LANGUAGE MODELS FOR AUTOMATIC SPEECH RECOGNITION OF CZECH LECTURES

Tomáš Mikolov
Doctoral Degree Programme (4), FIT BUT
E-mail: imikolov@fit.vutbr.cz

Supervised by: Jan Černocký
E-mail: cernocky@fit.vutbr.cz

ABSTRACT

This paper describes improvements in Automatic Speech Recognition (ASR) of Czech lectures obtained by enhancing language models. Our baseline is a statistical trigram language model with Good-Turing smoothing, trained on half billion words from newspapers, books etc. The overall improvement from adding more training data is over 10% in accuracy absolute, while using advanced language modeling techniques - mainly neural networks - yields another 3%. Perplexity improvements and OOV reduction are discussed too.

1 INTRODUCTION

Automatic speech recognition is relatively new technology, which is still not widely used because of several reasons. Probably the most painful is the overall low accuracy, mainly for languages with little training data (this is true for almost all languages except English). The other reason is high computational cost of current systems. Still, this technology is a must have in the future, mainly in cases where human annotator is too expensive (automatic generation of subtitles for films, news, meetings, ..), or too untrustworthy (terrorism, organized crime).

Building a Large Vocabulary Continuous Speech Recognition system (LVCSR) for Czech spontaneous speech, with highly specialized topic - university lectures - is therefore a very challenging task.

Typical ASR system uses so called acoustic models for detection of phonemes in the speech. Phoneme sequences are then translated into a word sequences using language model. For Czech, statistical language models are quite difficult to build. The main reason is simply insufficient training data - it is very hard to obtain written materials for modeling spontaneous speech and data that would contain technical expressions used in the lectures. Our main source of mistakes comes from the Out Of Vocabulary words (OOVs) - current systems are unable to recognize words that were not present in the training data. The other problem is the flexibility of Czech, resulting in many times bigger vocabulary than is usual for English.

2 BASELINE

Our baseline language model was trained using SRILM toolkit [2] on a huge corpora (half billion words) consisting mainly of newspaper articles, books, etc. Since these data are highly inappropriate for modeling spontaneous speech, the results are very poor. In fact, the recognizer is able to detect mostly only auxiliary words. The intended use of recognizer is opposite - we are highly interested in detection of technical terms, which will be used by students to find the appropriate lecture.

It is clear that the first step to improve our system is to add more training data, which will be suitable for modeling spontaneous speech and which will contain technical terms. With additional data, it may be necessary to use advanced language modeling techniques, to avoid sparse data problem caused by the inflectionality of Czech language.

Our baseline language model is a statistical trigram language model with Good-Turing smoothing (to see details about language modeling basics, see [1][4]). Language model (LM) usefulness may be basically estimated using validation data (data that should match closely the test data) - we can compute Out Of Vocabulary (OOV) rate, which determines how many words are not present in the LM. Since decoder can not output OOV word, these are our primary source of errors. Second, we can compute perplexity (PPL) - this number denotes average branching factor of the LM when applied on the data. In simple words, if we have perplexity 200, it means that the LM is on average choosing between 200 equally probable words when predicting the next one. Of course, language model that predicts the next word better has a lower perplexity value. Usually, lower perplexity should mean better recognition accuracy.

Note that despite the fact that statistical n-gram language models are widely used for decades, most attempts to improve them significantly have failed. It is out of scope of this simple article to fully discuss the reason, but basically - n-gram statistics is converging to optimum with more training data and the computational cost is so low that beating n-gram language models in domains where we have billions of words for training is very hard.

3 ADDITIONAL DATA

Since our baseline results proved to be quite poor, mainly because of inappropriate training data, we have chosen several possible sources of data that should be helpful for our task. First, since the lecture speech is usually spontaneous, we needed some data that would model this type of speech well. The spontaneous part of the speech was somehow covered by data from 'Prazsky mluveny korpus' and 'Brnensky mluveny korpus' (Prague and Brno spoken corpora).

Our most painful problem was high OOV rate - in most cases, we are most interested in detection of technical terms. If such term is not in the vocabulary, decoder can not produce it. Instead, it produces some word that is acoustically close - however, this can be even more than one word, usually changing the meaning of the utterance dramatically. Moreover, language models tend to get confused in these cases - it was estimated (for example by [5]) that one OOV in the utterance 'produces' 1.6 recognition mistakes.

To solve this problem, we have included some training data from Masaryk University's Faculty of Informatics and data from written lecture materials (BUT FIT). Although these data are not spontaneous, they reduce the OOV rate significantly.

Results in table 1 are obtained while recognizing one test lecture with basic acoustic models supervisedly adapted on the speaker. It is clear that while Prazsky & Brnensky spoken corpora (PBSC) are improving mainly the perplexity (as it was expected, since these data are sponta-

Data source	Perplexity	OOV rate	Accuracy
General data	1035.9	7.8%	60.1%
General + PBSC	690.6	7.2%	61.0%
General + MU	932.5	4.6%	62.6%
General + FIT	967.7	3.4%	63.2%
All data	606.9	2.6%	66.0%

Table 1: Speech recognition accuracy

neous, as is the lecture), the data from FIT BUT and FI MU are decreasing the OOV rate. Of course, the best results were obtained after using all available data.
It can be seen in table 2 that our general data corpus is huge in comparison to the others. The full vocabulary for this corpora was over 2 million of unique words, which is too much. After cutting off all words that have occurred less than 30 times, the 360K vocabulary was obtained. Despite the fact that new corpora are tiny in comparison to the general one, the improvements in the speech recognition accuracy show us important fact - it is much better to have small corpora that are similar to test data than to have huge off topic corpus.

