Language models for spontaneous speech recognition

Jiri Valicek, Petr Mizera
Czech Technical University, Faculty of Electrical Engineering
Technická 2, 166 27 Praha, Czech Republic
valicjir@fel.cvut.cz, mizerpet@fel.cvut.cz

Abstract.
The paper presents the creation of n-gram Language Models (LMs) for the purposes of spontaneous speech recognition with the special focus on the recognition performed with the data from the Nijmegen Corpus of Casual Czech (NCCCz). Required LMs which cover spontaneous or casual speech respectively were created using the data available in the collected corpus NCCCz. These models were combined with a general model built on the basis Czech National Corpus. The creation of LMs, their adaptation to specific thematic domain (casual speech) were created using SRILM toolkit. The quality of created LMs was measured on the basis Perplexity and Out-Of-Vocabulary words at the text level, same as they was used within the automatic speech recognition and the achieved Word Error Rates are also presented. The study shows that created LMs were suitable for given purpose and achieved results were related to other works presenting the spontaneous speech recognition.

Keywords
speech recognition, LVCSR, language models, n-grams, spontaneous speech, casual speech

1. Introduction

Automatic speech recognition (ASR) systems are used currently in our daily life as human to machine interfaces in many applications, e.g. instead of keyboard for text input to PC, to control a machine by voice or in other tasks. ASR systems for major languages are nowadays well developed and used in cell phones, smart televisions, personal computers. Systems for minor languages such as Czech were significantly improved over last decades and they work usually very accurately, similarly to systems for major languages.

Particular speech recognition tasks can be divided into three groups: the recognition of single spoken words, the recognition using small vocabulary, and mainly Large Vocabulary Continuous Speech Recognition (LVCSR). The main goal of current research teams is to build a system that can recognize words from ideally unlimited vocabulary, independently on the speaker, and which can perform well even if the speech recording conditions are not ideal. It means a presence of background noise in collected speech, some channel distortion, speech disorder, and also the case of informal and spontaneous speech.

Concerning LVSCRs, they works generally with two principal stochastic models: Acoustic Model (AM) modelling the particular characteristics for the speech waveform and Language Model (LM) modelling the order of words in a sentence, i.e. a relations between particular words. The main purpose of the work presented in this paper was to create suitable LMs for the task of spontaneous (Causal) speech recognition. The main goal was to find the optimal way of creating LMs for spontaneous speech recognition by the creating models specific for casual speech and combining them with a general language models. This work is a next logical step to the realization of casual speech recognition and it follows the work presented in [1]. Created language models should be then used for the recognition of causal speech and achieved accuracy of speech recognition will be analysed.

The paper is organized as follows. In Chapter 2 basics of language modelling in ASR is presented, the following chapter 3 addresses the creation of language models for our purposes, achieved results of LVCSR are presented and discussed in Chapter 4, and the paper is closed by conclusions.

2. Language modelling in LVCSR

As said above, the role of a LM in a speech recognition system is to model relations between words and consequently the general structure of sentence. The key role in the LM creation plays the source text corpus which is used for its building. Suitable training corpus has to be sufficiently large to cover the language and the given topic as much as possible. This is not possible to guarantee generally, so LMs are created usually from a general model which covers the language as whole and then it is combined (adapted) with suitable thematic LM. This is the standard approach used in LVCSR in particular thematic domains and it will be used also by us for casual speech recognition.

2.1. n-gram language models

Statistical n-gram language models are represented by a probability of a sequence of n words. In fact these models are lists of probabilities of an occurrence for each sequence of words from one up to the length n. Special kind of n-
gram language model is so called zero grams. In the case of zero grams, each word has the same probability and it is used only for some special purposes, e.g. when the quality of acoustic model purely should be analyzed. Within LVCSR systems, 2-grams (bigrams) and 3-grams (trigrams) represent the most frequently used LMs. For some special cases, the models based on 4-grams or higher can be used but they are very huge and their usage require higher computational power. n-grams of higher order are typically used only for some very specific and frequent word sequences.

