocabulary speech recognition systems is the large number of out-of-vocabulary words. This is especially the case for automatically transcribing broadcast news in languages other than English, that have a large number of inflections and compound words. We introduce a set of techniques to decrease the number of out-of-vocabulary words during recognition by using linguistic knowledge about morphology and a two-pass recognition approach, where the first pass only serves to dynamically adapt the recognition dictionary to the speech segment to be recognized. A second recognition run is then carried out on the adapted vocabulary. With the proposed techniques we were able to reduce the OOV-rate by more than 40% thereby also improving recognition results by an absolute 5.8% from 64% word accuracy to 69.8%.

1. INTRODUCTION

Let N be the maximum number of words a speech recognition engine can handle in decoding. For speed and memory reasons this number is limited in current state-of-the-art recognizers to be somewhere in the range of 20k to 60k words. Constraining the maximum number of words can be considered acceptable when building recognizers for languages like English, where the number of out-of-vocabulary words given N=60k vocabulary is below a percent. With the error rates in tasks like transcribing broadcast news or conversational speech (Switchboard) between 30% and 40% (due to highly disfluent speech, noisy environment, and overlapping speech, music etc.) an OOV-rate of less than a percent is not considered a major or significant source of errors[1].

For languages other than English the picture is very different. In order to achieve reasonable automatic transcription performance in the broadcast news domain for languages that are characterized by rapid vocabulary growth due to a large number of possible word inflections, we have to deal with out-of-vocabulary rates between 5% and 13%. This makes OOV-words a major source of recognition errors in the multilingual broadcast news task.

In this paper we focus on building large vocabulary recognizers on two such languages, namely German and Serbo-Croatian.

[1] Rule of thumb: one OOV-word causes about $1.5 - 2$ additional errors.

We present two different techniques for dealing with high OOV-rates by increasing the *effective dictionary size* to the disposal of the recognizer far beyond the *defacto* dictionary size of N.

In a first approach we define morphemes to be the fundamental units to recognize speech. The output of a morpheme based recognizer with a dictionary consisting of N morphemes is then post-processed to assemble words based on a vocabulary list of length $M \gg N$. Thus, the effective size of the dictionary is equal to the number of of words that can be decomposed into N different constituents.

A completely different and as it turned out more effective way increasing the limits of the recognition engine beyond N is a two-pass vocabulary adaptation and recognition approach. The idea is to use for each speech segment the lattice output of a word based recognizer to derive a list of potential words. The list of all words in the lattice is augmented by the most likely words (in terms of number of observations in a huge text corpus) which are acoustically similar to words observed in the lattice. In a second recognition run we make use of this segment-adapted vocabulary and language model respectively. This approach decreased the OOV-rate by 40% yielding an increase in word accuracy of about 5%-6% absolute. The effective dictionary size is estimated to be about three times the defacto size N.

2. DATABASES AND BASELINE SYSTEMS

The following two databases we refer to in this paper were collected at the Interactive Systems Laboratories in Germany and USA.

2.1. Serbo-Croatian

The Serbo-Croatian JanusRTk recognizer was trained on 18 hours of recorded speech of read newspaper articles by native Serbo-Croatian speakers and 15 hours of recorded broadcast news (see table 1). It is based on 35 phones that were modeled by left-to-right HMMs. The preprocessing of the system consists of extracting an MFCC based feature vector every 10 ms. The final feature vector is computed by a truncated LDA transformation of a concatenation of MFCCs and their first and second order derivatives.

Vocal tract length normalization and cepstral mean subtraction are used to extenuate speaker and channel differences.

	Recording Length	# Words	Vocabulary Size
Read Data	18 hours	89.000	17.000
Spontaneous Data	15.5 hours	125.000	25.000
Text Data	–	11.800.000	313.000

Table 1: **Available Training Data** (Serbo-Croatian).

The language models were trained on the hand-transcribed acoustic training data and an additional 11.8 million words of text data collected on the internet [2]. Baseline results given a comparatively high OOV-rate of 13.6% are shown in table 2 below.

	Vocabulary Size	OOV-Rate	Word Accuracy
Baseline	31.000	13.6%	62.4%
Interp.LM	31.000	13.6%	64.0%

Table 2: **Baseline Recognition Results** on Serbo-Croatian broadcast news comparing a baseline system and its interpolated language model rescored output.

