12

Recent Progress in Development of Language Model for Slovak Large Vocabulary Continuous Speech Recognition

Jozef Juhár, Ján Staš and Daniel Hládek
Technical University of Košice
Slovakia

1. Introduction

Speech technologies have a potentiality to simplify the human-machine interaction as well as the communication between people. The use of speech technology applications has nowadays continuously growing trend. Each speech recognition system, which stands in the heart of every speech application, besides an algorithmic complexity, is strongly language dependent. Therefore, one of the challenging tasks by the development of the *Slovak large vocabulary continuous speech recognition* (LVCSR) system is a creation of an efficient language model (LM).

Development of the Slovak language model, which belongs to a group of highly inflective languages, is more laboured than creation of an English language model. First reason is that the Slovak language is characterized by a relative free order of words in sentences. This consequently leads to the problem of *data sparseness of the text data* used for training of language models (LMs). Second reason is the *inflection in the language* itself due to the rich morphology which leads to a several times larger vocabulary than in English. Therefore, amount of text data that could statistically enough cover the Slovak language is substantially higher.

Contemporary modeling of the Slovak language is based on the knowledge of modeling of the related Slavic languages, such as Czech, Polish, Serbo-Croatian or Russian language (Nouza et al., 2010). From the field of statistics, Slovak language is very similar to the Czech language, especially in forming words into sentences and determining the sentence semantics. In the contrast, from linguistic point of view, mainly in phenomena of inflection and assimilation in voice, Slovak is more or less similar to the Polish. Therefore, for statistical language modeling it is appropriate to be limited with linguistic constraints as well.

This chapter describes results of the Slovak language model development for judiciary domain-specific LVCSR task and broadcast news transcription. During this process, we have coped with several problems in text preprocessing, selection of the basic statistical methods used in modeling of the other similar languages and adaptation into the area of application. Another important part in the Slovak language modeling has been optimization of the resultant model, which introduced phonetic and linguistic relations between words. These optimization steps have caused an improvement in quality of our LM as well as recognition accuracy of the LVCSR system itself.

This chapter is organized as follows. Section 2 introduces the process of text gathering and preprocessing text corpora used in training LMs. Section 3 describes the process of creation a vocabulary of the Slovak language. In the Section 4 the selection of appropriate smoothing technique, method for the adaptation to the given domain and optimal pruning algorithm are presented. Some proposed optimization approaches in modeling of the Slovak language are summarized in Section 5. Section 6 presents the setup of the Slovak LVCSR system used in real task of domain-oriented speech recognition. At the end of this chapter in Section 7, the experimental results are summarized. Section 8 closes this chapter with the discussion.

2. Text data and preprocessing

Small languages of Eastern Europe, such as the Slovak language, can be considered as under-resourced, because they usually suffer from the lack of audio databases and linguistic resources. Then, the main assumption in the process of creation an effective LM for any language is to collect and consistently process a large amount of text data entering into the process of training LM. Therefore, we have proposed an automatic system, called *webAgent* (Hládek & Staš, 2010a), which retrieves text data from various web pages written in Slovak language. Moreover, the text gathering system is able to detect the character encoding of the given web page, to collect links to other web pages and to retrieve text data from DOC (MS Word), RTF or PDF documents as well.

Before training LMs it has been necessary to transform the text data into pronunciation form. These text preprocessing steps include: (a) *word tokenization*, (b) *text normalization*, (c) *sentence segmentation* and (d) *filtering of grammatically incorrect sentences* (Hládek & Staš, 2010b).

The most important preprocessing operation is text normalization, for which the following rules has been proposed:

- each sentence is on exactly one line;
- all words were mapped to lowercase;
- all numerals (cardinal, ordinal, dates, mathematical items and others) were replaced by their pronunciation form according their surrounding context;
- compound words and numerals were divided to their separated form;
- selected frequent abbreviations, acronyms and names of titles were expanded to the pronunciation form according their surrounding context;
- numbered and alphabetical intends were transcribed to their pronunciation form;
- in judiciary documents hidden proper nouns and name entities, such as names, surnames, name of streets and cities were detected and replaced according their surrounding context using our proposed automatic generator of name entities;
- words with emphasized inter-character spaces were unified;
- all punctuation marks and symbols were replaced by their pronunciation form;
- spelled items were mapped to uppercase due to their better separation and their uniform phonetic transcription were determined;
- hypertext and email address were excluded from the text corpora.

When preprocessing the domain-specific text data from the field of judicature we had to resolve the problem of transcription of a large amount of specific abbreviations and numerals

text corpus		# sentences	# tokens
training data set	web corpus	54 765 873	946 958 508
	broadcast news	33 804 173	590 274 484
	judicial corpus	9 135 908	258 131 635
held-out data set	broadcast news	3 455 523	53 046 071
	judicial domain	1 782 333	55 163 941
annotations	broadcast news	124 733	925 912
	judicial domain	319 419	3 197 469
together		103 387 962	1 907 698 020

Table 1. Statistics of text corpora

as well (Staš et al., 2010b). Normalized documents are then stored in relational database based on PostgreSQL along with their titles, URIs of web pages, and names of sources where they were published. It should be noted that database is closely associated with the system for text gathering. In the process of insertion text data into database the duplicity verification is performed. Nowadays, we are dealing with text corpus of size about 1.9 billion of tokens in more than 100 million of sentences. The text corpus is divided into several different domain-related sub-corpora (see Table 1).