The reason of the popularity of n-gram LMs is mainly in the fact that they can be computed rather easily and current LVCSR toolkit support them standardly. To create n-gram LM, it is necessary to compute relative occurrences for every sequence of words of the length up to chosen n-gram order. The estimation of particular probabilities in trigram LM can be determined as the ratio of $N(w_{k-2}, w_{k-1}, w_k)$, the count of occurrence for sequence of words $w_{k-2}, w_{k-1}, w_k$, over $N(w_{k-2}, w_{k-1})$, the count of occurrence for sequence of words $w_{k-2}, w_{k-1}$ in training corpus [5].

2.2. LM smoothing

When creating n-gram LMs, of course, it is built typically from the finite text corpus so it is impossible to create LM which will contain all words that could be in the utterance which will be recognized and given finite text it is also known that the probability we would be giving to n-grams with lowest occurrence would not correspond with reality.

That is why when LM is built in is necessary to preserve some probability for n-grams which were not in our training corpus same as for n-grams which have appeared only few times or once. To achieve that some probability from higher frequent n-grams is reallocated. This procedure is called as smoothing of LM and currently the most used smoothing methods are Good-Turing(GT), Kneser-Ney(KN) and Witten-Bell(WB) one [5].

2.3. Thematic corpora and LM adaptation

When a LM is created from given corpus, it is clear that it covers a topic of source corpus. When creating speech recognition system it is vital to take in mind the topic of recognition. Speech recognition can be used in many fields, each working with different vocabulary, e.g. transcripts in court, technical lectures or dictation software. Since it would be nearly impossible to collect enough data to create good model for each field, combination of general and thematic model is used. While combining/adapting general LM with thematic LM weight has to be chosen. Generally thematic LM has higher weight but in some cases the weight can be same for each model.

2.4. LMs of casual speech

For the purpose of language modelling for casual speech recognition the LMs were built using the following procedure. Firstly, general LMs were used, i.e. bigram and trigram models from Czech National Corpus 2-gram and 3-gram. These models contained 240000 or 340000 unigrams, i.e. such number of words appearing within all n-grams. We used models created and analyzed within [1].

LMs for causal speech were created from NCCCz [3] (Nijmegen Corpus of Casual Czech) using the transcriptions of utterances from 59 speakers (both male and female ones). This corpus was divided into two groups. Always one sub-part was used as training corpus and the second one as testing corpus (in fact recognized utterances later), exactly 80% of corpus was used usually for the training and remaining 20% represented the evaluation corpus. The sentences for training and testing corpus were chosen from utterances having more than five words which should have eliminate less important utterances without meaning, e.g. “Ale no ale to”.

2.5. SRILM & ARPA format

To create above described LMs, SRILM [6] toolkit was used. It is nowadays used as a standard for the creation of n-gram LMs, same as for further manipulation with them. It contains large set of tools which are very easy to work with and which usage is well documented by good manual pages. Its additional benefit is that it can perform basic test of created models at the text level on the basis of perplexity and number out-of-vocabulary words.

Particular tools use various way of representation of n-grams, however, ARPA format of n-gram LMs is used (or at least supported) by almost all speech recognition toolkits. For our purposes we use this format to ensure easy interoperability when tools from various toolkits are used.

3. Experimental setup

In order to pick adequate language models we had to compare many models and to ease our work and in mind of future, set of scripts was created. These script were inspired by [7] and [10].

3.1. Data

All experiments were performed using the speech data from three Czech databases. All of them were recorded at 16kHz sampling frequency and saved to 16-PCM format. The SPEECON [4] and CZKCC (private car speech data) databases consisted of signals with of high-quality recordings and contained phonetically rich sentences suitable for creating acoustic models. The NCCCz contained the recordings of informal conversations [3]. The choice of train and test sets used for the evaluation is summarized in Tab. 1.

<table>
<thead>
<tr>
<th>Database</th>
<th>set</th>
<th>speakers</th>
<th>sentences</th>
<th>hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEECON</td>
<td>train</td>
<td>225</td>
<td>60877</td>
<td>53.6</td>
</tr>
<tr>
<td></td>
<td>train</td>
<td>302</td>
<td>12771</td>
<td>20.6</td>
</tr>
<tr>
<td>CZKCC</td>
<td>train</td>
<td>40</td>
<td>10975</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>20</td>
<td>890</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Tab. 1: Database for training and testing
3.2. Feature Extraction

The speech signal is segmented to frames by window of length 25 ms and shift 10 ms. The common 13 mel-frequency cepstral coefficients (MFCC) with $c[0]$, delta and delta-delta parameters are extracted using the Ctcucopy tool [2]. The time context of 11 frames was stacked to high dimension vector. The high dimension of vector is reduced by Linear discriminant analysis (LDA) and decorrelated by maximum likelihood linear transform (MLLT). Finally the feature vector of the size 40 was used.

