REDUCING THE OOV RATE IN BROADCAST NEWS SPEECH RECOGNITION

Thomas Kemp *Alex Waibel*
Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

ABSTRACT

The recognition of broadcast news is a challenging problem in speech recognition. To achieve the long-term goal of robust, real-time news transcription, several problems have to be overcome, e.g. the variety of acoustic conditions and the unlimited vocabulary. In this paper we address the problem of unlimited vocabulary. We show, that this problem is more serious for German than it is for English. Using a speech recognition system with a large vocabulary, we dynamically adapt the active vocabulary to the topic of the current news segment. This is done by using information retrieval (IR) techniques on a large collection of texts automatically gathered from the internet. The same technique is also used to adapt the language model of the recognition system. The process of vocabulary adaptation and language model retraining is completely unsupervised. We show, that dynamic vocabulary adaptation can significantly reduce the out-of-vocabulary (OOV) rate and improve the word error rate of our broadcast news transcription system View4You.

1. THE VIEW4YOU SYSTEM

The View4You project is a cooperation between the Interactive Systems Labs and the Carnegie Mellon University's Informedia group [5]. It aims at the automatic generation of a searchable multilingual video database. In the prototype system, German and Serbocroatian TV news shows are recorded daily and stored as MPEG compressed files. Using the acoustic signal, a segmenter chops the newscasts into acoustically homogeneous segments ranging from several seconds to few minutes in length. A speech recognition system generates transcriptions for the segments. The segmentation information and the automatic transcriptions are stored in a database.

The user of the system can give queries in natural language, e.g. 'Tell me everything about the peace talks between Mr Netanyahu and Mr Arafat'. Using the speech recognizer's transcriptions in the multimedia database, an information retrieval component computes a ranked order of relevant segments, which are displayed to the user. By clicking on a segment, an MPEG-player is activated that plays the corresponding video segment.

For more details on the View4You system, see [1].

2. MOTIVATION

The index into View4You's video database consists of the output of our speech recognizer. Therefore, only words that are in the vocabulary of the recognizer can be searched for. If a video contains keywords that are unknown to the recognizer, they cannot be found in the index, and the user can not retrieve the video by this keyword. OOV (out-of-vocabulary) words therefore pose a problem to the View4You system, and the vocabulary of the speech recognizer should be as large as possible to ensure low OOV rates. Currently, our speech recognizer is limited to a vocabulary of 64k words. On the North American Business News (NAB) corpus, a vocabulary of this size covers more than 99% of the text, and even with a 20k vocabulary, OOV rates on the NAB task do not exceed 3%.

We measured the OOV rate on German news shows for a 60k vocabulary which was derived from our language model corpus. The result is shown in figure 1.

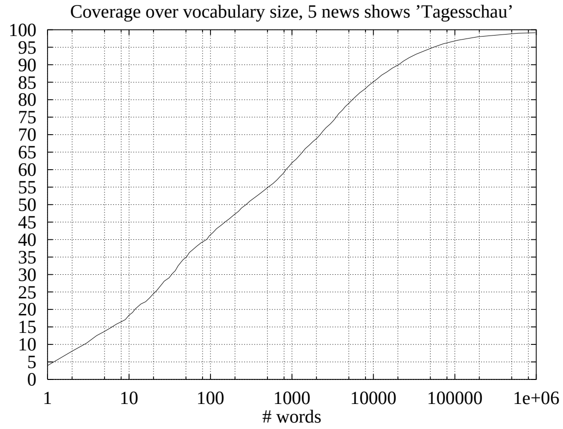

Figure 1. Coverage over vocabulary size for the test set

As can be seen from the graph, the average OOV rates in German broadcast news are approximately 3 to 4 times higher than in the English NAB task. There are two reasons for this. First, as news topics have a high variety, the vocabulary of broadcast news shows generally exceeds the vocabulary of business news text, regardless of the language used [5]. Second, the German language allows the free construction of compound nouns, like e.g. 'kindergarten' from 'kinder' (children) and 'garten' (garden), or 'dachshund' from 'dachs' (badger) and 'hund' (dog). Often, complex morphology rules make it difficult to treat compounds as separate words. Such compound words form an open set and can not be covered by a static vocabulary of any reasonable size.

An OOV rate of approx. 5% could be regarded as a minor problem. However, for the View4You system, only words that might be used in database queries are relevant. To determine the rate of unknown words within the relevant words of the testset, we manually categorized the unknown words in our testset into relevant (all nouns and infrequent verbs and verbforms) and irrelevant words (frequent verbs, conjunctions, etc). The result of this analysis is shown in table 1.

category	total count	OOV count
all words	10648	486 (4.5%)
relevant words	3462	318 (9.2%)
unique relevant words	1935	291 (15%)
proper names	-	69

Table 1. Composition of the unknown words

2/3 of the unknown words are categorized relevant, but only about 1/3 of the known words. The out-of-vocabulary rate in terms of relevant words is therefore 9.2%. If multiple occurrences of a relevant term are counted only once, more than 15% of all relevant keywords in the testset are unknown. Note that only 20% of all relevant unknown words are proper names (including street names, city names and country names), which is less than we expected.

