Exploiting Linguistic Knowledge in Language Modeling of Czech Spontaneous Speech

Pavel Ircing, Jan Hoidekr, Josef Psutka

Department of Cybernetics
University of West Bohemia, Plzeň, Czech Rep.
{ircing, hoidekr, psutka}@kky.zcu.cz

Abstract

In our paper, we present a method for incorporating available linguistic information into a statistical language model that is used in ASR system for transcribing spontaneous speech. We employ the class-based language model paradigm and use the morphological tags as the basis for world-to-class mapping. Since the number of different tags is at least by one order of magnitude lower than the number of words even in the tasks with moderately-sized vocabularies, the tag-based model can be rather robustly estimated using even the relatively small text corpora. Unfortunately, this robustness goes hand in hand with restricted predictive ability of the class-based model. Hence we apply the two-pass recognition strategy, where the first pass is performed with the standard word-based n-gram and the resulting lattices are rescored in the second pass using the aforementioned class-based model. Using this decoding scenario, we have managed to moderately improve the word error rate in the performed ASR experiments.

1. Introduction

It is widely known that the language modeling of spontaneous speech is an extremely challenging task due to the frequent occurrence of speech disfluencies, ungrammatical sentences and to the large extent also due to the lack of appropriate in-domain training data. The problem of scarce training data is even more serious for highly inflectional languages (such as the language in question - Czech), since such languages produce many words forms for a single lemma and thus the vocabulary and consequently also the number of language model parameters grow very rapidly.

The mentioned training data insufficiency of course negatively affects the robustness of the estimated word-based language model. Using additional language modeling data sources seems to be a natural way to overcome the data sparsity, however, our experience shows that finding such new data and combining them with the existing in-domain resources is a nontrivial task with uncertain results (Psutka et al., 2003). Thus we propose a technique that makes use of the linguistic information present in the in-domain data that we have at our disposal.[1]

2. The Task at Hand

Our task is to automatically transcribe the speech extracted from the Czech portion of the large video archives processed within the project MALACH (Byrne et al., 2004). Those archives consist of the testimonies given by the Holocaust survivors, therefore the speech contained in the recordings poses extreme challenges to the ASR system - the speakers are usually elderly, their speech is spontaneous, often heavily accented and emotional.

From the lexical point of view, the survivors often talk about the people and places that are hardly mentioned outside the archives (and it is therefore difficult to estimate their language model probabilities correctly). Moreover, in Czech part of the archives we face the problem of frequent usage of colloquial words - this issue and the ways of its partial alleviation have been discussed in (Psutka et al., 2004).

The entire archive contains almost 52 000 interviews in 32 languages, a total of 116 000 hours of audio and video. The Czech portion consists of 346 interviews that we divided into 336 training speakers and 10 test speakers. The 15-minute segment was transcribed from each of the training speakers, yielding a total of 84 hours of annotated speech. The testimonies of the test speakers were transcribed completely, yielding approximately 23 hours of transcribed speech. However, we randomly selected 500 sentences from the test set for the ASR test purposes - all the results reported in the paper were achieved on this 500-sentence set (we will call it the "ASR test set" in the rest of the paper).

The transcripts of the acoustic training data contain 606 thousands (606k) tokens and the vocabulary extracted from this text contains 45k different words. The OOV rate on the ASR test set is 5.21%. It can be seen that this number is substantially larger than the OOV rates usually observed in the similar tasks for English.

3. Available Linguistic Information

Fortunately, the automatic linguistic processing of the Czech language has been broadly investigated in the recent years and thus we can make use of several tools developed within these research projects. We have decided to use the system built by the team led by Jan Hajič (2004). The system encompasses two major components. First of them is the morphological analyzer, which generates all the possible lemmas and tags for a given word form, regardless of the context. The second component is the part-of-speech tagger that performs the lemma and tag disambiguation, taking into account the context of the word form in question.

[1] Note that these data usually consists of the manual transcripts of the speech used for the acoustic modeling - therefore the acquisition of the additional text in this manner is at least very costly, if not altogether prohibited due to the usually limited speech data resources.

Every tag is represented as a string of 15 symbols. Each position in the string corresponds to one morphological category in the following order - part of speech, detailed part of speech, gender, number, case, possessor's gender, possessor's number, person, tense, degree of comparison, negation and voice. Positions 13 and 14 are currently unused and finally position 15 is used for various special purposes (such as marking colloquial and archaic words or abbreviations). Non-applicable values are denoted by a single hyphen (-).

For example, the tag `VB-S---3P-AA---` denotes the verb (V) in either the present or the future tense (B), singular (S), in the third person (3), in the present tense (P), affirmative (A) and in the active voice (A).