2.2. German

As recognition engine for German a recognizer trained on spontaneous human-to-human dialogues was used and retrained with 2 hours of recorded broadcast news. The baseline result [2] was created by building a language model only on broadcast news texts (see table 4). Two other language models were used for interpolation: one was built on a large German newspaper corpus, the other on training texts from German radio news.

	Recording Length	# Words	Vocabulary Size
Acoustic Data	2 hours	21.000	6.000
Text Data	–	46.00.000	980.000

Table 3: **Available Training Data** (German).

3. VOCABULARY GROWTH

The following figures present a statement of the problem we are attacking in this paper. Figure 1 shows the number of words as a function of the number of tokens in broadcast news data for both languages considered.

In figure 2 we compare the self-coverage and cross-coverage as measured on newspaper and broadcast news text corpora.

[2] As test data a "clean" segment of speech was used, which is one of the reasons for better performance of the German system, although less training data had been available.

	Vocabulary Size	OOV-Rate	Word Accuracy
Baseline	17.000	9.3%	64.6%
Interp.LM	17.000	9.3%	74.2%

Table 4: **Baseline Recognition Results (German)** comparing a baseline system and its interpolated language model rescored output.

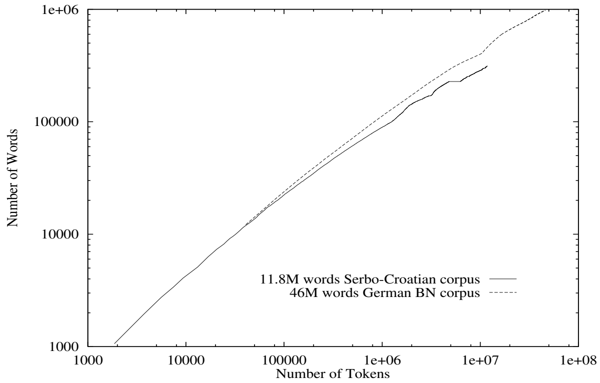

Figure 1: **Words per Token** shows the rapid vocabulary growth in German as well as Serbo-Croatian.

4. MORPHEME BASED RECOGNITION

Attempts have been made to reduce the high number of OOV-words by using morphemes or morpheme like base units throughout the recognition process [1]. Especially for Serbo-Croatian words do follow very systematic and easy-to-recognize patterns concerning inflections (see table 5). Thus, making use of a decomposition of words in word stem and suffix is straight forward.

Taking advantage of morphological knowledge like the one presented above, the training database of the recognizer was automatically decomposed into morphemes. For the baseline experiment the number of dictionary words N in the vocabulary thereby

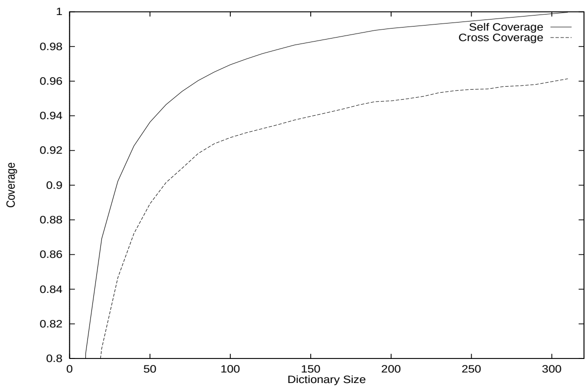

Figure 2: **Self- and Cross-Coverage** on Serbo-Croatian broadcast news data.

Word	Decomposition	Translation
hotela	hotel -a	(of the) hotel
hotelu	hotel -u	(to the) hotel
hotelom	hotel -om	(with the) hotel
govoriti	govor -iti	to speak
govorim	govor -im	I speak
govoris5	govor -is5	you speak (singular)
govori	govor -i	he speaks
govorimo	govor -imo	we speak
govorite	govor -ite	you speak (plural)
govore	govor -e	they speak

Table 5: **Examples for Serbo-Croatian Morphology.**

Word	Decomposition	Translation
Wahrheit	Wahr -heit	truth
Schwierigkeit	Schwierig -keit	difficulty
Kinder	Kind -er	children
Kindern	Kind -ern	children
gehen	geh -en	to go
(ich) gehe	geh -e	(I) go
(Du) gehst	geh -st	(you) go
(er) geht	geh -t	(he) goes

Table 6: **Examples for German Morphology.**

decreased from 31k words to 17k morphemes. The *effective dictionary size* of the 17k morpheme based experiment was computed to be equivalent to a word based OOV-rate of 7.5%. Trigram Language models were built on the morphologically decomposed text corpora. The acoustic models were not retrained for our experiments.

