Speech Recognition of Czech - Inclusion of Rare Words Helps

Petr Podveský and **Pavel Machek**
Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic
{podvesky,machek}@ufal.mff.cuni.cz

Abstract

Large vocabulary continuous speech recognition of inflective languages, such as Czech, Russian or Serbo-Croatian, is heavily deteriorated by excessive out of vocabulary rate. In this paper, we tackle the problem of vocabulary selection, language modeling and pruning for inflective languages. We show that by explicit reduction of out of vocabulary rate we can achieve significant improvements in recognition accuracy while almost preserving the model size. Reported results are on Czech speech corpora.

1 Introduction

Large vocabulary continuous speech recognition of inflective languages is a challenging task for mainly two reasons. Rich morphology generates huge number of forms which are not captured by limited-size dictionaries, and therefore leads to worse recognition results. Relatively free word order admits enormous number of word sequences and thus impoverishes n-gram language models. In this paper we are concerned with the former issue.

Previous work which deals with excessive vocabulary growth goes mainly in two lines. Authors have either decided to break words into sub-word units or to adapt dictionaries in a multi-pass scenario. On Czech data, (Byrne et al., 2001) suggest to use linguistically motivated recognition units. Words are broken down to stems and endings and used as the recognition units in the first recognition phase. In the second phase, stems and endings are concatenated. On Serbo-Croatian, (Geutner et al., 1998) also tested morphemes as the recognition units. Both groups of authors agreed that this approach is not beneficial for speech recognition of inflective languages. Vocabulary adaptation, however, brought considerable improvement. Both (Icring and Psutka, 2001) on Czech and (Geutner et al., 1998) on Serbo-Croatian reported substantial reduction of word error rate. Both authors followed the same procedure. In the first pass, they used a dictionary composed of the most frequent words. Generated lattices were then processed to get a list of all words which appeared in them. This list served as a basis for a new adapted dictionary into which morphological variants were added.

It can be concluded that large corpora contain a host of words which are ignored during estimation of language models used in first pass, despite the fact that these rare words can bring substantial improvement. Therefore, it is desirable to explore how to incorporate rare or even unseen words into a language model which can be used in a first pass.

2 Language Model

Language models used in a first pass of current speech recognition systems are usually built in the following way. First, a text corpus is acquired. In case of broadcast news, a newspaper collection or news transcriptions are a good source. Second, most frequent words are picked out to form a dictionary. Dictionary size is typically in tens of thousand words. For English, for example, dictionaries of size

Proceedings of the ACL Student Research Workshop, pages 121–126,
Ann Arbor, Michigan, June 2005.

of 60k words sufficiently cover common domains. (Of course, for recognition of entries listed in the Yellow pages, such limited dictionaries are clearly inappropriate.) Third, an n-gram language model is estimated. In case of Katz back-off model, the conditional bigram word probability is estimated as

$$P_1(w_i|w_{i-1}) = \begin{cases} \widetilde{P}(w_i|w_{i-1}) \text{ if } C(w_{i-1}, w_i) > k \\ BO(w_{i-1}) \cdot \widetilde{P}(w_i) \quad \text{otherwise} \end{cases} \tag{1}$$

where $\widetilde{P}$ represents a smoothed probability distribution, $BO()$ stands for the back-off weight, and $C(.)$ denotes the count of its argument. Back-off model can be also nicely viewed as a finite state automaton as depicted in Figure 1.

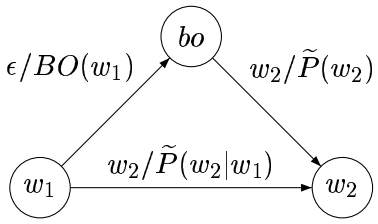

Figure 1: A fragment of a bigram back-off model represented as a finite-state automaton.

To alleviate the problem of a high OOV, we suggest to gather supplementary words and add them into the model in the following way.

$$P(w_i|w_{i-1}) = \begin{cases} P_1(w_i|w_{i-1}) & w_i \in D \\ BO(w_{i-1}) \cdot Q(w_i) & w_i \in S \end{cases} \tag{2}$$

$P_1()$ refers to the regular back-off model, D denotes the regular dictionary from which the back-off model was estimated, S is the supplementary dictionary which does not overlap with D.