In this paper, we describe a method that can be used to reduce the OOV rate on German TV news shows based on information retrieval (IR) techniques. The basic idea is as follows.

Many news sources are available electronically, e.g. as web newspapers, or captions of TV or radio news shows. This sources can be automatically loaded from different internet servers and can be used to construct a searchable news database.
The preliminary output of a speech recognizer on the testset, using a generic vocabulary, can then be used as a query for the news database. From the retrieved relevant articles of this database, a new vocabulary can be computed, which may have a lower OOV rate.
The rationale behind this is, that different kinds of events have different words associated with them, which are used to describe either the events themselves, consequences, or actions to be taken. If, for example, the word 'raft' is unknown to the recognizer in the description of a flood in India, this word may very well appear in a news article over a flood in Bangladesh in the web news database.

3. DATABASES

3.1. The View4You broadcast news database

For all described experiments we used the View4You database, which has been collected at the University of Karlsruhe since October 1996. A news program (called 'Tagesschau') is recorded daily and stored as MPEG-1 compressed file. The audio part is manually segmented and transcribed. The segmentation is done according to the acoustic condition of the audio signal. Therefore, each segment contains either field speech or clean speech from the anchor speaker. For our experiments, we used a set of 12 transcribed news shows totaling 3 hours of speech. 8 shows (approx. 2 hours of speech) were used for training.
There are captions available for the anchor part of the news shows. This captions are moderately accurate and cover about 45% of the speech signal. There are no captions available for the rest of the signal.

3.2. The ONLINE corpus

The ONLINE corpus consists of news text which is collected daily from the internet. ONLINE consists of three sub-corpora: GLIVE, BR5 and TGS/TT. GLIVE is a internet newspaper featuring reports and analysis of daily news. BR5 contains the transcription of radio news. The TGS/TT corpus contains the transcriptions of the anchor speaker part of the 'Tagesschau' and 'Tagesthemen' TV news shows. No captions for the rest of the data are available.
The size of the ONLINE corpus and its sub-corpora is summarized in table 2.

database	time covered	number of articles	size (MB)
BR5	20/06/96 to 01/10/97	8488	30
GLIVE	01/11/96 to 01/10/97	56883	84
TGS/TT	26/11/96 to 01/10/97	5683	5
ONLINE	20/06/96 to 01/10/97	71054	119
ONLINE-0	20/06/96 to 28/02/97	25121	44
ONLINE-1	01/03/97 to 30/04/97	11733	18
ONLINE-2	01/05/97 to 31/07/97	21292	31

Table 2. The ONLINE database

3.3. Vocabulary and language model

The View4You language model is a standard backoff trigram language model which has been built upon the concatenation of two corpora. The structure of the language model training corpus is summarized in table 3.

database	time covered	size (kWords)
ONLINE-0	20/06/96 to 28/02/97	6052
FAZ	1992-1994	39669
total	1992 to 28/02/1997	45721

Table 3. Corpora used for language modelling

The FAZ corpus contains text from a German newspaper ('Frankfurter Allgemeine Zeitung'). All data more recent than March 1, 1997, has been excluded both from the language model training and from the training of the acoustic models. The test data is taken from the period between March 30, 1997 and June 30, 1997.

4. THE INFORMATION RETRIEVAL (IR) ENGINE

We built our information retrieval engine using the Okapi similarity measure [4]. This measure has been evaluated thorougly in the context of NIST's TREC information retrieval contests [3], and has been found to be especially powerful. The Okapi measure can be parameterized to the special requirements of a task. We use a parameterization that has been found to be very good for short queries [6]:

$$d(q,d) = \sum_{t \in Q \wedge t \in d} \left(\frac{f_{d,t}}{f_{d,t} + \sqrt{\frac{f_d}{E(f_d)}}} \right) \log \left(\frac{N - f_t}{f_t} \right)$$
$$= \mathrm{Okapi}(k_1 = 1, k_2 = 0, k_3 = 0, b = 1, r = 0, R = 0)$$

where N is the number of documents in the collection, f_t is the number of documents containing term t, $f_{d,t}$ is the frequency of term t in document d, and f_d is the number of terms in document d, which is an approximation to the document length. A *term* in this context is the same as a word, however, the 500 most frequent words ('I', 'other' and the like) are excluded. The database engine computes the distance between a query and each article in the database and returns the articles sorted in decreasing order of similarity to the query. For longer queries of about 50 words, typically several thousand articles are found.

