On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech

W. Byrne\(^1\), J. Hajic\(^2\), P. Ircing\(^3\), F. Jelinek\(^1\)

1 Johns Hopkins University
Baltimore, MD 21218
USA
{byrne,sanjeev}@clsp.jhu.edu

2 Institute of Formal and Applied Linguistics
MFF UK
Malostranske n. 25, 118 00 Praha 1 Czech Republic
{hajic,krbec}@ufal.mff.cuni.cz

3 University of West Bohemia in Pilsen
Univerzitni 8, 306 14 Plzeň
Czech Republic
{ircing,psutka}@kky.zcu.cz

Abstract

A system for large vocabulary continuous speech recognition of highly inflectional language is introduced. Word-based recognition approach is compared with a morpheme-based recognition system. An experiment involving Czech N-best rescoring has been performed with encouraging results.

1. Introduction

In our paper we deal with the difficulties encountered in large vocabulary continuous speech recognition of highly inflectional languages such as Czech, Russian and other Slavic languages. Major problem encountered in these languages is excessive vocabulary growth caused by huge number of different word forms derived from one basic word form (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.

This paper compares results of speech recognition experiments performed with word-based and morpheme-based language models.

2. Speech and text corpora

2.1. Czech TV & Radio Broadcast News speech corpus

This speech corpus consists of news broadcasted on 3 TV channels and 4 radio stations during the period February 1, 2000 through April 22, 2000. Corpus contains over 50 hours of audio data stored on 347 files, which yield about 26 hours of pure transcribed speech.

The following TV channels and radio stations were recorded:

- **TV**: ČT1 (22 files), NOVA (30), PRIMA (23)
- **Radio**: RADIOŽURNAL (138 files), PRAHA (57), VLTAVA (14), FREKVENCE1 (30)

The broadcast news corpus does not contain weather forecasts, sports news and traffic announcements. The signal is single-channel, sampled at 22.05 kHz with 16-bit resolution.

22 hours of the transcribed material were used for acoustic model training and 4 hours were put aside as the test data and data designated for finding the optimum scaling factors.

2.2. Lidové Noviny 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.

3. Word-based recognition

3.1. Baseline system

Acoustic models were trained using HTK, the hidden Markov model toolkit [7]. The recognition system is based on a continuous density HMMs trained on approximately 22 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. Pronunciation dictionary contains 34 511 words.

Standard word bigram language model with Katz discounting was estimated utilizing the SRILM toolkit [8]. As the vocabulary for this language model we used 60K most frequent words from the Lidové Noviny corpus. Then we generated word lattices for the test utterances as well as for the utterances, which we used later for finding the optimum scaling factors (see the section 3.2). AT&T decoder [9] was employed for this purpose.

As you can see in the table above, word accuracy is not so great. It is partially due to the high OOV rate, but also due to the limited prediction power of the bigram language model. Therefore we created N-best lists (N=5000) from all generated lattice and rescored them with a better trigram language model.
We also computed the oracle word accuracy, i.e. the best accuracy that can be achieved by rescoring the given N-best lists. In our case the oracle word accuracy is 81.21%.

3.2. Trigram rescoring, optimization of scaling factors

We have used the technique of N-Best list rescoring. The language model which we used is an interpolated word forms trigram LM with all trigrams (even singletons) included. The interpolation coefficients are grouped into buckets based on histories.

The total score of a sentence in a N-Best list is computed as:

\[ \text{Sentence}_\text{score} = A M_{\text{score}} + L M_{\text{scaling}} \cdot L M_{\text{score}} - IP_{\text{scaling}} \cdot \text{Sentence}_\text{length} \]

We have obtained the optimum scaling factors from the Held-out data as shown on Figure 1.