The usage of the morphological tags in language modeling is motivated by the fact that the Czech language makes an extensive use of agreement. The strongest agreement is between a noun and its adjectival or pronominal attribute: they must agree in gender, number and case. There is also agreement between a subject (expressed by a noun, a pronoun or even an adjective) and its predicate verb in gender and number, and for pronouns, also in person. Verbal attributes must agree in number and gender with its related noun, as well as with its predicate verb (double agreement). Possessive pronouns exhibit the most complicated type of agreement - in addition to the abovementioned triple attributive agreement with the possessed thing they must also agree in gender and number with the possessor.

All those morphological categories (gender, number, etc.) are included in the morphological tags. Therefore there should exist some dependencies between adjacent tags that can be captured even by a relatively simple n-gram language model.

4. Proposed Language Model

Since we want to exploit the dependencies existing between morphological tags and at the same time we still need to have words in the output of the ASR system, the usage of the class-based language model seems to be a natural choice. Note that in this particular case a word-to-class mapping is defined by the morphological tag of the word in question - therefore one word can obviously fall into multiple classes. Thus we need to utilize the class-based model where the probability of the word w_i given the history h_i is defined by

$$P(w_i|h_i) = \sum_{c_i} P(w_i|c_i).P(c_i|c_{i-n+1}, \ldots, c_{i-1}) \quad (1)$$

where $P(w_i|c_i)$ denotes the probability of the word w_i given the class c_i and $P(c_i|c_{i-n+1}, \ldots, c_{i-1})$ denotes the n-gram probability that the class c_i will follow the previous $(n-1)$ classes $c_{i-n+1}, \ldots, c_{i-1}$.

The incorporation of the model (1) into the ASR decoder is slightly challenging. In our experiments, we employ the decoder developed by AT&T. This decoder is based on the weighted finite-state transducers (WFST) (Mohri et al., 2002) and can be represented by the so-called recognition cascade depicted in Figure 1.

Each component of the cascade is a weighted finite-state transducer - H represents an acoustic model, C transduces

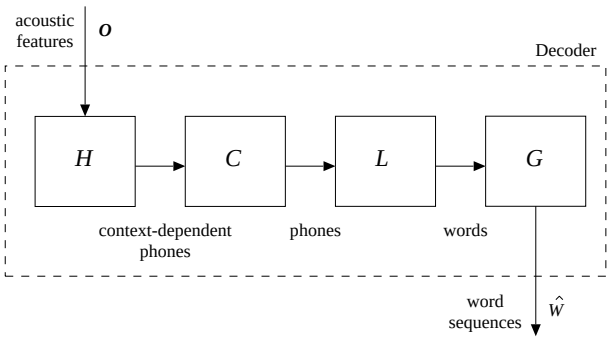

Figure 1: Recognition cascade

context-dependent phones to context-independent ones, L represents a pronunciation lexicon and finally G is a (word-based) n-gram language model. The decoder task of finding the best word sequence $\hat{W}$ given the observed acoustic sequence $\boldsymbol{O}$ can be then reduced to finding the best path in the composite transducer

$$O \circ H \circ C \circ L \circ G$$

where O is a transducer originating from the transformation of an input acoustic feature sequence $\boldsymbol{O}$ to a trivial finite-state machine. The $C \circ L \circ G$ part of the cascade is constructed beforehand whereas the composition with O and H is performed during the decoder run.

It was shown in (Ircing and Psutka, 2003) that the model (1) can be represented within the WFST framework as a composition of two transducers $T \circ V$ where T realizes a mapping from words to classes with the appropriate probability $P(w_i|c_i)$ and V is the class variant of the n-gram transducer G (it assigns the probability $P(c_i|c_{i-n+1}, \ldots, c_{i-1})$ to the sequences of classes). Thus we can simply replace the word-based language model G with $T \circ V$ in the recognition cascade. In that case we obtain the best class sequence instead of the best word sequence in the output of the decoder. However, the AT&T decoder is built so that it produces not only the best sequence but also a so-called lattice. The lattice is an acyclic finite-state transducer containing the most probable paths through the recognition cascade for a given utterance. It has context-dependent phones on the input side and output labels from the rightmost transducer in the cascade on the output side. Therefore we can retrieve the best word sequence $\hat{W}$ even from the class-based lattice using a set of trivial WFST operations. However, the transducer $C \circ L \circ T \circ V$ often becomes too large due to the many-to-many word-to-class mapping. Thus the first recognition run is usually performed with a simple word-based n-gram G. Then the language model score is stripped from the output lattices and the resulting lattices are rescored with $T \circ V$.

