SELECTION CRITERIA FOR HYPOTHESIS DRIVEN LEXICAL ADAPTATION

P. Geutner

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

M. Finke and A. Waibel

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

ABSTRACT

Adapting the vocabulary of a speech recognizer to the utterance to be recognized has proven to be successful both in reducing high out-of-vocabulary as well as word error rates. This applies especially to languages that have a rapid vocabulary growth due to a large number of inflections and composita. This paper presents various adaptation methods within the Hypothesis Driven Lexical Adaptation (HDLA) framework which allow speech recognition on a virtually unlimited vocabulary. Selection criteria for the adaptation process are either based on morphological knowledge or distance measures at phoneme or grapheme level. Different methods are introduced for determining distances between phoneme pairs and for creating the large fallback lexicon the adapted vocabulary is chosen from. HDLA reduces the out-of-vocabulary-rate by 55% for Serbo-Croatian, 35% for German and 27% for Turkish. The reduced out-of-vocabulary rate also decreases the word error rate by an absolute 4.1% to 25.4% on Serbo-Croatian broadcast news data.

1. INTRODUCTION

A high rate of out-of-vocabulary words is one of the main problems when transcribing broadcast news shows in highly inflected languages such as Serbo-Croatian, German and Turkish. As a large number of unknown words automatically leads to degradation in recognition performance, a method that reduces the number of out-of-vocabulary words will also improve recognition results.

Our technique of adapting the recognition dictionary to the utterance to be recognized, Hypothesis Driven Lexical Adaptation (HDLA), is able to reduce both out-of-vocabulary and word error rates significantly. The multipass procedure uses the outcome of a first recognition run on a baseline dictionary to determine an adapted vocabulary for the second run. Our HDLA approach offers a variety of different selection criteria for the adaptation procedure. This paper presents an overview over all applied methods that either use morphological knowledge of a language or distance-based measures at phoneme or grapheme level. Moreover all approaches rely on the availability of a large fallback lexicon that can either be derived from a large text database or created artificially through morphological rules. Results on reduced out-of-vocabulary rates on Serbo-Croatian and German broadcast news data as well as Turkish dictation data are introduced and also improved Serbo-Croatian recognition results are presented.

2. THE SPEECH RECOGNITION ENGINE

The speech recognition system used to perform all experiments for transcribing Serbo-Croatian broadcast news shows is 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 an MFCC based feature vector 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 are trained on the hand-transcribed acoustic training data and an additional 11.8 million words of text data collected on the internet. Performance of the baseline system with an out-of-vocabulary rate of 8.7% as well as results achieved by using HDLA are shown in table 1 below.

	Vocabulary Size	**OOV-Rate**	**Word Error**
Baseline	49k	8.7%	29.5%
Morphology-Based Adaptation	49k	4.8%	26.0%
Phonetic-Distance Based Adaptation	49k	4.0%	25.4%

Table 1: **Serbo-Croatian Recognition** Results based on Adapted Vocabulary using Morphological Knowledge and Phonetic Distances.

	Vocabulary Size	**OOV-Rate**	**Word Error**
Baseline	60k	4.4%	24.7%

Table 2: **German Recognition** Baseline Results.

3. HYPOTHESIS DRIVEN LEXICAL ADAPTATION

Adapting the vocabulary of a speech recognizer to a particular utterance to be recognized has proven to be successful both in reducing high out-of-vocabulary as well as word error rates. This applies especially to languages that have a rapid vocabulary growth and consist of a large number of inflection endings and composita. Figure 1 shows the vocabulary growth of highly inflected languages like Serbo-Croatian, German, Turkish and Korean compared to English. The algorithm below describes the **Hypothesis Driven Lexical Adaptation** process:

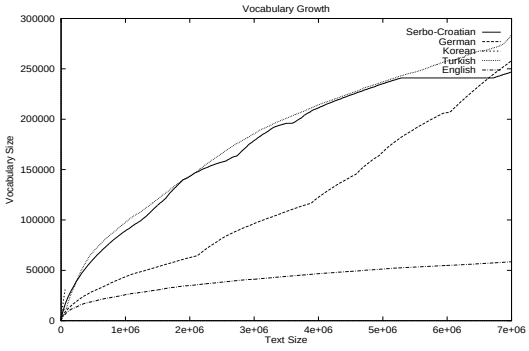

Figure 1: **Vocabulary Growth** of several Languages.