3.3. Acoustic modelling

Our LVCSR system was based on standard GMM-HMM architecture. The GMM-HMM system was built from 45 Czech phones which was then expanded to the context-dependent crossword triphones. The GMM-HMM models were trained using the embedded Viterbi algorithm. The final trained AM consisted of 15029 number of Gaussian mixture components. The more conventional system was built using the feature-space maximum likelihood linear regression (FMLLR) based on speaker adaptive training (SAT). The all processes of creating AMs was performed by modern KALDI toolkit [10]. The following AMs were created:

- tri2 - GMM-HMM crossword triphones system,
- GMM-HMM-SAT system.

3.4. Used language models and lexica

As said above to create and work with LMs SRILM toolkit was used. Most used tools were ngram-count and ngram. This two tools cover both creation and testing (on text basis) of LMs.

ngram-count tool – it is used for creation of LM from text corpus or n-gram counts and for counting n-grams from text corpus. The output can be LM or file with n-gram counts.

ngram tool – When LM was created this tool can be used to evaluate it using text testing corpus or mix it with other models using different weights. List of unigrams can be extracted from LM using this tool to create vocabulary file for speech recognition system.

Acronyms for created LMs

- CNK240 - Czech National Corpus, 240 000 unigrams
- CNK340 - Czech National Corpus, 340 000 unigrams
- N - created from whole corpus NCCCz
- N(80-20) - created from training corpus NCCCz

Acronyms for used lexica

- C - Czech National Corpus, 340 000 unigrams
- S - Speecon, 25 500 unigrams
- N - created from whole corpus NCCCz
- N(80-20) - created from training corpus NCCCz

3.5. Evaluation criteria

The quality of created LMs will be evaluated no the basis of Perplexity (PPL), Out-Of-Vocabulary words (OOV) and Word Error Rate (WER). Of course, the most important criterion for the performance of LMs in LVCSR is WER which presents the real results of LM used in LVCSR [8]. The PPL is the criterion evaluated on text testing corpus. To compute PPL it is not needed to use LVCSR [5]. Advantage over WER is that PPL is faster to compute than WER. During the process of creating adequate LM, the models were firstly compared using PPL and then the best models to use in speech recognition were chosen. Last measure important for the comparison of LMs is OOV [9]. It represents the number of unknown words in percent for LM which appear in text testing corpus.

4. Results and discussion

Thematic LMs from NCCCz and general LMs from CNC (created using KN smoothing only) were tested on testing corpus. These results for particular LMs testing are presented in Tab. 2. Results of this test corresponds with presumptions for thematic and general LMs. General LMs have lower OOV but their PPL is lot higher because of difference in topic. Difference between thematic LMs using different smoothing options were rather small (aprox. 7 % in PPL).

In the second step, combined LMs were tested. The goal was to lower OOV as the perplexity of thematic LMs seemed reasonable. Three weights for combining LMs were chosen - 25%, 50% and 75%[11]. Since thematic LMs scored much lower PPL, their combination with general LMs scored worse PPL but OOV was reduced by about 4-5%. Results presented in Tab. 3 were obtained for combination of CNK 240/340 and NCCCz(80-20) with WB smoothing.

Secondly, LMs were evaluated within LVCSR performance. Achieved results for created LMs are summarized in Tab. 4. As we expected, the best results were achieved for thematic LMs from all utterances of NCCCz. The value of WER for NCCCz setups represents the top achieved accuracy in this very difficult task of casual speech recognition in comparison to achieved results from general LMs. The impact of combination of LMs with various weight on achieved WER are summarized in Tab. 5 and 6. The best results about 40.7% of WER was achieved for CNK-340-N-wb setup . The complexity (2-gram vs. 3-gram) of LMs on final WER was also analysed. The best results were always achieved for 3-gram LMs.