5. EXPERIMENTS

5.1. The View4You speech recognizer

The speech recognizer of the View4You system is based on the JANUS-3 speech recognition toolkit. It is a standard HMM based, continuous mixture decision-tree clustered triphone base model system. For the experiments described in this paper, we used speed-optimized system that had a word error rate of 25% on the anchor speaker portion of the data.

For a detailed description, refer to [1] [2].

5.2. Experiments on recorded news shows

We ran our View4You recognizer with the generic 60k dictionary on 4 news shows, assuming perfect (manual) segmentation. For the sake of speed, we used a stripped-down recognizer setup that performs about 5 times faster than our baseline system but has a higher word error rate.
We excluded the anchor speaker portions of each news show from the evaluation. The remainder is the most challenging part of the data: it contains field speech, often over telephone lines, with all kind of background noise, and often several speakers at once. The results in terms of word error rate and the OOV rates are shown in table 4. Only approximately every other word in the hypothesis is correctly recognized.

show (date)	error rate	OOV (words)
30/03	48.9%	56 (4.2%)
13/04	49.2%	65 (5.3%)
28/05	46.1%	70 (5.9%)
27/06	48.5%	60 (4.3%)

Table 4. Recognition results (worst segments) and OOV rates

We built an information retrieval system containing the complete ONLINE corpus (71054 articles). The recognition result of each of the segments, which were on the average 30 seconds in length, was used as a query to the IR system. The number of retrieved articles ranged between three and 10,000. In a first experiment, we discarded all retrieved articles which were more recent than the training corpus, that is more recent than 28/02/97. All words in the found articles were added to the vocabulary, until either no more articles were available, or 5 percent of the vocabulary had been exchanged in this way. Note that by adding a word the size of the vocabulary was kept constant by eliminating the token with the fewest counts in the baseline language model corpus. Therefore, by adding too much or inappropriate new data, the OOV rate can rise again (see figure 2).

The resulting OOV rates when adding data in the described way are shown in table 5.

date	OOV after lookup	change in OOV rate
30/03	46	-18%
13/04	52	-20%
28/05	68	-3%
27/06	60	0%

Table 5. OOV changes by IR lookup into the ONLINE-1 corpus

Adding words from all different segments into one global vocabulary is clearly suboptimal. Therefore, in a contrast experiment, we computed distinct vocabularies for each of the segments for the 30/03 news show. The overall number of OOV words dropped from 52 to 50, which is only a slight improvement. Since the effort to compute distinct vocabularies for each of the segments is very high, we decided to use only one global vocabulary adaptation per complete news show.

The results from table 5 indicate, that the benefit from adding data from the external news corpus is the higher, the more recent this data is with respect to the show.

Therefore, in order to improve the results achieved on the two most recent shows, we added the sub-corpora ONLINE-1 and ONLINE-2 to the database. We conducted the same experiment as described above both with this large database and with the ONLINE-2 portion of it alone, which is closest to the date of the test shows. The results are shown in table 6.

date	ONLINE-0	ONLINE-2	ONLINE-0,1,2
28/05	-3%	-31%	-34%
27/06	0%	-30%	-30%

Table 6. OOV rate changes using different databases

The addition of the older data does not yield substantial improvements. Therefore, we used only the ONLINE-2 corpus for the 28/05 and 27/06 shows and the ONLINE-1 subcorpus for the remaining two test broadcasts. The final results are summarized in the third column of table 7. Unsupervised information-retrieval based vocabulary adaptation is capable to reduce the OOV rate by 33.5%, or 1.7% absolute.

5.3. Linear text interpolation

To compare the IR-based method of OOV reduction with a simpler scheme, we tried to interpolate the existing base training corpus with the new data (the ONLINE-1 and the ONLINE-2 corpus, respectively). Since the added corpora are much smaller than the base training corpus, their impact on the vocabulary is limited. Therefore, we weighted the words in the added corpus with a interpolation weight λ. For high values of λ, the resulting vocabulary will be the vocabulary of the added sub-corpus. For very low values of λ, the existing vocabulary will remain unchanged. We tried this method with a variety of values for λ. The results are shown in figure 2.

