PHONETIC-DISTANCE-BASED HYPOTHESIS DRIVEN LEXICAL ADAPTATION FOR TRANSCRIBING MULTLINGUAL BROADCAST NEWS

Petra Geutner, Michael Finke and Alex Waibel

pgeutner@ira.uka.de, finkem@cs.cmu.edu, ahw@cs.cmu.edu
Interactive Systems Laboratories
University of Karlsruhe (Germany)
Carnegie Mellon University (USA)

ABSTRACT

High out-of-vocabulary (OOV) rates are one of the most prevailing problems for languages with a rapid vocabulary growth due to a large number of inflections. Especially when transcribing Serbo-Croatian and German broadcast news, the OOV-rate is between 8.7% and 4.5%. Hypothesis Driven Lexical Adaptation (HDLA) has already been shown to decrease high OOV-rates significantly by using morphology-based linguistic knowledge. This paper introduces another approach to dynamically adapt a recognition lexicon to the utterance to be recognized. Instead of morphological knowledge about word stems and inflection endings, distance measures based on Levenstein distance are used. Results based on phoneme and grapheme distances will be presented. Compared to the use of morphological knowledge, our distance-based approach offers the distinct advantage that no expert knowledge about a specific language is required, no definition of complex grammar rules is necessary. Instead, grapheme sequences or the phoneme representation of words are sufficient to apply our HDLA algorithm easily to any new language. With our proposed technique we were able to decrease OOV-rates by more than half from 8.7% to 4%, thereby also improving recognition performance by an absolute 4.1% from 29.5% to 25.4% word error rate.

1. INTRODUCTION

One of the most prevailing problems in transcribing broadcast news shows is the out-of-vocabulary problem. High out-of-vocabulary rates automatically lead to a decrease in recognition performance, thus a method that is able to reduce the number of out-of-vocabulary words will also improve recognition results. Our Hypothesis Driven Lexical Adaptation (HDLA) algorithm is able to reduce out-of-vocabulary rates significantly by adapting the dictionary of a speech recognition system to the speech data to be recognized. This paper presents several methods to deal with problems arising from rapid vocabulary growth due to a large number of inflections and the resulting high out-of-vocabulary rates. As representatives for languages with a large number and wide variety of inflection endings, Serbo-Croatian and German were chosen to illustrate the problems. Both languages were tested on a domain that requires an almost unlimited vocabulary: automatic transcription of broadcast news.

Major goal of all methods presented here is to allow speech recognition on a virtually unlimited vocabulary by adapting the used dictionary to the utterance to be recognized. Especially when transcribing broadcast news, this should keep the out-of-vocabulary rate limited and thus improve the word error rate. All methods are based on a two-pass recognition approach. A word list generated in the first run is used as basis for the adaptation of the recognition dictionary for the second recognition run. Adaptation of the lexicon is performed through various approaches that either use lingustic knowledge about morphology or distance-based measures on grapheme or phoneme level. Using any of the proposed methods OOV-rates drop by 30% to 55%, thereby also decreasing word error rate by 4.1% absolute. The effective dictionary size for this experiments is estimated to be more than three times the defacto size N.

2. THE SPEECH RECOGNITION ENGINE

The speech recognition system used to perform all experiments for transcribing Serbo-Croatian broadcast news shows [3] was trained on 12 hours of recorded speech of read newspaper articles and 18 hours of recorded broadcast news. It is based on 35 phones that are modeled by left-to-right HMMs. The preprocessing of the system consists of extracting MFCC based feature vectors every 10ms. 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.

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. Some results, including the baseline system (B5) for our experiments with an OOV-rate of 8.7%, are shown in table 1 below.

System	Vocabulary Size	OOV-Rate	Word Error
B0	29k	14.0%	43.6%
B4	31k	13.6%	36.0%
B5	49k	8.7%	29.5%

Table 1: Recognition Results on Broadcast News.