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 full 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 replacing 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 is created to perform a second recognition run.

Within this lexical adaptation procedure various criteria for selecting the adapted vocabulary can be applied. Depending on the special characteristics of the respective language different procedures lead to optimal performance. Figure 2 summarizes the methods that have been applied within our Hypothesis Driven Lexical Adaptation procedure and are presented in this paper:

1. Linguistic knowledge about morphology
2. Phonetic distance
3. Grapheme distance
4. The combination of phonetic distance and an artificially created fallback lexicon.

Whereas the first three methods rely on the availability of a large fallback lexicon obtained through a huge text database for the language in question, the last method uses linguistic knowledge to create the fallback lexicon artificially. When using the second method, phonetic distance measures, to determine the adapted vocabulary for the second recognition run, three different sources to obtain the distance of two phonemes can be used (see section 5). All methods as well as their results achieved on Serbo-Croatian and German broadcast news and Turkish dictation data will be described in the following sections.

Hypothesis Driven Lexical Adaptation

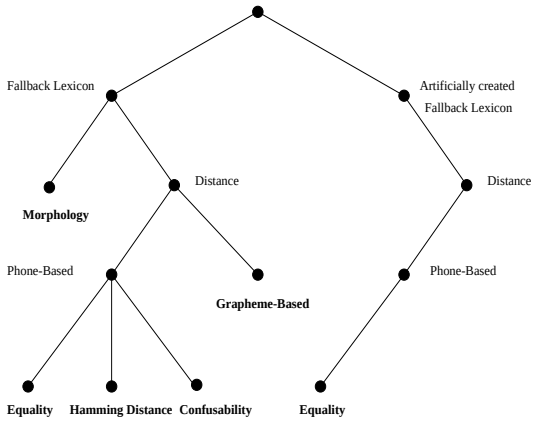

Figure 2: **Overview** over all HDLA Methods.

	Serbo-Croatian	German	Turkish
	OOV-Rates		
Baseline	8.7%	4.4%	14.9%
Morphology-Based	4.8%	**2.9%**	10.9%
Equality	**4.0%**	3.3%	–
Hamming distance	5.4%	–	–
Acoustic Confusability	4.4%	–	–
Grapheme-Based	**4.0%**	–	–
Artificially created Fallback Lexicon	5.8%	–	–

Table 3: **OOV-rates** for Serbo-Croatian, German and Turkish data on all HDLA approaches.

4. MORPHOLOGY-BASED LEXICAL ADAPTATION

For the morphology-based approach linguistic knowledge about inflection endings and suffixes is used as adaptation criterion as already introduced in [1]. Basically this method takes advantage of the fact that for a lot of misrecognized words the word stem is recognized correctly and only the ending is wrong.

Experiments on different wordstem and suffix lengths were performed (see table 4). For our Serbo-Croatian broadcast news data this procedure was able to reduce the out-of-vocabulary rate by 45% from 8.7% to 4.8% and also decrease word error rate by an absolute 3.5% from 29.5% to 26%. The same applies to the German broadcast news data. Starting off with a baseline out-of-vocabulary rate of 4.4% a 35% improvement is achieved which yields an out-of-vocabulary rate of 2.9%. Additional tests on Turkish read newspaper articles with a baseline out-of-vocabulary rate of 14.9% resulted in a 27% decrease leading to an out-of-vocabulary rate of 10.9%.