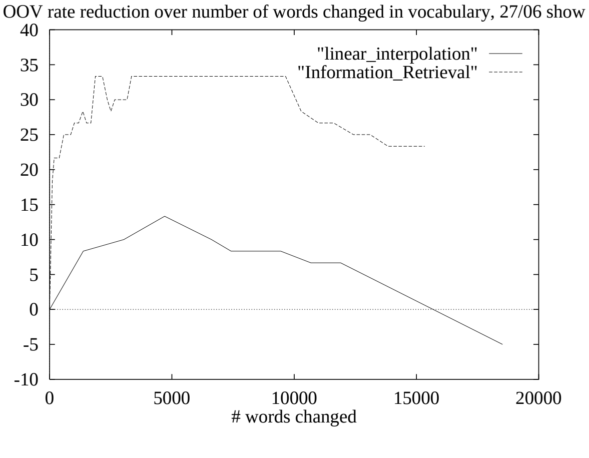

Figure 2. Coverage over number of changed words for linear interpolation and IR

Using this data as crossvalidation, we conclude that the optimal size of the replaced vocabulary is somewhere around 10% of the words, or 6k. With this value, we computed new vocabularies with both the information retrieval algorithm and with linear interpolation. The results are summarized in table 7. The column 'best linear' refers to a cheating experiment, where the interpolation factor λ was chosen for each news show as to minimize the OOV rate. This result can therefore be regarded as an upper bound for the linear interpolation.

The last column shows the upper bound for the OOV rate reduction on the given corpora. Only about one half of the OOV words in the test set is contained in the new data. The IR algorithm is capable to detect 59% of them.

date	linear	IR	best linear	best possible
30/03	-34%	-50%	-39%	-64%
13/04	-20%	-24%	-20%	-40%
28/05	-14%	-31%	-16%	-55%
27/06	-10%	-33%	-13%	-53%
average	-19.1%	-33.5%	-21.5%	-53%

Table 7. OOV changes by IR lookup and linear interpolation

In a final analysis, we examined the type of the unknown words which were found by the IR algorithm. This was done by manually categorizing each found unknown word. More than 90% of the found new words were categorized relevant.

5.4. Recognition results

To evaluate the effect of IR based vocabulary adaptation, we ran our baseline recognizer on one news show both with the baseline vocabulary and with the adapted vocabulary. The pronounciations for the new words were taken from a large background pronounciation dictionary.
In an additional experiment, we added all articles retrieved by IR to the base language modeling text corpus and recomputed the language model. No weighting of the new data took place. The results of the two experiments are summarized in table 8. Using both vocabulary and language model adaptation, the word error rate on the most difficult part of the news shows dropped by 11.2% relative.

System	word error rate	improvement
baseline	41.0%	-
vocabulary adapted	38.3%	6.4%
plus LM adapted	36.4%	11.2%

Table 8. Recognition rates (baseline system)

6. CONCLUSION

In this work we have shown, that it is possible to reduce the OOV rate on German news shows with a 60k vocabulary by approximately one third, using unsupervised corpus collection from the internet and information retrieval (IR) techniques to select relevant articles from the collection. This reduction is possible in spite of a rather low recognition accuracy of the recognition system on the tested sub-part of the broadcast news. Comparing the result achieved with IR techniques with linear interpolation of the vocabulary with the vocabulary of the internet corpus collection, we find that the IR techniques reduce the OOV rate significantly better.

Using the adapted vocabulary, the word error rate dropped by 6.4% relative. If the retrieved articles are added to the language model, the word error rate improves by another 4.9% relative.

Analysis on the unknown words that were found by the IR-based algorithm showed, that more than 90% of these words are relevant terms with regard to information retrieval applications. This makes the algorithm especially useful for such applications.

7. ACKNOWLEDGEMENTS

This work was carried out at the Interactive Systems Labs, Karlsruhe. The authors would like to thank all members of the Interactive Systems Labs for helpful discussions and support. We also thank Manfred Weber for providing the manual transcriptions for the news shows.
The views and conclusions contained in this document are those of the author.

REFERENCES

[1] T. Kemp, P. Geutner, M. Schmidt, B. Tomaz, M. Weber, M. Westphal, A. Waibel, *The interactive systems labs View4You video indexing system*, elsewhere in these proceedings

[2] T. Kemp, A. Waibel, *Unsupervised training of a speech recognizer using TV broadcasts*, elsewhere in these proceedings

[3] http://www-nlpir.nist.gov/TREC/

[4] M.M. Beaulieu, M. Gatford, X. Huang, S.E. Robertson, S. Walker, P. Williams, *Okapi at TREC-5*, Proc. of the 5th Text Retrieval Conference, NIST, Gaithersburg, MD, January 1997

[5] H. Wactlar, A. Hauptmann, M. Witbrock: *Informedia: news-on-demand experiments in speech recognition*, Proc. of ARPA SLT workshop, 1996.

[6] R. Wilkinson, J. Zobel, R. Sacks-Davis: *Similarity measures for short queries*, in Proc. of TREC-4, NIST, November 1995