It should be noted that for filtering of grammatically incorrect words we have used our spellcheck lexicon, created by merging available Open Source dictionaries such as *aspell*, *hunspell* and *ispell* (sk-spell, 2010) with lists of proper nouns, geographical items and various name entities available on the Internet. The size of our lexicon for spell-checking is about 1.25 million of unique words (Staš et al., 2011a).

3. Vocabulary

Vocabulary which have been used in language modeling was selected from collected text corpora using standard methods based on the *highest occurrence* words in the training corpora and *maximum likelihood* approach (Venkataraman & Wang, 2003) for selection domain-specific words from the field of judicature. The vocabulary was then extended to the number of names and surnames, geographical items, names of various institutions and some other name entities in the Slovak Republic, as can be seen in the Table 2.

description		# words
348*k*	base vocabulary	348 255
names	female (inflected forms)	1 060
	male (inflected forms)	824
surnames	female (inflected forms)	55 774
	male (inflected forms)	82 388
name entities	geographical items	22 050
	names of institutions	2 331
	legal terms	2 548
	multiword expressions	3 000

Table 2. Vocabulary

We have also proposed an automatic tool for generating inflective word forms for names and surnames which were used in modeling of the Slovak language, (a) in the on-line dictation LVCSR system as an *independent model of names and surnames*, and later (b) in *modeling of names and surnames using word classes* conditioned by their grammatical category.

We have found that in modeling of the Slovak language with currently available text data the optimal results were achieved if the vocabulary size is about $100 - 150$ thousand of words for the domain-specific and about $300 - 350$ thousand of words in general domain task of speech recognition. It should be noted that all words in vocabulary were manually checked and corrected by linguistic experts.

4. Statistical modeling of the Slovak language

In the following sections, selected methods like smoothing, adaptation, combination and pruning are summarized. The most suitable algorithms were later used in training of the reference Slovak language model, as described in Section 6.1.

4.1 Language model

In general, language model determines the probability of the sequence of words as well as the word itself, what consequently helps the decoder to find the most probable sequence of words, which corresponds to the acoustic information pronounced by the user. Contemporary language modeling is based on the use of n-grams, which mainly consider the statistical dependency between n individual words.

Formally, the main aim of the n-gram model is to determine a priori probability $P(W)$ of a sequence of words $W = \{w_1 w_2 \dots w_{n-1}\}$ and to provide the quickest and the most exact estimation of this sequence of words in decoding process of a LVCSR system. This probability can be defined as follows

$$P(W) = P(w_1 w_2 \dots w_{n-1}) = \prod_{i=1}^{n} P(w_i | w_1 w_2 \dots w_{i-1}), \tag{1}$$

where $P(w_i|w_1 w_2 \dots w_{i-1})$ is the conditional probability of word w_i conditioned by its history $\{w_1 w_2 \dots w_{i-1}\}$. Such process of decomposition allows us to recognize for LVCSR system a sequence of words during its pronunciation and determines the probability $P(W)$ for searching strategy in decoding process gradually.

The main advantage of using n-gram models in LVCSR lies in relative easy computating their probability estimations based on computating the relative occurrence of words, or word sequences in the training data set using *maximum likelihood* approach (Jurafsky & Martin, 2009; Manning & Schütze, 1999).

4.2 Smoothing

As it was mentioned earlier, for dealing with problem of *data sparseness*, some re-estimation methods such as discounting, interpolation or backing-off also called smoothing are used in statistical language modeling (Jurafsky & Martin, 2009).

Due to the fact that a speaker can also pronounce a sentence does not occuring in the training data set, cause that the probability of such events can lead to the zero. Therefore, the problem of zero probabilities leading to errors in the recognition is resolved by smoothing of the language model. Smoothing uniformly redistributes parts of probabilities of observed *n*-grams among *n*-grams which are not observed in training data set. Nowadays, there exist several different smoothing techniques, such as *additive Add-One* or *Add-δ smoothing*, *Ristad natural law*, *Good-Turing estimation*, *Katz back-off model*, *absolute* and *linear discounting*, *Witten-Bell model* (Manning & Schütze, 1999), or *Kneser-Ney smoothing* and its modifications, which use counting of *n*-grams or counting these counts in computing discounting constants in smoothing LMs (Chen & Goodman, 1996).

We observed that among all smoothing techniques in modeling of the Slovak language, the optimal results produce followed algorithms:

- *Katz model* in smoothing LMs trained on small text corpora (approx. hundrets of MB);
- *modified Kneser-Ney algorithm* in smoothing LMs trained on huge text corpora (approx. tenths of GB);
- *Witten-Bell smoothing* in modeling of the Slovak language from text corpora with more regular structure of sentences.