![Figure 1: Held-out data accuracy depending on scaling factors. Maximum point - \(LM_{\text{scaling}} = 24, IP_{\text{scaling}} = 13.5\)](image)

Table 2: Results of word-based recognition

<table>
<thead>
<tr>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>Rescored</td>
</tr>
</tbody>
</table>

Table 3: Examples of the word decomposition

<table>
<thead>
<tr>
<th>Word</th>
<th>Decomposition</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>tráva</td>
<td>tráv+ -a</td>
<td>the grass</td>
</tr>
<tr>
<td>trávy</td>
<td>tráv+ -y</td>
<td>(without the) grass</td>
</tr>
<tr>
<td>trávě</td>
<td>tráv+ -ě</td>
<td>(about the) grass</td>
</tr>
<tr>
<td>trávu</td>
<td>tráv+ -u</td>
<td>(for the) grass</td>
</tr>
<tr>
<td>trávou</td>
<td>tráv+ -ou</td>
<td>(with the) grass</td>
</tr>
<tr>
<td>pít</td>
<td>pí+ -t</td>
<td>to drink</td>
</tr>
<tr>
<td>píj</td>
<td>píj+ -u</td>
<td>I drink</td>
</tr>
<tr>
<td>píjě</td>
<td>píj+ -ěš</td>
<td>you drink (singular)</td>
</tr>
<tr>
<td>píje</td>
<td>píj+ -e</td>
<td>he drinks</td>
</tr>
<tr>
<td>píjeme</td>
<td>píj+ -eme</td>
<td>we drink</td>
</tr>
<tr>
<td>píjete</td>
<td>píj+ -ete</td>
<td>you drink (plural)</td>
</tr>
<tr>
<td>píjou</td>
<td>píj+ -ou</td>
<td>they drink</td>
</tr>
</tbody>
</table>

Table 4: Comparison of the word vocabulary and the morpheme vocabulary

<table>
<thead>
<tr>
<th>Vocabulary size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word vocabulary</td>
</tr>
<tr>
<td>Morpheme vocabulary</td>
</tr>
<tr>
<td>– stems</td>
</tr>
<tr>
<td>– endings</td>
</tr>
</tbody>
</table>

4. Morpheme-based recognition

4.1. Word decomposition

As we already mentioned, major cause of enormous vocabulary growth is the highly inflected nature of Czech language. In order to demonstrate that this is really correct we decomposed our baseline 60k vocabulary into stems and endings (morphemes) using Czech Morphological Analyzer [6]. Examples of the decomposition are given in table 3.

Table 5: Baseline results of morpheme-based recognition

<table>
<thead>
<tr>
<th>Vocabulary size</th>
<th>OOV Rate</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>62 740</td>
<td>4.62%</td>
<td>63.14%</td>
</tr>
</tbody>
</table>

Table 4 shows the effect of the word decomposition on the vocabulary size, which was reduced by 58%. You can also see that word forms in the 60k vocabulary are created using quite small set of endings.

Even though the OOV rate substantially decreased, recognition results are slightly worse. The problem is in the qual-
ity of the our morpheme-based language model. If we look more closely at the probabilities assigned to the particular morphemes, we can see the possible flaw. 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 the 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 we generated N-best lists (again N=5000) and rescored them using more sophisticated language model. The oracle word accuracy of morpheme-based N-best lists is 82.33%.

4.3. Trigram rescoring, optimization of scaling factors

The morpheme-based language model has the following form: Prediction of the $i^{th}$ stem is based on the knowledge of the two preceding stems. 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 form trigrams, possibly even better, because we overcome the data sparseness problem.

Prediction of the ending is more complicated. Firstly, ending $e_i$ should depend on the corresponding stem $s_i$. 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 $e_{i-1}$.

$$
\frac{P(s_i \mid s_{i-1}, e_{i-2})}{P(e_i \mid s_i, e_{i-1})}
$$

Both trigram models use the interpolation coefficients which are grouped into the buckets based on histories again. The same N-Best list rescoring technique is performed as with the wordforms. We have also computed a new set of the optimum scaling factors.

<table>
<thead>
<tr>
<th></th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>63.14%</td>
</tr>
<tr>
<td>Rescored</td>
<td>70.38%</td>
</tr>
</tbody>
</table>

Table 6: Results of morpheme-based recognition

5. Conclusions and future work

Morpheme-based recognition performs nearly as good as the word based recognition. We believe that a more sophisticated morpheme-based language model can be implemented in the future. In any case, we consider the N-best rescoring approach combined with a Morpheme-based recognition a correct way how to handle the high WER for the inflectional languages such as Czech.

6. Acknowledgements

This work was supported by the projects KONTAKT No. ME293 and No. F917 of the Ministry of Education in Czech Republic.

7. References