Data source	# words	Vocabulary size
General data	500 milion	360 900
PBSC	1 170 000	69 300
FI MU	184 000	23 300
FIT BUT	1 732 000	70 100

Table 2: Number of words in training data and vocabulary sizes

The results obtained after adding more data are encouraging, but the accuracy is still insufficient for practical use. As we have seen, the more the training data match testing data, the better accuracies of ASR system we may expect. The most relevant data in our case are lecture annotations. The problem is that obtaining these data is very costly, and we can not hope to gather more than just a few hundred thousand words. On the other hand, general corpus is very huge, making it's use quite impractical.

4 NEURAL NETWORKS

As we already know, the better data we have, the better the final system shall be. On the other hand, the better data we want, the more costly they would be, and the less of them we shall be able to obtain. So it is natural that we want to 'mine out' the most possible information from the best corpora. There are several ways how to do this - for example, class based language models assume that words form clusters in high dimensional space, and try to exploit this fact. Structural language models try to capture long distance relations between words. For more information about advanced language modeling, see [4].
For language modeling of Czech, we have chosen to use neural networks. There are several reasons - neural networks are able to work similarly as class based language models, but it was assumed that their possibilities are even greater, since they are able to decompose words into 'features'. In other words, they are able to discover relations between word endings in Czech. Moreover, they should be able to model better long context relations, because they don't use the

word history in the same way as standard back-off models, which are unable to assign properly probability to unseen n-gram (although it is solved by backing off to n-1 gram, this approach is supposed to be insufficient). For more details about neural networks for language modeling, see [1][3][6].
Our baseline data for experiments with neural networks include annotations of lectures, since these data are the most valuable and we are primarily interested in modeling them better. All data that have already proven to be useful (PBSC, FI MU, FIT BUT) are also included. The general corpus is not included, since it's size is too big for our experiments. Besides, general corpus was found to be not much useful, and after discarding it, the decrease in accuracy is less than 1%. For better modeling of spontaneous speech, new corpora containing some general semispontaneous speech was also included.

	Accuracy
Trigram baseline	70.9%
Trigram + NN	73.6%
AR trigram + NN	74.2%

Table 3: Accuracy with more training data & NN

As is seen in table 3, more accurate training data provide much better baseline results than in the previous experiments - we were able to improve results from 66% to almost 71% just by using additional data, while our setup is now very small - total size of the training data is 7.3 milion words, which is much less than with general corpus. On the other hand, since the included lecture annotations are close to the test data, we can expect slightly worse results for some test data from another subject.
The baseline OOV rate is 1.8% and perplexity 377. AR trigram in table 3 denotes automatically reweighted language model according to the first pass decoded data. In practice this means that the data recognized in the first pass are used as validation data, for which new LM is built, using interpolation weights that are optimal for the validation data. So this language model works as a foreground language model, while neural network, which is trained on unweighted data, works like background model.
Neural networks were used here only for N-best list rescoring (in these experiments, N=3000). It can be expected that including NNs into the decoder would provide another improvement. It was observed that NNs help mostly by fixing endings of words, thus smoothing the output. Note that although the results may still be improved by gathering new data, our baseline system is already quite good, having low OOV rate and perplexity.

5 CONCLUSION

This paper described improvements of Czech lecture ASR, with overall improvement more than 14% in accuracy absolute. Most of the improvement was achieved by gathering and using more training data (almost 11%), while neural network based language models together with automatically reweighted language model improve the results further by another 3%.
Future work may focus on improving both sources of improvements: gathering new data would be for sure very important for any recognizer. In cases where new data are too costly, neural networks may provide useful, so future work should focus on them as well. For example,

neural networks may be extended with morphological information about the word structure, as it is done in factored language models.

REFERENCES

[1] Mikolov, T.: Language Modeling for Speech Recognition in Czech, Master's thesis, Brno, FIT BUT, 2007

[2] Stolcke A.: SRILM - an extensible language modeling toolkit. International Conference on Speech and Language Processing, 2002

[3] Yoshua Bengio, Réjean Ducharme and Pascal Vincent. 2003. A neural probabilistic language model. Journal of Machine Learning Research

[4] Joshua Goodman, Eugene Charniak. 1999. The State of The Art in Language Modeling.

[5] Ircing Pavel. Large Vocabulary Continuous Speech Recognition of Highly Inflectional language (Czech). 2003. Ph.D thesis, University of West Bohemia in Pilsen Department of Cybernetics

[6] Holger Schwenk and Jean-Luc Gauvain. Training Neural Network Language Models On Very Large Corpora. Published in Joint Conference HLT/EMNLP, oct 2005