4.3 Adaptation and combination

In the process of enhancing the performance of the LVCSR system, the *language model adaptation* (LMA) plays an important role in case of domain-specific speech recognition. The basic idea of LMA is to use a small amount of domain-specific text data to adjust LMs to reduce the impact of languages differences between the training and testing text data and set the parameters for independent topic-dependent LMs to correspond domain as much as possible with the real conditions of LVCSR application. LMA includes not only statistical dependencies between words in given language, but also the frequency of word occurrences, structure of the text data and further additional information that usually come from the field of linguistics and phonology (Staš et al., 2010a).

The LMA is usually performed by combining several (different) topic-dependent LMs when adaptation text (held-out data set) is used for adjusting the parameters of these LMs. In recent years, many different techniques have been designed for adaptation and combining LMs, including *maximum a posteriori* (MAP) approaches such as *count merging* and *linear*, *log-linear* or *generalized linear interpolation* (Gao et al., 2006; Hsu, 2009) and some discriminative methods such as LMA based on *minimum discriminative information*, *boosting* and *perceptron algorithm* or *minimum sample risk method* (Gao et al., 2006), which come from *maximum entropy* approach.

We have observed that algorithms producing significant results for strong statistically dependent languages such as English language, do not bring notable improvement in modeling the Slovak language. Based on detailed analysis experimental results methods for adaptation and combination LMs published in (Staš et al., 2010a), we also achieved that usage of the *linear interpolation* or its generalized alternative for the Slovak language is more than sufficient and interpolation weights should be adjusted using *expectation-maximization* (EM) algorithm by minimization of perplexity on held-out data set.

4.4 Pruning

Typically, an uncompressed LM in highly inflective language is comparable in size to the text data on which it has been trained. To build LMs for the task of real-time application it is necessary to limit the size of the resultant LM. In highly inflective languages, with using a large vocabulary increases the number of *n*-grams in LM which may occur in the training set just once or twice and do not have a big impact on the quality of LM or accuracy of the recognition system. Therefore, these *n*-grams can be excluded from the LM using pruning.

There exist several criteria for pruning LMs. To create an efficient and compact model of the Slovak language for using in real-time application of LVCSR system we observed the influence on the quality of LM of following pruning methods: (a) *cutoff counts*, (b) *weighted difference method* (Seymore & Rosenfeld, 1996), and (c) *pruning based on relative-entropy* (Stolcke, 1998). We found out that the *relative entropy-based pruning* achieved the best results.

5. Model optimization

Several different techniques and principles have been used and proposed in order to get an efficient model of the Slovak language for off-line and on-line speech recognition. These so-called optimization techniques which include the statistical, linguistical and phonetical principles and practices and lead to the increasing the quality of language models, decreasing errors in LVCSR system and usability of these models in real conditions of speech recognition in Slovak are described in following sections.

5.1 Spelling pronunciation

One of the main problems in speech recognition having the significant influence on the overall result of the speech recognition is how to implement the best phonetic transcription of words contained in dictionary. The *transcription of words from orthoepic to the ortographical form* concerns also such words as *abbreviations* or *acronyms* usually spelled character-by-character, for example: IBM, PhD., P. O. Box, etc. These events were necessary to unify, also to define their transcription to the Slovak phonetic alphabet (Cerňak et al., 2003) and to assign them all possible pronunciation variants. Regarding to the Slovak language, we detect about 620 abbreviations and acronyms (510 alternative pronunciations) in the text corpora mentioned in Section 2 and manually modified their transcription under linguistic rules used in the Slovak language.

5.2 Modeling of noise events

Spontaneous speech is also characterized by various *non-speech sounds* or *expressions* which are mainly generated by the speaker or surrounding environment. On the analysis of the resulting hypotheses obtained from the output of our dictation LVCSR system, we encountered relatively a lot of mistakes at the beginning of the speech or after long pause, in situations where the speaker paused, coughed, lip smacked, etc. We have decided to explore such ways in which it would be possible to model these so-called *noise events* in Slovak language modeling without having a knowledge of their occurrences in the training data set and false increase the estimate of their probabilities. Since the locations of the noise events are usually tagged by annotators during transcription or annotation of speech recordings into

text by special tags, we have decided to include these *annotations of speech recordings* with noise tags into the process of training LMs and model the Slovak language by using selected noise events as well.

First, we had to map all noise tags contained in annotations into five groups: (a) *short pause*, (b) *long pause*, (c) *filled pause*, (d) *background* and (e) *speaker noise* (Staš et al., 2010b) and these were later included into the dictionary and used in language modeling. It is important to say that after recognition, these noise events have appeared in output like *transparent words*.