3. HYPOTHESIS DRIVEN LEXICAL ADAPTATION

The procedure of Hypothesis Driven Lexical Adaptation has been introduced in [1]. HDLA is a two-pass approach where a first recognition run on a baseline dictionary is followed by a second recognition run with a dynamically adapted dictionary of the same size but a smaller out-of-vocabulary rate.

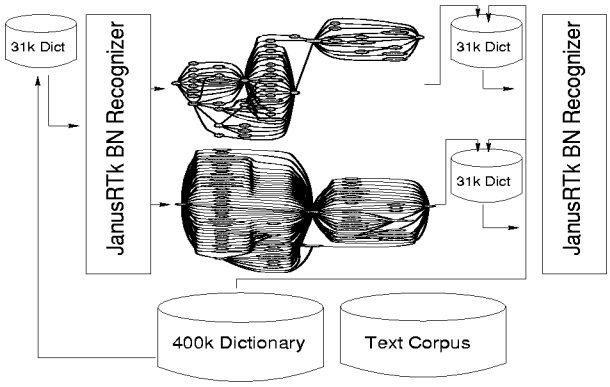

Figure 1: Lexical Adaptation based on Lattices. Two-pass recognition and vocabulary adaptation.

The basic idea is that in many cases errors are not due to misrecognitions but because the correct word was not part of the dictionary of the recognizer, thus constituting an out-of-vocabulary word. Most erroneous words tend to be acoustically very similar to the words that have actually been uttered. Our adaptation process makes use of this acoustic similarity: 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 the lattice leads to an utterance-specific vocabulary. This vocabulary list together with a fallback lexicon is used to create the lexicon for the second recognition run. The fallback lexicon contains all words that were found in the largest available text database for the task. Usually the language model training text is used for this purpose. Not only are all words seen in the text database included in the lexicon, but is also annotated with the frequencies of the words within this corpus. Based on this fallback lexicon and the generated word list the dynamically adapted vocabulary for the second recognition run is determined. The different criteria for selecting new vocabulary entries will be described in the following sections.

The algorithm below shows the overall **Hypothesis Driven Lexical Adaptation** process:

1. A first recognition run gives word lattices and an utterance-specific vocabulary list.
2. This vocabulary list is then used to look up all similar words in the fallback dictionary consisting of all words that were observed in the largest available text corpus.
3. All similar words 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).
4. In an automatic procedure a new dictionary and language model are created to perform a second recognition run.

4. MORPHOLOGY-BASED LEXICAL ADAPTATION

Results for lexical adaptation based on morphological knowledge were previously shown in [2]. The main idea is that for a large number of misrecognitions just the inflection ending is wrong, whereas the word stem was recognized correctly. As a consequence word stem equivalence is used as similarity criterion to generate the dictionary for the second recognition run. This leads to the following modification of the algorithm presented in the previous section:

2. a) The vocabulary list derived from the word lattice of the first recognition run is split into word stems and suffixes (where different combinations of word stem and suffix lengths were tested, see table 2). Note that the word stem length had to be at least two letters long.

2. b) The resulting word stem list is then used to look up all similar words in the fallback dictionary consisting of all words that were observed in the language model training text.

4.1. Results on Serbo-Croatian Data

This vocabulary adaptation procedure applied to Serbo-Croatian broadcast news data yields a significant improvement in terms of the out-of-vocabulary rate, which is reduced by 40% (see table 2), and in terms of accuracy by reducing the error rate by 5.8% (see table 3).

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

The same experiments as described above for system B4 were also performed on our latest B5 system. Starting off with a baseline performance of 29.5% word error and an out-of-vocabulary rate of 8.7%, we were able to reduce the number of out-of-vocabulary words to 4.8%. The 3.9% improvement in out-of-vocabulary rate was also reflected in a 3.5% improvement in word error rate yielding a performance of 26% word error.

	Vocabulary Size	OOV-Rate	Word Error
Baseline (B4)	31k	13.6%	36.0%
Adapted (B4)	31k	7.9%	30.2%
Baseline (B5)	49k	8.7%	29.5%
Adapted (B5)	49k	4.8%	26.0%

Table 3: Serbo-Croatian Recognition Results based on Adapted Vocabulary using Morphological Knowledge.