Several sources can be exploited to obtain supplementary dictionaries. Morphology tools can derive words which are close to those observed in corpus. In such a case, $Q(w_i)$ can be set as a constant function and estimated on held-out data to maximize recognition accuracy.

$$Q(w_i) = const \quad \text{for } w_i \text{ generated by morphology} \tag{3}$$

Having prior domain knowledge, new words which are expected to appear in audio recordings might be collected and added into S. Consider an example of transcribing an ice-hockey tournament. Names of new players are desirably in the vocabulary. Another source of S are the words which fell below the selection threshold of D. In large corpora, there are hundreds of thousands words which are omitted from the estimated language model. We suggest to put them into S. As it turned out, unigram probability of these words is very low, so it is suitable to increase their score to make them competitive with other words in D during recognition. $Q(w_i)$ is then computed as

$$Q(w_i) = \mathit{shift} \cdot f(w_i) \tag{4}$$

where $f(w_i)$ refers to the relative frequency of w_i in a given corpus, *shift* denotes a shifting factor which should be tuned on some held-out data.

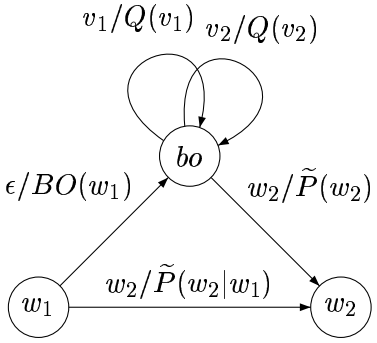

Figure 2: A fragment of a bigram back-off model injected by a supplementary dictionary

Note that the probability of a word given its history is no longer proper probability. It does not adds up to one. We decided not to normalize the model for two reasons. First, we used a decoder which searches for the best path using Viterbi criterion, so there's no need for normalization. Second, normalization would have involved recomputing all back-off model weights and could also enforce re-tuning of the language model scaling factor. To rule out any variation which the re-tuning of the scaling factor could bring, we decided not to normalize the new model.

In finite-state representation, injection of a new dictionary was implemented as depicted in Figure 2. Supplementary words form a loop in the back-off state.

3 Experiments

We have evaluated our approach on two corpora, Czech Broadcast News and the Czech portion of MALACH data.

3.1 Czech Broadcast News Data

The Czech Broadcast News (Radová et al., 2004) is a collection of both radio and TV news in Czech. Weather forecast, traffic announcements and sport news were excluded from this corpus. Our training portion comprises 22 hours of speech. To tune the language model scaling factor and additional LM parameters, we set aside 100 sentences. The test set consists of 2500 sentences.

We used the HTK toolkit (Young et al., 1999) to extract acoustic features from sampled signal and to estimate acoustic models. As acoustic features we used 12 Mel-Frequency Cepstral Coefficients plus energy and delta and delta-delta features. We trained a triphone acoustic model with tied mixtures of continuous density Gaussians.

As a LM training corpus we exploited a collection of newspaper articles from the Lidové Noviny (LN) newspaper. This collection was published as a part of the Prague Dependency Treebank by LDC (Hajič et al., 2001). This corpus contains 33 million tokens. Its vocabulary contains more than 650k word forms. OOV rates are displayed in Table 1.

Dict. size	OOV
60k	8.27%
80k	6.92%
124k	5.20%
371k	2.23%
658k	1.63%

Table 1: OOV rate of transcriptions of the test data. Dictionaries contain the most frequent words.

As can be readily observed, moderate-size vocabularies don't sufficiently cover the test data transcriptions. Therefore they are one of the major sources of poor recognition performance.

The baseline language model was estimated from 60k most frequent words. It was a bigram Katz back-off model with Knesser-Ney smoothing pruned by the entropy-based method (Stolcke, 1998).

As the supplementary dictionary we took the rest of words from the LN corpus. To learn the impact of injection of infrequent words, we carried out two experiments.

First, we built a uniform loop which was injected into the back-off model. The uniform distribution was tuned on the held-out data. Tuning of this constant is displayed in Table 2.

Uniform scale	WER
12	18.89%
11	18.68%
10	18.40%
9	21.00%

Table 2: Tuning of uniform distribution on the held-out set. *WER* denotes the word error rate.

Second, we took relative frequencies multiplied by a shift coefficient as the injected model scores. This shift coefficient was again tuned on held-out data as shown in Table 3.

Unigram shift	WER
no shift	19.52%
e^3	18.54%
e^4	17.91%
e^5	18.75%

Table 3: Tuning of the shift coefficient of unigram model on the held-out set.