5.3 Multiwords in Slovak language modeling

As it was mentioned in previous section, the most common mistakes in speech recognition arise at the beginning of the speech or after long pause and these also can be caused by *misrecognition of short monosyllabic words* consisting of no more than three or four characters. These words are often added to the following or preceding word, recognized as a noise or ignored (Kolorenč et al., 2006). To avoid this problem, it is suitable to model these events using *multiword expressions* (MWEs).

It has been showed that MWEs in the form of connection of short (monosyllabic) word with long (di-, tri- or polysyllable) word, which is usually more recognizable, can help increasing the recognition accuracy of the given short word. Moreover, using MWEs increases the order of *n*-gram LM and decreases the number of pronunciation variants depending on the context of the given word, because in an inflective language some of the words are pronounced differently in different context.

The extraction of MWEs in the Slovak language was performed by following selection criteria (Staš et al., 2011b):

1. both words forming the MWE and MWE itself must occur frequently in the text corpus;
2. MWE is formed by at least one short word, consisting from no more than three characters;
3. final selection is conditioned by additional linguistic constraints.

For the process of selection multiwords, we have used the standard statistical measures based on *absolute* and *relative co-occurrence* and *pointwise mutual information* (PMI) of these word pairs in the text corpora limited by the linguistic constraints. Selection measures was intentional. Absolute frequency expresses the most frequented events in given language. Relative frequency in the context of the first word extract MWEs such part-of-speech in Slovak as prepositions, conjunctions or pronouns usually occuring in the first place of given MWE. PMI reflects collocations which do not occur in language frequently but usually have certain meaning.

Linguistic constraints come from the observations of the behaviour of a LVCSR system in the process of testing LMs. It have been discovered that our LVCSR system is often mistaking in following cases: (a) there was an *assimilation of voicing* on a word boundaries and (b) if a first word in MWE ended with *same letter* as the second word begins.

Using mentioned and proposed methodology for extraction MWEs from the text corpora we obtained about 3 000 word pairs (561 pronunciation variants) which were included into dictionary with phonetic transcription and into the process of training Slovak language models.

5.4 Class-based models

Another problem by using LVCSR system is a possibilty to insert new words into the dictionary and LM. The similar problem arises also in recognizing proper nouns such as names, surnames or geographical names and other name entities. The recognition of names and surnames is one of the key properties of the on-line dictation LVCSR system which has noticeable influence on its usability in real conditions. There are different suboptimal solutions, by which we can cover a large part of the vocabulary in given language and also we can deal with the problem of insertion new words without overtraining LM. One of these solutions are *class-based* LMs which have great importance in dealing with certain problematic tasks, because they generalize context dependency of also such words, which have not occurred in the training corpora yet. We decided to use class models in modeling names and surnames in Slovak, in order to easily extend the class of words just for this case and resolve the problem of insertion new words into the dictionary.

For this purpose, we developed the *rule-based morphological tagger for names and surnames*, which is based on pattern matching principle from predefined set of names and surnames (patterns) conditioned by their grammatical category. The accuracy of this approach is then limited just by the number of patterns and selected rules. In this case the principle based on semantic similarity of formal expressions and syntactic knowledge contained in grammatical category of a proper noun is used. Using this approach we replaced thus approximately 24 818 unique inflected forms of names and surnames with one from the set of 20 morphological tags which have been depended on the case of given proper noun.

Also, it is important to say that for increasing of recognition accuracy we have created an *independent model for names and surnames* which can be used in special dictation mode in our dictation LVCSR system in the Slovak language as a parallel model of a primary domain-specific LM from the field of judicature.

5.5 Morphology

The inflection in Slovak language usually occurs on the border of stem and endings. This knowledge can help modeling of unknown words or words with a low occurrence in training corpus using *morpheme-based models* (Byrne et al., 2000; Creutz et al., 2007). Dividing singletons or words with a low frequency in the training corpus into morphemes, it is statistically possible to cover such events which do not occur in dictionary and LM. The knowledge of morphology of the given language then allows to also generate new word forms, for example as it was in the case of declination of names and surnames described in Section 3 or Section 5.4.

5.6 Augmentation statistics of *n*-grams

Nowadays, research in the language modeling is oriented on the augmentation of statistics of bigrams or trigrams from other resources than by gathering a large amount of text data of given language. Statistics of seen or unseen *n*-grams can be obtained by using:

1. statistics of *n*-grams contained in free available *academical* or *national text corpora*;
2. *web search engines* by copying statistics of *n*-grams on the Internet (Creutz et al., 2009; Oger et al., 2010; Zhu & Rosenfeld, 2001);

3. machine translation systems in *translation n-grams* from other (similar) languages.

At the end of this section, it is important to say that contemporary modeling of the Slovak language uses only the text data (trigrams) obtained from the Slovak National Corpus (SNC) (Šimková, 2006) to the augmenting the statistics of *n*-grams used in training LMs. However, the research and development in the other mentioned areas does not lag.

6. Speech recognition setup

In the following sections, the setup of our LVCSR system and description about proposed methodology of training Slovak LMs, used annotated speech databases, acoustic modeling and data for testing LMs is presented. The setup of LVCSR system was adjusted to the testing of LMs oriented to the judicial domain and broadcast news transcription in the Slovak language.