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

Table 4: German OOV-rates with different Splitting Methods. The baseline OOV-rate is 9.3%.

4.2. Results on German Data

Table 4 shows that the same result holds for German news data. Again a significant reduction of the out-of-vocabulary rate was observed. For German a fixed list of suffixes was used to create the word stems. Using this linguistic knowledge for decomposition also resulted in a huge out-of-vocabulary rate reduction from 9.3% to 6.0% (see table 4).

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).

5. PHONETIC-DISTANCE-BASED LEXICAL ADAPTATION

Dependence on knowledge about a specific language, especially linguistic knowledge about morphology, is not desirable. Speech recognition systems might be built by a person that is not an expert in the language to be recognized. The definition of classes of inflection endings and complex grammar rules would have to be supplied by a language expert. However, even if the knowledge is available, coming up with suitable grammatical rules and classification schemes is a tedious and extremely time-comsuming job. A better alternative would be to use knowledge inherent in the data itself (see the grapheme-based approach in section 6) or knowledge that can be acquired through tools that are included in the recognition system anyway.

As some kind of grapheme-to-phoneme conversion is necessary to generate the dictionary of a speech recognition system, usually a grapheme-to-phoneme tool is used to get a first baseline dictionary. In most cases this dictionary is then hand-corrected by human experts. For our experiment the available grapheme-to-phoneme tool was used to generate phoneme representations not only for the baseline dictionary of the system, but also for all words of the fallback dictionary retrieved from the web texts. The phonetic distance between words obtained from the word lattice and the words of the fallback lexicon was then used as similarity criterion to decide if a word was added to the dictionary for the second recognition run. As phonetic distance measure the Levenstein distance was used.

Modifying the already presented adaptation procedure for these needs resulted in the following step 2 of the HDLA algorithm:

step 2. The vocabulary list derived from the word lattice of the first recognition run is compared with all words of the fallback dictionary based on phonetic distances.

	Maximum Distance				
Minimum Length	1	2	3	4	5
3	8.7%	5.6%	4.5%	4.4%	–
4	8.7%	5.7%	4.5%	4.4%	–
5	8.7%	5.8%	4.3%	4.2%	–
6	8.7%	6.1%	4.3%	**4.0%**	4.0%
7	8.7%	6.7%	4.9%	4.1%	4.2%
8	8.7%	7.1%	5.5%	4.6%	4.3%
9	8.7%	7.6%	6.3%	5.7%	5.0%

Table 5: Serbo-Croatian OOV-rates with different minimum word lengths based on phonetic distances. The baseline OOV-rate is 8.74%.

5.1. Results on Serbo-Croatian Data

Again, a minimum length for a word had to be fixed and also a limit for a maximum distance had to be defined, in order to prevent HDLA from creating word lists where almost every word would be "similar" to the other. Different parameter combinations were tried (see table 5) where the optimum was found for a minimum length of 6 and a maximum distance of 4. For this combination the OOV-rate could be decreased by 55% from 8.7% to 4%.

Class	Phonemes
NOISES	+QK +hGH +hBR +nGN
CONSONANT	B C C1 C5 D D1 DZ5 F G H . . .
VOWEL	A E I O U
VOICED	B D D1 DZ5 G J L LJ M N NJ . . .
UNVOICED	C C1 C5 F K P S S5 T H
COMPACT	C1 D1 S5 Z5 K G H J
DIFFUSE	P B F M V

Table 6: Examples of **Serbo-Croatian Phoneme Classes**.

Within this experiment the distance of two phonemes was either 0 – if the phonemes were equal – or 1.0 otherwise. In a second experiment we considered a distance measure between phones that also takes similarity between different phonemes into acount. To this end we computed the Hamming distance with respect to a binary vector of phonetic features (see table 6) for each pair of phonemes. If for example two phonemes share the same phonetic features their distance is defined to be 0. If they have no features in common

their distance corresponds to the number of used phoneme classes. Examples of the phoneme classes used in our recognizer are given in table 6. Distances were normalized to 1.0 and the best parameter combination turned out to be a minimum word length of 4 and maximum distance of 0.7.