Then, we took the best parameters and used them for recognition of the test data. Recognition results are depicted in Figure 4. The injection of supplementary words helped decrease both recognition word error rate and oracle word error rate. By oracle WER is meant WER of the path, stored in the generated lattice, which best matches the utterance regardless the scores. In other words, oracle WER gives us a bound on how well can we get by tuning scores in a given lattice. Injection of shifted unigram model brought relative improvement of 13.6% in terms of WER over the 60k baseline model. Uniform injection brought also significant improvement despite its simplicity. Indeed, we observed more than 10% relative improvement in terms of WER. In terms of oracle WER, unigram injection brought more than 30% relative improvement.

Model	WER	OWER
Baseline 60k	29.17%	15.90%
Baseline 80k	27.44%	14.31%
60k + Uniform injection	26.12%	11.10%
60k + Unigram injection	25.21%	11.03%

Table 4: Evaluation on 2500 test sentences. *OWER* stands for the oracle error rate.

It's worthwhile to mention the model size, since it could be argued that the improvement was achieved by an enormous increase of the model. We decided to measure the model size using two factors. The disk space occupied by the language model and the disk space taken up by the so-called *CLG*. By *CLG* we mean a transducer which maps triphones to words augmented with the model scores. This transducer represents the search space investigated during recognition. More details on transducers in speech recognition can be found in (Mohri et al., 2002). Table 5 summarizes the sizes of the evaluated models.

Model	CLG size	G size
Baseline 60k	399MB	106MB
60k + Uniform	405MB	115MB
60k + Unigram	405MB	115MB
Baseline 80k	441MB	116MB

Table 5: Model size comparison measured in disk space. *G* denotes a language model compiled as a finite-state automaton. *CLG* denotes transducer mapping triphones to words augmented with model scores.

Injection of supplementary words increased the model size only slightly. To see the difference in the size of injected models and traditionally built ones, we constructed a model of 80k most frequent words and pruned with the same threshold as the 60k LM. Not only did this 80k model give worse recognition results, but it also proved to be bigger.

3.2 MALACH Data

The next data we tested our approach on was the Czech portion of the MALACH corpus (`http://www.clsp.jhu.edu/research/malach`). MALACH is a multilingual audio-visual corpus. It contains recordings of survivors of World War II talking about war events. 600 people spoke in Czech, but only 350 recordings had been digitized till end of 2003. The interviewer and the interviewee had separate microphones, and were recorded on separate stereo channels. Recordings were stored in the MPEG-1 format. Average length of a testimony is 1.9 hours.

30 minutes from each testimony were transcribed and used as training data. 10 testimonies were transcribed completely and used for testing. The acoustic model used 15-dimensional PLP cepstral features, sampled at 10 msec. Modeling was done using the HTK Toolkit.

The baseline language model was estimated from transcriptions of the survivors' testimonies. We worked with the standardized version of the transcriptions. More details regarding the Czech portion of the MALACH data can be found in (Psutka et al., 2004). Transcriptions are 610k words long and the entire vocabulary comprises 41k words. We refer to this corpus as *TR_41k*.

To obtain a supplementary vocabulary, we used Czech morphology tools (Hajič and Vidová-Hladká, 1998). Out of 41k words we generated 416k words which were the inflected forms of the observed words in the corpus. Note that we posed restrictions on the generation procedure to avoid obsolete, archaic and uncommon expressions. To do so, we ran a Czech tagger on the transcriptions and thus obtained a list of all morphological tags of observed forms. The morphological generation was then confined to this set of tags.

Since there is no corpus to train unigram scores of generated words on, we set the LM score of the generated forms to a constant.

The transcriptions are not the only source of text data in the MALACH project. (Psutka et al., 2004) searched the Czech National Corpus (CNC) for sentences which are similar to the transcriptions. This additional corpus contains almost 16 million words, 330k types. CNC vocabulary overlaps to a large extent with TR vocabulary. This fact is not surprising since the selection criterion was based on a lemma unigram probability. Table 6 summarizes OOV rates of several dictionaries.

We estimated several language models. The baseline models are pruned bigram back-off models with Knesser-Ney smoothing. The baseline word error

Dictionary		OOV
Name	Size	
TR41k	41k	5.07 %
TR41k + Morph416k	416k	2.74 %
TR41k + CNC60k	79k	3.04 %
TR41k + CNC100k	114k	2.62 %
TR41k + CNC160k	171k	2.25%
TR41k + CNC329k	337k	1.76 %
All together	630k	1.46 %

Table 6: OOV for several dictionaries. *TR*, *CNC* denote the transcriptions, the Czech National Corpus, respectively. *Morph* refers to the dictionary generated by the morphology tools from from *TR*. Numbers in the dictionary names represent the dictionary size.

rate of the model built solely from transcriptions was 37.35%. We injected constant loop of morphological variants into this model. In terms of text coverage, this action reduced OOV from 5.07% to 2.74%. In terms of recognition word error rate, we observed a relative improvement of 3.5%.

