Morpheme Based Language Models for Speech Recognition of Czech ⋆

W. Byrne[1], J. Hajič[1,2], P. Ircing[3], P. Krbec[1,2], J. Psutka[3]

[1] Johns Hopkins University, Baltimore MD, USA
[2] Charles University, Prague, Czech Republic
[3] University of West Bohemia in Pilsen, Czech Republic

Abstract. In our paper we propose new technique for language modeling of highly inflectional languages such as Czech, Russian an other Slavic languages. Our aim is to alleviate main problem encountered in these languages, which is enormous vocabulary growth caused by great number of different word forms derived from one word (lemma). We reduced the size of the vocabulary by decomposing words into stems and endings and storing these subword units (morphemes) in the vocabulary separately. Then we trained morpheme based language model on the decomposed corpus. This paper reports perplexities, OOV rates and some speech recognition results obtained with new language model.

1 Introduction

In this paper we propose new approach to language modeling which would be suitable for highly inflectional languages such as Czech, Russian and other Slavic languages. When we try to build large vocabulary speech recognition system for these languages, we encounter one major problem - excessive vocabulary growth caused by great number of different word forms derived from one word (lemma). Since state-of-the-art speech recognition engines are capable of handling vocabularies whose size is somewhere in the range from 20K to 60K words, we have to restrict the vocabulary and therefore we obtain very high OOV (Out Of Vocabulary) rate.

Inflectional change of the word mostly affects word ending, whereas stem remains constant and moreover the number of different endings is relatively very small(approximately 700). So it seems to be a good idea to decompose words into stems and endings and to treat these units (let's call them morphemes) separately as if they were independent words.

Our paper presents both characteristics of such morpheme based language models (perplexities, OOV rates) and results of speech recognition experiments performed with these models.

⋆ This work was supported by the projects KONTAKT No. ME293, No. VS97159 and No. VS96151 of the Ministry of Education in Czech Republic and by NSF Grants No. IIS-9810517 and No. IIS-9820687.

2 Czech Speech and Text Corpora, Baseline System

2.1 CZBN Broadcast News Corpus

CZech language Broadcast News (CZBN) corpus was collected for acoustic training. Satellite transmissions of Voice of America broadcasts were recorded by the Linguistic Data Consortium (LDC) and transcribed at the University of West Bohemia according to protocols recommended by LDC for use in Broadcast News LVR evaluation. The recordings span the period February 8 through May 31, 1999. The corpus consists of 62 recordings of 30 minute news broadcasts yielding in total over 20 hours of broadcast material. Portions of the shows containing music, speech in music, or other non-speech material are marked, but speech found during those periods is not transcribed. Therefore corpus yields approximately 19:30 minutes of transcribed material from each 30 minute broadcast.

2.2 Text Corpus

For language modeling purposes we collected texts from newspapers Lidove Noviny [5] spanning the period 1991 through 1995. Corpus contains approximately 33 million tokens (650K distinct words). Text data had to go through several preprocessing stages in order to obtain clear and unambiguous data.

2.3 Baseline System

Acoustic training and speech recognition experiments were conducted on the CZBN corpus using HTK, the hidden Markov model toolkit [6]. The recognition system is based on a continuous density HMMs trained on approximately 18 hours of speech. The speech features parameterization employed in training and test are mel-frequency cepstra, including both delta and delta-delta sub-features; cepstral mean subtraction is applied to all features on a per utterance basis. Triphone state clustering was carried out using broad acoustic phonetic classes similar to those used for English.

Baseline bigram language model was trained on Lidove Noviny corpus using CMU-Cambridge Statistical Modeling Toolkit [3]. As the vocabulary for this language model we used 20K most frequent words from the corpus. Witten Bell discounting was applied. Baseline results are shown in the table 1.

Vocabulary size	OOV-Rate on the test set	Test set perplexity	Word Accuracy
20 036	15.18%	584.34	47.89%

Table 1. Baseline results.

3 Morpheme Based Language Models

3.1 Word Decomposition

As we already mention, major cause of enormous vocabulary growth is the highly inflected nature of Czech language. Therefore we decomposed text corpus and 20K language model vocabulary into stems and endings (morphemes) using Czech Morphological Analyzer [4]. Stems and endings were marked with different marks (stems with '+' sign and endings with '−' sign) in order to allow them to be conjoined back after the recognition. Endings of zero length were discarded since they are not necessary for reconstruction of the original words. Examples of the decomposition are given in table 2.

Word	Decomposition	Translation
biskup	biskup+	the bishop
biskupa	biskup+ −a	(for the) bishop
biskupovi	biskup+ −ovi	(about the) bishop
biskupem	biskup+ −em	(with the) bishop
jet	je+ −t	to go
jedu	jed+ −u	I go
jedeš	jed+ −eš	you go (singular)
jede	jed+ −e	he goes
jedeme	jed+ −eme	we go
jedete	jed+ −ete	you go (plural)
jedou	jed+ −ou	they go

Table 2. Example of decomposition into stem and ending.

We obtained quite interesting results. As can be seen in table 3, size of the vocabulary was reduced by 52%. Moreover endings observed in 20K cover endings from the test set almost completely. Even the word OOV-rate dropped by 19%, because now some words which were originally unknown (and therefore constituted an OOV) can be composed of known stem and known ending.

	Vocabulary size	OOV-Rate on the test set
Word vocabulary	20 036	15.18%
Morpheme vocabulary	9 621	12.46%
– stems	9 258	12.32%
– endings	363	0.20%

Table 3. Comparison of the word vocabulary and the morpheme vocabulary.

3.2 Basic Morpheme Based Language Model

In our first morpheme based language model morphemes were treated as if they were independent words. No distinction between stems and endings was made. Bigram language model was trained on decomposed corpus using decomposed vocabulary. Witten Bell discounting was applied as in the word language model training. Results for this language model are shown in table 4.