The recognition results of the morpheme based recognizer are shown in table 7. The output of the recognizer was postprocessed by merging morphemes to words based on a list of all words observed in the text corpora. It turns out that the morpheme based approach is significantly worse than the word based. Since even getting the morphemes right the word based recognizer does a better job we are now retraining the recognizer based on the morpheme decomposition to make the acoustic models fit better to the new units of recognition.

	Vocab. Size	OOV-Rate	Morph. Acc.	Word Acc.
Baseline (word-based)	31.000	13.6%	65.1%	55.1%
morpheme-based (17k)	17.000	7.5%	60.5%	46.7%
morpheme-based (31k)	31.000	5.5%	63.3%	48.6%

Table 7: **Morpheme-Based Recognition Results** (Serbo-Croatian).

In a second experiment not only the text of the acoustic training material with a vocabulary of 31k was decomposed, but also another 50k words from Web texts were split and combined with the 17k vocabulary resulting in a 31k morpheme dictionary. This resulted in a word based OOV-rate of 5.5% instead of 7.5% and improved the recognition result, but still was worse than the baseline experiment.

Similar experiments were performed on the German database resulting in the following numbers (table 8).

	Vocabulary Size	OOV-Rate	Morph. Acc.	Word Acc.
Baseline (word-based)	17.000	9.3%	72.5%	64.6%
morpheme-based	17.000	6.5%	62.8%	54.4%

Table 8: **Morpheme-Based Recognition Results** (German).

5. VOCABULARY ADAPTATION

A cheating experiment has been performed, pretending all information about the news vocabulary of a certain day would be accessible. Even when all important keywords of the day of transmission of the news broadcast would be known, the OOV-rate would only decrease to 7.8% in Serbo-Croatian news. The same cheating experiment as described above, was done on German data and resulted in an OOV-rate of 5.5%. This means that a significant portion of the OOV-words are not necessarily day or event related (new events cause new words to show up).

Therefore, the following vocabulary adaptation approach makes use of acoustic similarity instead of semantic similarity to reduce the OOV-rate. A first recognition run on a general baseline dictionary is followed by a second recognition run with a dynamically adapted dictionary of the same size but a smaller OOV-rate. Especially in a time uncritical process like the recognition of broadcast news this seems to be a practical idea.

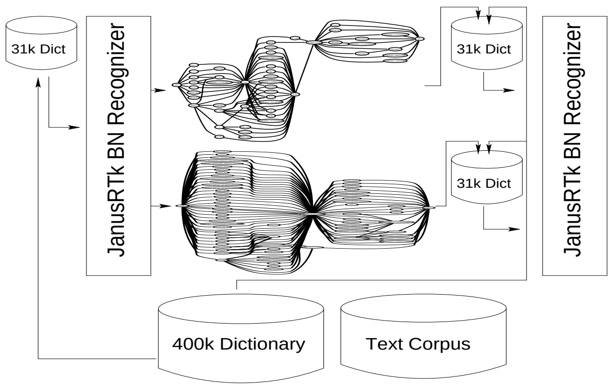

Figure 3: **Vocabulary Adaptation** based on Lattices. Two pass recognition and vocabulary adaptation.