5. PHONETIC-DISTANCE-BASED LEXICAL ADAPTATION

As dependence on linguistic knowledge is not desirable, in a second experiment phonetic distances between phonemes are used as

	Wordstem Length		
Suffix Length	4	5	6
Serbo-Croatian (1+2+3+4)	–	**4.8%**	–
German (fixed)	3.2%	**2.9%**	3.1%
Turkish (fixed)	–	**10.9%**	11.2%

Table 4: **OOV-rates** using Morphology-Based Adaptation. The baseline OOV-rate for Serbo-Croatian is 8.7%, for German 4.4% and for Turkish 14.9%.

similarity criterion for the HDLA procedure (see also [2]). To determine the distance between two phonemes three different methods were experimented with:

1. Equality
2. Hamming distance with respect to a binary vector of phonetic features
3. Acoustic confusability.

As distance measure the Levenshtein distance is used in all three cases.

5.1. Equality

Using the equality criterion the distance of two phonemes p and q is either 0, if the phonemes are identical, or 1 otherwise. Results of this experiment are shown in table 5 for Serbo-Croatian and German.

$$distance(p,q) = (1 - \delta(p,q)) \tag{1}$$

where

$$\delta(p,q) = \begin{cases} 1 & : \text{ if } p = q \\ 0 & : \text{ if } p \neq q \end{cases}$$

Serbo-Croatian

	Maximum Distance				
Minimum Length	1	2	3	4	5
5	8.7%	5.8%	4.3%	4.2%	–
6	8.7%	6.1%	4.3%	**4.0%**	4.0%
7	8.7%	6.6%	4.9%	4.1%	4.2%

German

	Maximum Distance			
Minimum Length	1	2	3	4
5	3.9%	3.5%	3.2%	3.3%
6	3.8%	3.6%	3.1%	3.4%
7	3.8%	3.6%	3.2%	**3.1%**

Table 5: **Serbo-Croatian and German OOV-rates** with different minimum word lengths based on phonetic distances according to equation 1. The baseline OOV-rates are 8.7% for Serbo-Croatian and 4.4% for German.

5.2. Hamming distance

In a second experiment a distance measure that also considers similarity between different phonemes is used. To this end the Hamming distance with respect to a binary vector of phonetic features $f_i(p)$ for each pair of phonemes is computed. Results can be found in table 6.

$$distance(p,q) = \frac{1}{n}\sum_{i=1}^{n}(1 - \delta(f_i(p), f_i(q))) \tag{2}$$

where $f_i(p)$ is the phonetic feature representation of phoneme p and n the number of phonetic features.

	Maximum Distance					
Minimum Length	0.4	0.5	0.6	0.7	0.8	0.9
3	6.1%	5.7%	5.5%	5.5%	5.5%	5.5%
4	6.1%	5.7%	5.5%	**5.4%**	5.4%	5.4%
5	6.2%	6.0%	5.7%	5.6%	5.7%	5.7%

Table 6: **Serbo-Croatian OOV-rates** with different minimum word lengths based on phonetic distances using phone-wise Hamming distances. The baseline OOV-rate is 8.7%.

5.3. Acoustic Confusability

The last method for estimating distances between two phonemes is their acoustic confusability. All possible confusion pairs are determined by conducting a recognition experiment on several broadcast news shows of the training material. The phonetic distance between two phonemes is measured through the phonetic alignment of the hypothesized recognizer output and the actually uttered words. To allow the calculation of distances between phonemes that have never been confused in the training data a backoff scheme like in [3] is applied. Results of this experiment on Serbo-Croatian data can be found in table 7.

$$distance(p,q) = -\log(prob(p,q)) \tag{3}$$

where $prob(p,q)$ is the probability of confusing phoneme p with phoneme q derived from recognition experiments.

	Maximum Distance					
Minimum Length	8	9	10	11	12	13
5	4.7%	4.6%	4.7%	4.7%	4.7%	4.7%
6	4.7%	4.4%	4.5%	**4.4%**	4.4%	4.5%

Table 7: **Serbo-Croatian OOV-rates** with different minimum word lengths based on phonetic distances using speech recognizer acoustic confusability and a backoff scheme. The baseline OOV-rate is 8.7%.