	Maximum Distance			
Minimum Length	0.5	0.6	0.7	0.8
3	5.7%	5.5%	5.4%	5.4%
4	5.7%	5.5%	**5.4%**	5.4%
5	6.0%	5.7%	5.6%	5.7%
6	6.2%	5.8%	5.6%	5.7%

Table 7: Serbo-Croatian OOV-rates with different minimum word lengths based on phonetic distances using phone-wise Hamming distances. The baseline OOV-rate is 8.74%.

Experiments at a minimum OOV-rate of 4.0% were performed on our baseline system with 29.5% word error. The reduction in OOV-rate was reflected in a 4.1% improvement of the word error rate to 25.4% (see table 8).

	Vocabulary Size	**OOV-Rate**	**Word Error**
Baseline (B5)	49k	8.7%	29.5%
Adapted (B5)	49k	4.0%	25.4%

Table 8: Serbo-Croatian Recognition Results based on Adapted Vocabulary using Phonetic Distances.

6. GRAPHEME-DISTANCE-BASED LEXICAL ADAPTATION

As Serbo-Croatian orthography closely matches its pronunciation, especially for the Serbo-Croatian language the use of literary language instead of phoneme representations provides a solution for the HDLA procedure.

Very similar to the last section distances between words of the word lattice and the fallback dictionary were calculated based on the letter sequences of these word pairs. The algorithm for calculating the distance was the same as the one used for the phonetical-distance-based approach.

6.1. Results on Serbo-Croatian Data

Results were almost as good as using phoneme distances. This is due to the easy-to-formulate rules of grapheme-to-phoneme conversion in Serbo-Croatian. In this experiment not even conversion had to be done, but the only thing needed was the large fallback lexicon retrieved from web texts. Best results were achieved for a parameter combination of minimum word length 6 and a maximum distance of 4. Here the OOV-rate was divided in half from 8.7% to 4.4% which is very close to the 4% achieved when using phoneme distances.

	Maximum Distance			
Minimum Length	1	2	3	4
3	8.6%	5.5%	4.5%	**4.4%**
4	8.6%	5.5%	4.4%	4.4%
5	8.6%	6.1%	4.8%	4.7%
6	8.6%	6.4%	5.0%	4.6%
7	8.6%	6.7%	5.4%	4.7%
8	8.6%	7.1%	5.9%	5.0%
9	8.7%	7.5%	6.6%	6.0%

Table 9: Serbo-Croatian OOV-rates with different minimum word lengths based on grapheme distances. The baseline OOV-rate is 8.74%.

7. CONCLUSIONS

With respect to the problem of encountering excessive growth of vocabularies in heavily inflected languages like Serbo-Croatian and German, Hypothesis Driven Lexical Adaptation turned out to be a very effective means of reducing the rate of out-of-vocabulary words. Morphological knowledge or distance measures can be used as similarity criterion within this procedure. The best results are obtained when using phonetic distances based on the Levenstein distance. For Serbo-Croatian the OOV-rate is reduced from 8.7% to 4% which results in a significant decrease of the word error rate from 29.5% to 25.4%. Compared to the use of morphological knowledge as similarity criterion for the HDLA procedure, the use of phoneme distances offers the distinct advantage of being easily applicable to any new language without the need of expert knowledge on the particular language.

8. 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.

9. REFERENCES

1. P. Geutner, M. Finke, and P. Scheytt. *Adaptive Vocabularies for Transcribing Multilingual Broadcast News.* Proceedings of the IEEE 1998 International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 1998. Seattle, Washington.
2. P. Geutner, M. Finke, P. Scheytt, A. Waibel, and H. Wactlar. *Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation.* Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, February 1998. Lansdowne, Virginia.
3. P. Scheytt, P. Geutner, and A. Waibel. *Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain.* Proceedings of the IEEE 1998 International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 1998. Seattle, Washington.