In the next experiment we took as the baseline LM a linear interpolation of the LM built from transcriptions and a model estimated from the CNC corpus. Into this model, we injected a unigram loop of all the available words. That is the rest of words from the CNC corpus with unigram scores and words provided by morphology which were not already in the model. Table 7 summarizes the achieved WER and oracle WER. Given the fact that the injection only slightly reduced the OOV rate, a small relative reduction of 2.3% matched our expectations.

Model	Acc	OAcc
TR41k	37.35%	14.40%
TR41k + Uniform_Morph	36.06%	12.48%
TR41k + CNC_100k	34.47%	11.95%
TR41k + CNC_100k + Inj	33.67%	10.79%
TR41k + CNC_160k	34.19%	11.65%

Table 7: Word error rate and oracle WER for baseline and injected models. *Uniform_Morph* refers to the constant uniform loop of the morphology-generated words. *Inj* denotes the loop of the rest of words of the CNC corpus and the morphology-generated words.

To learn how the injection affected model size, we measured size of the language model automaton and the optimized triphone-to-word transducer. As in the case of the LN corpus, injection increased the model size only moderately. Sizes of the models are shown in Table 8.

model	CLG	G
TR41k	38MB	5.6MB
TR41k + Morph	54MB	11MB
TR41k + CNC_100k	283MB	53MB
TR41k + CNC_100k + Inj	307MB	61MB
TR41k + CNC_160k	312MB	59MB

Table 8: Disk usage of tested models. *G* refers to a language model compiled into an automaton, *CLG* denotes triphone-to-word transducer. *CNC* and *Morph* refer to a LM estimated from transcriptions and the Czech National Corpus, respectively. Morph represents the loop of words generated by morphology. *Inj* is the loop of all words from *CNC* which were not included in *CNC* language model, moreover, *Inj* also contains words generated by the morphology.

4 Conclusion

In this paper, we have suggested to inject a loop of supplementary words into the back-off state of a first-pass language model. As it turned out, addition of rare or morphology-generated words into a language model can considerably decrease both recognition word error rate and oracle WER in single recognition pass. In the recognition of Czech Broadcast News, we achieved 13.6% relative improvement in terms of word error rate. In terms of oracle error rate, we observed more than 30% relative improvement. On the MALACH data, we attained only marginal word error rate reduction. Since the text corpora already covered the transcribed speech relatively well, a smaller OOV reduction translated into a smaller word error rate reduction. In the near future, we would like to test our approach on agglutinative languages, where the problems with high OOV are even more challenging. We would also like to experiment with more complex language models.

5 Acknowledgements

We would like to thank our colleagues from the University of Western Bohemia for providing us with acoustic models. This work has been done under the support of the project of the Ministry of Education of the Czech Republic No. MSM0021620838 and the grant of the Grant Agency of the Charles University (GAUK) No. 375/2005.

References

W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka. 2001. On large vocabulary continuous speech recognition of highly inflectional language - Czech. In *Eurospeech 2001*.

P. Geutner, M. Finke, and P. Scheytt. 1998. Adaptive Vocabulariesfor Transcribing Multilingual Broadcast News. In *ICASSP*, Seattle, Washington.

Jan Hajič and Barbora Vidová-Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the Conference COLING ACL '98*, pages 483-490, Mountreal, Canada.

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. Prague dependency treebank 1.0. Linguistic Data Consortium (LDC), catalog number LDC2001T10.

P. Icring and J. Psutka. 2001. Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary. In *TSD*, Železnaá Ruda, Czech Republic.

M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16:69-88.

J. Psutka, P. Ircing, V. Radova, and J. V. Psutka. 2004. Issues in annotation of the Czech spontaneous speech corpus in the MALACH project. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Vlasta Radová, Josef Psutka, Luděk Müller, William Byrne, J.V. Psutka, Pavel Ircing, and Jindřich Matoušek. 2004. Czech broadcast news speech. Linguistic Data Consortium (LDC), catalog number LDC2004S01.

A. Stolcke. 1998. Entropy-based pruning of backoff language models. In *In Proceedings of the ARPA Workshop on Human Language Technology.*

S. Young et al. 1999. *The HTK Book*. Entropic Inc.