In a first recognition run, word lattices for all test utterances are created. The lattice is then used to determine, which words are most likely uttered in the segment (namely all words represented in the lattice). For each utterance to be recognized this lattice leads to an utterance specific vocabulary. This vocabulary is then used to dynamically adapt the recognition dictionary. The basic idea is, that a large number of words in the recognized hypothesis are recognized incorrectly because only the inflection ending is wrong whereas the stem was recognized correctly. In many cases this was not due to misrecognition but because the right word was not even in the dictionary of the recognizer, so constituting an OOV-word. The algorithm below shows the whole **Vocabulary Adaptation** process:

1. A first recognition run gives word lattices and an utterance-specific vocabulary list.
2. This vocabulary list is then split into word stems and suffixes (where different combinations of word stem and suffix lengths were tested, see table 9). Note that the word stem length had at least to be 2 letters long.
3. The resulting word stem list is then used to look up all similar words in the full dictionary consisting of all words that were observed in the language model training text.
4. All words with the same stem are then incorporated into the dictionary by being replaced with the least frequent words that did not show up in the lattice. (so that the dictionary size of the recognizer remains N)
5. In an automatic procedure a new dictionary and language model is created to perform a second recognition run.

This vocabulary adaptation procedure applied to Serbo-Croatian broadcast news data yields a significant improvement in terms of the OOV-rate, which is reduced by 40% (see table 9), and in terms of the accuracy by reducing the error rate by 5.8% (see table 10).

	Wordstem Length				
Suffix Length	2	3	4	5	6
1	9.7%	9.0%	8.7%	8.4%	9.0%
1+2		8.9%	8.2%	8.2%	8.6%
1+2+3			8.1%	8.0%	8.4%
1+2+3+4			8.2%	7.9%	8.3%

Table 9: **Serbo-Croatian OOV-rates** with different Splitting Methods. The baseline OOV-rate is 13.6%.

	Vocabulary Size	OOV-Rate	Word Accuracy
Baseline	31.000	13.6%	64.0%
Adapted	31.000	7.9%	69.8%

Table 10: **Serbo-Croatian Recognition** Results Based on Adapted Vocabulary.

Table 11 shows that the same result holds for German news data, again a significant reduction of the OOV-rate. For German a fixed list of suffixes was used to create the word stems. Some examples for the used suffixes are given in table 6. Using this linguistic knowledge for decomposition also resulted in a huge OOV-rate reduction from 9.3% to 6.0% (see table 11).

	Wordstem Length				
Suffix Length	2	3	4	5	6
fixed	–	–	7.7%	6.0%	6.5%

Table 11: **OOV-rates with different Splitting Methods (German)**: The baseline OOV-rate is 9.3%.

In both languages it turned out to be a good choice to fix the stem length to 5 which is correlated with the distribution of word lengths (50% of the words are longer than 5 letters). Figure 4 shows the distribution of different word lengths in Serbo-Croatian and German.

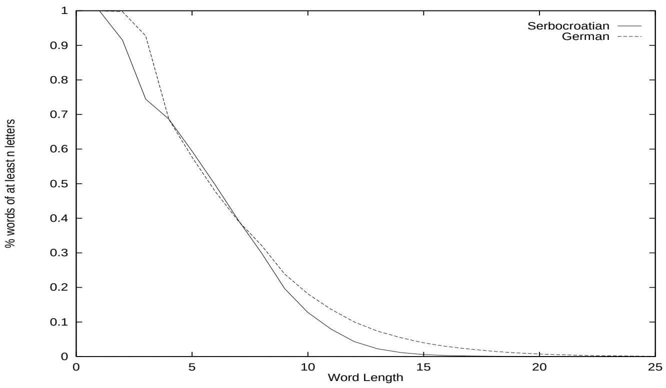

Figure 4: **Cummulative Distribution of Word Length** for German and Serbo-Croatian language.

6. CONCLUSIONS

On two different broadcast news corpora – Serbo-Croatian and German – several methods to decrease OOV-rates have been applied. The presented two-pass recognition approach reduces the number of out-of-vocabulary-words by 35% in German and 42% in Serbo-Croatian. Recognition results with the Serbo-Croatian speech recognizer also show tremendous performance improvements of an absolute 5.8%.

7. ACKNOWLEDGEMENTS

This research was partly funded by the Advanced Research Projects Agency under contract No. N66001-97-D-8502. The views and conclusions contained in this document are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

8. REFERENCES

[1] P. Geutner. Using Morphology towards better Large-Vocabulary Speech Recognition Systems. *Proceedings of the ICASSP 95*, pages 445–448, May 1995. Detroit, Michigan.

[2] P. Scheytt, P. Geutner, and A. Waibel. Serbocroatian LVCSR on the Dictation and Broadcast News Domain. submitted for publication.