6.1 Language modeling

Experiments have been performed with trigram LMs which were created using tools contained in the *SRI Language Modeling* (SRILM) *Toolkit* (Stolcke, 2002) with vocabulary mentioned in Section 3. The complete process of building the reference LM of the Slovak language can be resumed into following steps:

- *extraction the statistics of trigram counts* from each of the domain-specific corpora;
- *calculation the statistics of counts-of-counts* for estimating Good-Turing discounts needed in the process of smoothing LMs;
- *calculation the discounting constants* used in smoothing LMs by the modified Kneser-Ney algorithm from obtained discounts;
- *computing the perplexity* of each domain-specific LM for each sentence on held-out (development) data set;
- *computing the parameters* (interpolation weights) for individual LM by minimization of perplexity on held-out data set using EM algorithm from obtained files with PPL;
- *creation the final domain-adapted LM* by the weighted combination of particular domain-specific trigram LMs combined by *linear interpolation*;
- *pruning the resulting LM using algorithm based on relative entropy* in order to use it in the real-time application in domain-specific task of Slovak LVCSR.

6.2 Acoustic modeling

The triphone context-dependent acoustic models based on the *hidden Markov models* (HMM) have been used, where each state have been modeled by 32 Gaussian mixtures. The models have been generated from feature vectors containing 39 *mel-frequency cepstral* (MFC) *coefficients*. They have been trained on two databases of annotated speech recordings.

The first broadcast news speech database contains about 60 hours of readings mostly by professionally trained speakers recorded from Slovak TV broadcast news from 2007 to 2009 year. The database is characterized by gender balanced speakers, contains read, spontaneous and in a small amount also telephone speech with 48 *kHz* sampling frequency and 16 *bit* resolution.

The second judiciary speech database contains about 120 hours of reading real adjudgments from the court with personal data changed, recorded in studio conditions and about 130 hours of read phonetically rich sentences, newspaper articles, internet texts and spelled items, recorded in offices and conferencial rooms. The database, total size of 250 hours, was recorded from 250 gender balanced speakers with 48 *kHz* sampling frequency and 16 *bit* resolution. It has been then extended with about 100 hours of 90% male spontaneous speech, recorded from 120 speakers at council hall with 44 *kHz* sampling frequency and 16 *bit* resolution.

All recordings were later downsampled to 16 *kHz* for training and testing. The databases were annotated by team of trained annotators using the *Transcriber annotation tool* (Barras et al., 2001), slightly adapted to our need, twice checked and corrected.

For acoustic modeling rare triphones the *effective triphone mapping* algorithm was used (Darjaa et al., 2011). With reference to the authors, this knowledge-based triphone tying, which allows the synthesis of unseen triphones, outperforms standard tree-based state tying for acoustic models with 4 000 states and more, whereas for acoustic models with smaller number of states the performance is equal.

6.3 Phonetic transcription

Phonetical transcription selected words contained in vocabulary was performed using *data-driven approach to orthoepic transcription in the Slovak language* (Cerňak et al., 2003) with slight modifications. It has been trained using the phonetically rich sentences from the SpeechDat-E and MobilDat-SK Slovak speech databases (Rusko et al., 2006) with a new sentence-based pronunciation lexicon, and additional sentences with manually annotated pronunciation from a regional broadcast news speech corpus.

6.4 LVCSR decoder

For decoding, the *high-performance LVCSR engine Julius* (Lee et al., 2001) with recognition algorithm based on the two-pass strategy has been used. The input data using this algorithm are processed in the first pass with left-right bigram LM, and the final search for reverse right-left trigram model is performed again using the result of the first pass to narrow the search space.

6.5 Test data set

The first test data set was represented by 240 minutes of recordings obtained by randomly selected segments from broadcast news speech database. These segments were not used in the training acoustic model and contain 40 656 words in 4 343 sentences.

The second test data from the field of judicature were represented by 315 minutes of recordings obtained also by randomly selected segments from each speaker contained in the second read (250 hours) speech database. As well as in the first case, these segments were not used in training and contain 41 878 words in 3 426 sentences and phrases. We have decided to use also phrases in the second test set, because in real conditions, people make pause not only on the sentence boundaries, but also on phrase boundaries, usually before conjunctions.

6.6 Evaluation

Two standard measures have been used for evaluation of the LM: (a) extrinsic evaluation using *word error rate* (WER) and (b) intrinsic evaluation based on *perplexity* (PPL) calculated on a test data set. WER is a standard measure of the performance of the LVCSR system, computed by comparing reference text read by a speaker against the recognized result and takes into account insertion, deletion and substitution errors. If the LVCSR system is not available, the perplexity is often used for evaluation. It is defined as the reciprocal of the (geometric) average probability assigned by the LM to each word in the test set. This measure does not necessarily evaluate the accuracy of recognition itself, but usually highly corelates with it.