On the other hand, it is generally known (and our experiments proved it) that class-based language models yield more robust probability estimates than word-based models but at the same time they have worse discrimination ability ("sense of detail"). Thus word-based and class-based language models are usually combined in some manner. The WFST framework offers a natural way of model combina-

tion - we can simply retain the word language model score in the output lattices and then compose the lattices with $T \circ V$. We have found out empirically that better results are achieved when the $P(w_i|c_i)$ component is omitted from the class-based model. This final language model setting then corresponds to the usage of the word probability computation according to the following formula

$$P(w_i|h_i) = P(w_i|w_{i-l+1}, \ldots, w_{i-1}) \cdot \cdot \left\{ \max_{c_i} P(c_i|c_{i-n+1}, \ldots, c_{i-1}) \right\} \quad (2)$$

Note that the change from the sum of probabilities in Formula (1) to the max operator on Formula (2) is related to the use of Viterbi approximation in the AT&T decoder framework. The detailed explanation has to be omitted due to the space limitations and can be found in (Ircing and Psutka, 2003).

5. Experimental Evaluation

5.1. Acoustic models

As was already mentioned in Section 2., the acoustic models were trained using 84 hours of transcribed speech. The data was parameterized as 17-dimensional PLP cepstral features including their delta and delta-delta derivatives (resulting into 51-dimensional feature vector). These features were computed at a rate of 100 frames per second. Cepstral mean subtraction was applied on a per-utterance basis. The resulting cross-word-triphone-based models were trained using the HTK toolkit (Young et al., 2000) and had approximately 6k states and 107k Gaussians.

5.2. Language modeling experiments

All n-gram language models described in this section are bigrams with Katz's backing-off scheme, estimated using the SRILM toolkit (Stolcke, 2002).

We have decided not to use the original form of the transcripts as a basis for our language modeling experiments, since, as was already briefly mentioned in Section 2., there is a frequent occurrence of colloquial word forms in Czech spontaneous speech. The Czech colloquial words are not regarded as simply pronunciation variants, as they have well-defined spelling that differs from the corresponding standard word forms. This results into the presence of many spelling variants of a single standard word in the vocabulary, which obviously leads to further fragmentation of the already sparse training data.

Thus we utilized the "standardized" version of the text resources that was created by manually appending additional column with the spelling of the corresponding standard variant to the pronunciation lexicon. Such approach allows us to automatically "standardize" the transcripts and consequently use the resulting parallel corpora (original and standardized) for counting of the relative frequencies of the individual colloquial variants. Here we present an excerpt from the standardized lexicon (columns contain standard form, colloquial form, phonetic baseform of the colloquial form and the relative frequency of the colloquial form, respectively):

```
ODJET   ODEJET   o d e j e t   0.0161
ODJET   ODJEC    o d j e c     0.0161
ODJET   ODJECT   o d j e c t   0.0483
ODJET   ODJET    o d j e t     0.2741
ODJET   VODJET   v o d j e t   0.0967
```

The column with standard word forms then constitutes the vocabulary of the word-based language model (which is of course estimated using the standardized transcripts) and of the decoder. The standardized transcripts also serve as the input for the morphological analysis and tagging. Basic properties of the vocabularies corresponding to the (standardized) word and tag training corpora are given in Table 1 (The number of tokens is of course identical for both corpora (606k) and the OOV rate is evaluated on the standardized transcripts of the ASR test set).

	Vocabulary size	OOV rate
Words	41,249	5.07%
Tags	1,180	0.11%

Table 1: Basic properties of the word and tag vocabulary

Colloquial word forms are treated as "pronunciation variants" of the standard forms during the decoding. Two different types of the pronunciation lexicon were employed in our experiments - one of them considers all the "pronunciation variants" to equally likely (unweighted lexicon -**UL**), the other weights them according to the aforementioned relative frequencies (weighted lexicon - **WL**).

The recognition scenario was the same for both lexicons. First, we estimated the word-based and tag-based bigram language models. Then we put the word-based model into the recognition cascade, run the decoder, generated the word lattices and evaluated the baseline word error rate (**BASE**). The lattices were then rescored by composition with the tag-based model according to Formula (2) - that is, both the word-based and the tag-based model contributed to the final language model score (**RESC**). The results are summarized in Table 2.

	BASE	**RESC**
UL	43.38%	40.92%
WL	41.16%	40.00%

Table 2: Recognition results - word error rates (WER)

The outcomes of the experiment illustrate that both the weighting of the pronunciation variants and the lattice rescoring by tag-based language model brought a moderate improvement of the recognition performance. Moreover, this improvement seems to be additive since the best result was achieved by the combination of both approaches.