Test set perplexity of this language model is much lower than the perplexity of the baseline model, but there is a catch in this. One has to bear in mind that the perplexity depends on the number of tokens in the test set in the sense that higher number of tokens makes the perplexity lower. In our basic morpheme based model morphemes were consider to be separate words and therefore number of tokens in the test set increased in comparison with word based model. So although in the morpheme based experiment the test set was generated by the language model with lower probability, perplexity quite considerably decreased. However, word accuracy is a fair measure of quality of the language model and we achieve excellent improvement with our new language model – over 37 %.

This result is in contradiction with the results obtained by other authors who tried to use word decomposition for language modeling of highly inflectional languages [2]. Their word accuracy decreased when they used morpheme based model. There can be several causes of such contradictory results. First of all, paper [2] deals with Serbo-Croatian. Although both Czech and Serbo-Croatian belong to the Slavic language family, they are of course not identical. Second reason could be that we used decomposition which better represents the nature of the language. And finally, improvement could be task-dependent, because CZBN corpus has some characteristics different from other Czech speech corpora. For example it contains high number of foreign words, since Voice of America news report mainly on international issues.

Vocabulary size	OOV-Rate on the test set	Test set perplexity	Word Accuracy
9621	12.46%	161.15	65.96 %

Table 4. Results for basic morpheme based language model.

Possible flaw of this morpheme based language model emerges if we look more closely at the probabilities assigned to the particular morphemes. Prediction of the i^{th} word ending e_i is based on the knowledge of the corresponding stem s_i, because language model assigns the probability $P(e_i|s_i)$ to the ending e_i. Such dependency is strong, because particular stem can be followed by relatively small set of endings.

On the other hand prediction of the stem s_i depends on the preceding ending e_{i-1}, probability assigned to such stem is $P(s_i|e_{i-1})$. In this case the dependency is very weak, because the ending bears information about grammatical properties of the word (such as case of the noun, person of the verb etc.), not information about the word itself. So bigrams are in fact reduced to unigrams. Therefore our morpheme based language model should be modified.

3.3 Modified Morpheme Based Language Model

As we argued in the previous section, prediction of the stem s_i shouldn't be based on the knowledge of the preceding ending e_{i-1}. Common sense hints us to use the dependency on the previous stem s_{i-1}, i.e. the language model should assign the probability $P(s_i|s_{i-1})$ to stem s_i. Since the word stem gives us the major part of the information about the word (in most cases almost the same information as the lemma), quality of such dependency should be comparable to word bigrams.

Prediction of the ending is more complicated. Firstly, ending e_i should depend on the corresponding stem s_i (see previous section). In addition, the Czech language makes extensive use of agreement. It means that for example a noun and its adjectival or pronominal attribute must agree in gender, number and case. Since these morphological categories often affect word ending, prediction of the ending e_i should also be based on preceding ending e_{i-1}.

Let's now write down our modified language model more formally. Model assigns the probability:

$$P(s_i|s_{i-1})$$

to the stem s_i and the probability:

$$\xi \cdot P(e_i|s_i) + (1 - \xi) \cdot P(e_i|e_{i-1})$$

to the ending e_i, where ξ is the parameter in the range $0 \leq \xi \leq 1$.

Text corpus had to be modified before training of this new morpheme based language model. Now we have to incorporate also endings of zero length, into the corpus as well as into the language model vocabulary. Perplexity results for various ξ are given in table 5. Low values of the perplexity are due to the fact that number of tokens in the test set further increased because of endings of zero length, which were not present in the previous language model. However, it should be mentioned that modified morpheme based language model with the parameter $\xi = 0.0$ generated the test set with higher probability than basic morpheme based model and we could consider this model to be the best of all morpheme based models.

As can be seen in the table 5, model with the parameter $\xi = 0.0$ has the lowest perplexity. So the knowledge of the preceding ending e_{i-1} probably doesn't give us much information about the ending e_i. Looking into the test set, one discovers a possible source of this problem. Endings of zero length constitute almost a half of all endings (48%) in the test set and such percentage holds for training corpus as well. Therefore one could expect that endings of zero length are followed by other endings more or less with the same frequency. So there is no chance to get strong dependencies between the endings in the language model.

ξ	Test set perplexity
0.0	40.13
0.25	48.07
0.5	57.59
0.75	68.98
1.0	82.64

Table 5. Comparison of the test set perplexity for various ξ.

In order to obtain fair comparison between basic and modified morpheme based language model, speech recognition experiments have to be performed. That would involve some kind of lattice rescoring and HTK toolkit isn't very well suited to handle such large lattices. However, results of the experiment performed on the task with 600 word vocabulary were promising.

4 Conclusion

In this paper we introduced new language models based on morphemes. Using these models we manage to significantly reduce the size of the language model vocabulary, OOV rate and even to greatly improve word accuracy. In our future work we should test our techniques on large test sets from various domains to prove their robustness. Furthermore we have to perform speech recognition experiments using modified morpheme based language model described in section 3.3.

References

1. W. Byrne et al.: Large Vocabulary Speech Recognition for Read and Broadcast Czech. TSD. 1999
2. P. Geutner, M. Finke, P. Scheytt: Adaptive Vocabularies for Transcribing Multilingual Broadcast News. ICASSP. 1998.
3. P. Clarkson, R. Rosenfeld: The CMU-Cambridge Statistical Language Modeling Toolkit
4. J. Hajic: Computational Morphology of Czech. Karolinum Press, Charles University. Prague. In prep.
5. `http://www.lidovenoviny.cz`
6. S. Young et al.: The HTK Book. Entropic Inc. 1999