7. Experimental results

The experiments were oriented on the evaluation of WER and PPL on the test data set to discover the effect of proposed optimization techniques and principles in Slovak language modeling on the overall recognition accuracy of the LVCSR system. As it was mentioned in Section 6.1, the experimental results were performed with trigram LMs created with vocabulary size of 348 255 unique words or more, listed in the Table 3, and smoothed by using *modified Kneser-Ney algorithm* in any case. For adaptation and combination LMs trained independently on text corpora mentioned in the Table 1, standard *linear interpolation* have been used, where interpolation weights were adjusted to the selected domain using EM algorithm. The experiments were oriented to the off-line testing of LMs, where the emphasis is focused on the best recognition accuracy than to the memory requirements of application as in on-line speech recognition, where it is necessary to use one of the pruning techniques of LMs. In the case of pruned models, it would be difficult to find appropriate pruning threshold, to maintain the equal number of n-grams in LM and compare the contribution of given LM to the speech recognition.

To observe the impact of selected optimization techniques and principles to the area of speech recognition training and testing of LMs were performed in two independent areas: (a) for broadcast news transcription task and (b) in judicial domain. This step also includes the usage of appropriate acoustic model and speech recordings for testing, described in Section 6.2 and Section 6.5, respectively. Experimental results for both tasks in Slovak LVCSR are described in following sections.

7.1 Broadcast mews transcription

Broadcast news transcription task is directed to the general area of the speech recognition, usually for recognition and transcription of a continuous spontaneous speech. In modeling of the Slovak language and adaptation to this domain we achieved following results. As we can see in the Table 3, using adaptation into the general area of speech recognition represented by randomly selected sentences from broadcast news text corpora not used in training process, we achieved almost 1.39% decreasing in WER and 17.36% of PPL, relatively. In the next step, modifying rules of phonetic transcription for spelled abbreviations, we observed moderate improvement rather in subjective than in objective point of view. This fact is caused also by the undesirable shortening of the history for some abbreviations such as P. O. Box, M. D., etc., and reducing predictive ability of the LM. Extending the training data set by the text data obtained

language model	size of vocabulary	broadcast news PPL_{test}	broadcast news WER [%]	judicial domain PPL_{test}	judicial domain WER [%]
base without adaptation	348 255	401.105	10.78	126.720	7.92
domain adaptation	348 255	331.478	10.63	100.383	6.96
pronunciation modification	348 468	332.974	10.62	96.1422	6.97
added noise events (1)	348 473	326.519	10.54	57.2970	6.26
added multiwords (2) + (1)	351 473	336.558	10.52	62.7111	6.22
added classes (3) + (2) + (1)	351 493	302.711	10.50	56.1970	6.05
augmented (1)	348 473	308.113	10.31	56.7245	6.27
statistics (2) + (1)	351 473	319.578	10.37	64.2670	6.26
of n-grams (3) + (2) + (1)	351 493	287.815	10.18	55.5591	6.05

Table 3. Experimental results for off-line testing of the Slovak LVCSR system

from annotations of speech recordings we achieved additional decreasing, relatively 0.75% WER and about 2% of PPL. Taking into account that the testing data from general domain contained only small amount of selected MWEs and names or surnames, the contribution to the speech recognition of established multiwords and word classes into LM was too small. Variations were observed only in perplexity, which was increased due to the shortening of the history for MWEs and on the contrary decreased by more fixed connections between word classes. The significant improvement we achieved mainly in the case of augmentation of trigrams from the SNC database. Decreasing of about 3% WER and 5% of PPL relatively, results in the fact that the SNC database contained mostly the text data from newspapers or fictions. The impact of selected optimization techniques to the broadcast news transcription task in Slovak LVCSR brought overall reduction approximately 5.57% WER and 28.24% of PPL, relatively.

7.2 Speech recognition in judicial domain

This domain was selected as one of the most challenging acoustic and linguistic environments from the research point of view, and based on market demand, from the development point of view. Regarding adaptation into the judicial domain, we achieved significant improvement, relatively 12.12% in WER and 20.78% of PPL even if a small amount of adaptation data was added. As it was in the previous case of broadcast news transcription, by modifying pronunciation of spelling items, there were not observed any notable variations in WER or PPL. The impact of the text data from annotations of speech recordings results in significant decreasing of both values, more than 10% in WER and 40% of PPL, relatively. This fact is caused mainly by larger amount of text data (more hours) from annotations of speech recordings from judicial domain than in broadcast news transcription task. Multiwords brought an improvement in just about 5% of cases at the beginning of the speech or after long pause, what did not produce significant changes in the overall result of the speech recognition. Due to the fact that the testing data contained a large amount of names and surnames, we achieved additional decreasing, relatively 3% WER and more than 10% of PPL in the case of word classes. Augmentation statistics of trigrams did not improve resultant LM, because mentioned database does not contain any text data from the field of judicature. The contribution of mentioned optimization steps to the domain-specific task of Slovak LVCSR yield overall reduction approximately 24% in WER and 56% of PPL, relatively.