The number of "classes" in the tag-based language model (1,180) seems to be rather high from the class-based language modeling perspective. Thus we have experimented with reducing this number using both the knowledge-based approach (taking into account only a subset of the tag positions) and the data-driven clustering (Brown et al., 1992).

The experiments with the various meaningful tag shortenings, while substantially reducing the number of classes, did not cause any noticeable change in the WER. On the other hand, the data driven clustering caused a significant degradation of the recognition performance. Thus we conclude that, first, the existing tagset is well-suited to the language modeling (although it was originally designed for different purposes) and second, the amount of data that we have at our disposal is already sufficient to train a robust tag-based model even with such a relatively high number of classes. The later claim is supported by the analysis of tag unigram frequencies when we found out that they are only 14% of singletons in the tag vocabulary.

Originally, we have assumed that the positive effect of the tag-based language model will increase with the degree of data sparsity. However, this assumption has not proved to be true. A possible reason is illustrated in Figure 2. As can be seen from the graph, both WER curves follow the shape of the OOV curve quite closely; therefore the increasing WER in the systems trained using a small amount of data is caused chiefly by the frequent occurrence of the OOV words and thus even more robustly trained tag-based models are unable to help.

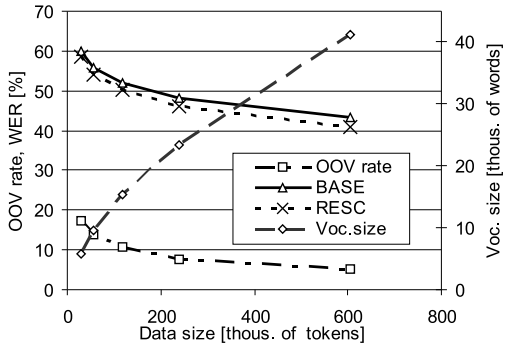

Figure 2: Effect of training data size

We have also planned to employ the tag-based language models trained using large out-of-domain data. However, it turned out that they did not improve the recognition performance. Most probably the discrepancy between spontaneous speech transcripts and written texts lies deeper than in the different vocabularies - we suspect that they are also substantial differences in word ordering of the sentences. Such conclusion is inspired by the perplexity values of the two tag-based language models showed in Table 3.

	Train	Test
Trans	34.8	38.7
Written	36.8	69.0

Table 3: Training and test set perplexities of different tag models

The first model (Trans) was trained using the speech transcripts and the perplexity was computed on this training data (Train) and on the ASR test set (Test). The second model (Written) was train using a large (33 million tokens) newspaper data and again tested on the train data itself and the ASR test set. As can be seen form the table, while the training data perplexity is roughly the same for both models, the in-domain tag model is clearly much better suited for modeling the tag sequences occuring in the test set.

6. Conclusion

We have managed to achieve a moderate reduction of WER by employing morphological information in the language model construction. We are still restricting ourselves to the relatively simple n-gram language models in the presented experiments; however, a more sophisticated language model structure could bring further improvement.

7. Acknowledgments

This work was supported by the Ministry of Education of the Czech Republic project No. 1P05ME786.

8. References

Peter Brown, Vincent Della Pietra, Peter deSouza, Jenifer Lai, and Robert Mercer. 1992. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajič, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. 2004. Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435.

Jan Hajič. 2004. *Disambiguation of Rich Inflection. (Computational Morphology of Czech)*. Karolinum, Prague.

Pavel Ircing and Josef Psutka. 2003. Fitting Class-Based Language Models into Weighted Finite-State Transducer Framework. In *Proceedings of Eurospeech 2003*, pages 1873–1876, Geneva, Switzerland.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted Finite-State Transducers in Speech Recognition. *Computer Speech and Language*, 16(1):69–88.

Josef Psutka, Pavel Ircing, Josef V. Psutka, Vlasta Radová, William Byrne, Jan Hajič, Jiří Mírovský, and Samuel Gustman. 2003. Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. In *Proceedings of Eurospeech 2003*, pages 1821–1824, Geneva, Switzerland.

Josef Psutka, Pavel Ircing, Jan Hajič, Vlasta Radová, Josef V. Psutka, William Byrne, and Samuel Gustman. 2004. Issues in annotation of the Czech spontaneous speech corpus in the MALACH project. In *Proceedings of LREC 2004*, pages 607–610, Lisbon, Portugal.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of ICSLP 2002*, Denver.

Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. 2000. *The HTK Book*. Entropic, Cambridge.