6. GRAPHEME-DISTANCE-BASED LEXICAL ADAPTATION

As Serbo-Croatian orthography closely matches its pronunciation, also distances based on literary language instead of phonetic representations were used. As expected the best result is identical to the phoneme-based approach (see also table 8).

	Maximum distance			
Minimum Length	1	2	3	4
5	8.6%	5.7%	4.3%	4.3%
6	8.6%	6.0%	4.3%	**4.0%**
7	8.6%	6.5%	4.9%	4.2%

Table 8: **Serbo-Croatian OOV-rates** with different minimal word lengths based on grapheme distances. The baseline OOV-rate is 8.7%.

7. ARTIFICIAL CREATION OF THE FALLBACK LEXICON

All methods presented so far need a huge fallback dictionary for choosing the vocabulary of the second recognition run. For some languages large enough text databases might not be available to build a lexicon like this. Therefore, we used linguistic knowledge about morphology to create the fallback lexicon artificially. Similar to the morphology-based approach language-specific rules for morphological inflection endings are defined. The artificially generated fallback lexicon is created through the following procedure: For each word of the baseline dictionary used for the first recognition run the appropriate rule is applied. New words are generated and included into a fallback dictionary.

Interpretation of rules can be done very strictly or more loose. Also some rules might not only generate lexically correct words but also words that are correct according to the syntax of the rule but have no semantic meaning. Various levels of strictness to interprete rules have been experimented with. The employed level of strictness influences the size of the fallback dictionary. The more importance is attached to generating legal words only, the smaller is the size of the resulting word list. If more freedom is allowed, the size of the fallback lexicon increases, thereby also including a higher percentage of "illegal" words. Different rule sets have been applied and the generated words were then compared with the 300k lexicon containing legal words only.

Systems	# of Rules	Rule Interpretation
ART-1	28	loose
ART-2	28	strict
ART-3	18	loose
ART-4	18	strict

Table 9: **Variations** on the artificial creation of the Fallback Lexicon.

Table 9 shows different combinations on the number of rules applied and different interpretation levels. It turned out that the best strategy was to be very loose in rule interpretation and allow to generate a high percentage of illegal words. Table 10 shows the resulting out-of-vocabulary rates of experiments ART-1 - ART-4.

	Systems			
	ART-1	ART-2	ART-3	ART-4
# of artificially created words	487.643	105.300	140.272	24.798
size of fallback lexicon	491.132	126.873	121.053	54.609
% legal words generated	11.6%	31.9%	24.2%	35.4%
% legal words in lexicon	12.2%	19.7%	33.4%	66.2%
OOV-rate	**5.8%**	6.8%	6.5%	7.1%

Table 10: **Results** on different methods to create the Fallback Lexicon.

8. CONCLUSIONS

We have shown that reduction of out-of-vocabulary rates for highly inflected languages, like Serbo-Croatian, German and Turkish, by up to 55% also yield significant reductions in word error rate. With a 41k vocabulary in Serbo-Croatian, e.g., vocabulary adaptation achieves an out-of-vocabulary rate that would require a recognition dictionary of 170k. Our HDLA approach offers different methods where the optimal one varies depending on the particular language. Best results in Serbo-Croatian are achieved by phoneme- or grapheme-based distance measures, whereas the out-of-vocabulary rate reduction in German and Turkish reaches its optimum by using the morpheme-based approach.

9. 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. Thanks to all members of the Interactive System Labs that contributed to this paper, especially Peter Scheytt.

10. REFERENCES

[1] P. Geutner, M. Finke, and P. Scheytt. Adaptive Vocabularies for Transcribing Multilingual Broadcast News. In *ICASSP'98*, Seattle, Washington.

[2] P. Geutner, M. Finke, and A. Waibel. Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News. In *ICSLP'98*, Sydney, Australia.

[3] R. Kneser and H. Ney. Improved Clustering Techniques for Class-Based Statistical Language Modelling. In *Eurospeech'93*, Berlin, Germany.