7.3 Discussion

Using selected methods, principles and approaches in statistical modeling of the Slovak language and proposed optimization techniques we achieved the recognition accuracy of our LVCSR system almost 94% in domain-specific task from the field of judicature and approximately 90% in the case of broadcast news transcription. The vocabulary used in experiments covers about 99% commonly used words in the Slovak language.

As regards the experimental results, the recognition accuracy could be increased by extending word classes with names of cities, streets, institutions, and other name entities in their inflected form. Regarding memory requirements, it could be more suitable to use only class-based approach in Slovak language modeling. However, absence of any available morphological tagger for Slovak language limits the utilization of this approach, although first steps in this area have already been done.

Contemporary research in Slovak language modeling is also oriented on different areas such as vocabulary selection in specific domain, topic detection in web corpora, augmentation statistics of the LM using machine translation systems or web engines, on-line adaptation of LMs, modeling of unknown words in spontaneous speech, morphologically motivated class-based modeling, discovering the influence of the morpheme-based models, and eliminating errors caused by used vocabulary or language modeling in speech recognition.

As regards the real application of domain-oriented speech recognition, nowadays, a new version of our LVCSR system for the purpose of the Ministry of Justice of the Slovak Republic is being finalized, in which these knowledges about the modeling of the Slovak language and LMs described in this chapter have been used. It is important to say, that at the time of the preparation of this chapter proposed LVCSR system has been installed and used by more than 50 persons (judges, court assistants and technicians) at 9 different institutions belonging to the Ministry of Justice for testing. The results of tests will be taken into consideration in the final version of the Slovak LVCSR system coming into everyday use at the organizations belonging to the Ministry of Justice of the Slovak Republic by the end of the year 2011.

8. Conclusion

In this chapter a brief summary of current methods and principles used in Slovak language modeling has been presented. By combination of standard statistical methods and proposed language dependent optimization techniques bringing an additional information into training process of LM, often linguistic regularities as well, we achieved notable improvement in recognition accuracy of our LVCSR system of the Slovak language in the task of broadcast news transcription as well as in domain-specific speech recognition from the field of judicature. We have discovered that using several different approaches oriented to the specific problem in language modeling, we can better eliminate errors arising in the speech recognition of such inflective language as is the Slovak language. The major contribution in the area of Slovak language modeling is the fact that current language models are also used in development and application of the Slovak automatic transcription and dictation LVCSR system for the judicial domain.

9. Acknowledgement

The research presented in this paper was supported by the Ministry of Education under research projects VEGA-1/0065/10 and MŠ SR 3928/2010-11 and by EU ICT Project INDECT (FP7–218086).

10. References

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production, *Speech Communication* 33(1-2): 5–22.

Byrne, W., Hajič, J., Krbec, P., Ircing, P. & Psutka, J. (2000). Morpheme based language models for speech recognition of Czech, *Proceedings of 3rd International Workshop on Text, Speech and Dialogue, TSD'2000*, Brno, Czech Republic, pp. 211–216.

Cerňak, M., Rusko, M., Trnka, M. & Daržagín, S. (2003). Data-driven versus knowledge-based approaches to orthoepic transcription in Slovak, *ICETA'2003: The 2nd International Conference on Emerging Telecommunications Technologies and Applications and the 4th Conf. on Virtual University*, Košice, Slovakia, pp. 95–97.

Chen, S. F. & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling, *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL'96*, Santa Cruz, CA, USA, pp. 310–318.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkänen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraclar, M. & Stolcke, A. (2007). Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages, *Proceedings of HLT-NAACL'2007*, Rochester, NY, USA, pp. 380–387.

Creutz, M., Virpioja, S. & Kovaleva, A. (2009). Web augmentation of language models for continuous speech recognition of SMS text messages, *Proceedings of the 12th Conference of the European Chapter of the ACL, EACL'2009*, Athens, Greece, pp. 157–165.

Darjaa, S., Cerňak, M., Trnka, M., Rusko, M. & Sabo, R. (2011). Effective triphone mapping for acoustic modeling in speech recognition, *Proceedings of INTERSPEECH'2011*, Florence, Italy, pp. 1717–1720.

Gao, J., Suzuki, H. & Yuan, W. (2006). An empirical study on language model adaptation, *ACM Transaction on Asian Language Information Processing, TALIP'2006* 5(3): 209–227.

Hládek, D. & Staš, J. (2010a). Text gathering and processing agent for language modeling corpus, *Proceedings of the 12th International Conference on Research in Telecommunication Technologies, RTT'2010*, Veľké Losiny, Czech Republic, pp. 200–203.

Hládek, D. & Staš, J. (2010b). Text mining and processing for corpora creation in Slovak language, *Journal of Computer Science and Control Systems* 3(1): 65–68.

Hsu, J. B. (2009). *Language modeling for limited-data domains*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Jurafsky, D. & Martin, J. H. (2009). *An introduction to natural language processing, computational linguistics, and speech recognition (2nd edition)*, Prentice Hall.

Kolorenč, J., Nouza, J. & Červa, P. (2006). Multi-words in the Czech TV/radio news transcription system, *Proceedings of the 11th International Conference Speech and Computer, SPECOM'2006*, Sankt Peterburg, Russia, pp. 70–74.

Lee, T., Kawahara, T. & Shikano, K. (2001). Julius - An Open Source real-time large vocabulary recognition engine, *Proceedings of EUROSPEECH'2001*, Aalborg, Denmark, pp. 1961–1694.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press.

Nouza, J., Zdansky, J., Cerva, P. & Silovsky, J. (2010). Challenges in speech processing of Slavic languages (Case studies in speech recognition of Czech and Slovak), *A. Esposito et al. (Eds.): Development of Multimodal Interface: Active Learning and Synchrony*, LNCS 5967, Springer-Verlag, Heidelberg, pp. 225–241.

Oger, S., Popescu, V. & Linarès, G. (2010). Combination of probabilistic and possibilistic language models, *Proceedings of INTERSPEECH'2010*, Makuhari, Japan, pp. 1808–1811.

Rusko, M., Trnka, M. & Daržagín, S. (2006). MobilDat-SK - A mobile telephone extension to the SpeechDat-E SK telephone speech database in Slovak, *Proceedings of the 11th International Conference Speech and Computer, SPECOM'2006*, Sankt Peterburg, Russia, pp. 485–488.

Seymore, K. & Rosenfeld, R. (1996). Scalable backoff language models, *Proceedings of the 4th International Conference on Spoken Language Processing, ICSLP'96)*, Philadelphia, PA, USA, pp. 232–235.

sk-spell (2010). *Slovak support in Open Source applications*, Projekt sk-spell. (in Slovak). URL: *http://www.sk-spell.sk.cx/*

Staš, J., Hládek, D. & Juhár, J. (2010a). Language model adaptation for Slovak LVCSR, *AEI'2010: International Conference on Applied Electrical Engineering and Informatics*, Venice, Italy, pp. 101–106.

Staš, J., Hládek, D., Pleva, M. & Juhár, J. (2011a). Slovak language model from Internet text data, *A. Esposito et al. (Eds.): Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, LNCS 6456, Springer-Verlag, Heidelberg, pp. 340–346.

Staš, J., Hládek, D., Trnka, M. & Juhár, J. (2011b). Automatic extraction of multiword expressions using linguistic constraints for Slovak LVCSR, *Proceedings of the 6th International Conference on NLP, Multilinguality, SLOVKO'2011*, Modra, Slovakia, pp. 1–8.

Staš, J., Trnka, M., Hládek, D. & Juhár, J. (2010b). Text preprocessing and language modeling for domain-specific task of Slovak LVCSR, *Proceedings of the 7th International Workshop on Digital Technologies, Circuits, Systems and Signal Processing, DT'2011*, Žilina, Slovakia, pp. 1–4.

Stolcke, A. (1998). Entropy-based pruning of backoff language models, *Proceedings of DARPA Broadcast News and Understanding Workshop*, Lansdowne, VA, pp. 270–274.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit, *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP'2002*, Denver, Colorado, USA, pp. 901–904.

Venkataraman, A. & Wang, W. (2003). Techniques for effective vocabulary selection, *Proceedings of EUROSPEECH'2003*, Geneva, Switzerland, pp. 245–248.

Šimková, M. (2006). Slovak National Corpus - History and current situation, *M. Šimková (Ed.): Insight into the Slovak and Czech Corpus Linguistics* VEDA - Publishing House of Slovak Academy of Sciences, Bratislava, pp. 151–159.

Zhu, X. & Rosenfeld, R. (2001). Improving trigram language modeling with the world wide web, *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing, ICASSP'2001*, Salt Lake City, Utah, USA, pp. 533–536.

New Technologies - Trends, Innovations and Research

Edited by Prof. Constantin Volosencu

ISBN 978-953-51-0480-3
Hard cover, 396 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

The book "New Technologies - Trends, Innovations and Research" presents contributions made by researchers from the entire world and from some modern fields of technology, serving as a valuable tool for scientists, researchers, graduate students and professionals. Some practical applications in particular areas are presented, offering the capability to solve problems resulted from economic needs and to perform specific functions. The book will make possible for scientists and engineers to get familiar with the ideas from researchers from some modern fields of activity. It will provide interesting examples of practical applications of knowledge, assist in the designing process, as well as bring changes to their research areas. A collection of techniques, that combine scientific resources, is provided to make necessary products with the desired quality criteria. Strong mathematical and scientific concepts were used in the applications. They meet the requirements of utility, usability and safety. Technological applications presented in the book have appropriate functions and they may be exploited with competitive advantages. The book has 17 chapters, covering the following subjects: manufacturing technologies, nanotechnologies, robotics, telecommunications, physics, dental medical technologies, smart homes, speech technologies, agriculture technologies and management.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821