5. Conclusions

The goal of this paper was to create new LMs for casual speech recognition using NCCCz and compare them with available general LMs. To achieve best results it was chosen to combine these LMs using three weights - 25, 50 and 75%. Best results were obtained for the combination of CNK340 with NCCCz using WB smoothing. The difference between 2-gram and 3-gram LMs were rather small, WERs differed by approximately 1% (55.4% – 54.3%). It is also important to mention that results obtained for casual speech recognition were with LMs covering even testing corpus. We can
Tab. 2: Results for not combined LMs on text test set - NCCCz 80_20

<table>
<thead>
<tr>
<th>LM</th>
<th>OOV [%]</th>
<th>2-gram</th>
<th>3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNK240</td>
<td>5.29</td>
<td>231.51</td>
<td>239.26</td>
</tr>
<tr>
<td>CNK340</td>
<td>4.65</td>
<td>248.70</td>
<td>219.30</td>
</tr>
<tr>
<td>N(80_20)</td>
<td>7.23</td>
<td>1831.44</td>
<td>1652.8</td>
</tr>
</tbody>
</table>

Tab. 3: Results for combined LMs on text test set - NCCCz 80_20

<table>
<thead>
<tr>
<th>LM</th>
<th>λ</th>
<th>OOV [%]</th>
<th>2-gram</th>
<th>3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNK240-N(80_20)</td>
<td>25</td>
<td>2.51</td>
<td>342.04</td>
<td>301.59</td>
</tr>
<tr>
<td>CNK340-N(80_20)</td>
<td>25</td>
<td>2.21</td>
<td>353.32</td>
<td>311.60</td>
</tr>
<tr>
<td>CNK240-N(80_20)</td>
<td>50</td>
<td>2.51</td>
<td>386.53</td>
<td>342.29</td>
</tr>
<tr>
<td>CNK340-N(80_20)</td>
<td>50</td>
<td>2.21</td>
<td>398.46</td>
<td>352.94</td>
</tr>
<tr>
<td>CNK240-N(80_20)</td>
<td>75</td>
<td>2.51</td>
<td>510.37</td>
<td>458.58</td>
</tr>
<tr>
<td>CNK340-N(80_20)</td>
<td>75</td>
<td>2.21</td>
<td>525.18</td>
<td>471.99</td>
</tr>
</tbody>
</table>

Tab. 4: Speech recognition on test set NCCCz 80_20 with various LM - WER [%]

<table>
<thead>
<tr>
<th>lex</th>
<th>sLM</th>
<th>tri2</th>
<th>tri3</th>
<th>tri2</th>
<th>tri3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C+S</td>
<td>CNK-240</td>
<td>63.9</td>
<td>57.7</td>
<td>63.4</td>
<td>56.3</td>
</tr>
<tr>
<td>C+S</td>
<td>CNK-340</td>
<td>63.9</td>
<td>57.9</td>
<td>63.4</td>
<td>56.7</td>
</tr>
<tr>
<td>N</td>
<td>N-kn</td>
<td>64.8</td>
<td>58.0</td>
<td>64.6</td>
<td>57.4</td>
</tr>
<tr>
<td>N</td>
<td>N-wb</td>
<td>62.3</td>
<td>54.4</td>
<td>60.4</td>
<td>53.5</td>
</tr>
</tbody>
</table>

Tab. 5: Speech recognition on test set NCCCz 80_20 with merged LM - WER [%]

<table>
<thead>
<tr>
<th>lex</th>
<th>sLM</th>
<th>tri2</th>
<th>tri3</th>
<th>tri2</th>
<th>tri3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C+S</td>
<td>CNK-340-N-kn</td>
<td>62.0</td>
<td>55.1</td>
<td>58.6</td>
<td>51.0</td>
</tr>
<tr>
<td>C+S</td>
<td>CNK-340-N-wb</td>
<td>65.1</td>
<td>58.2</td>
<td>60.4</td>
<td>53.5</td>
</tr>
</tbody>
</table>

Tab. 6: Speech recognition on test set NCCCz 80_20 with merged and weighted LM - WER [%]

<table>
<thead>
<tr>
<th>lex</th>
<th>sLM</th>
<th>λ</th>
<th>tri2</th>
<th>tri3</th>
<th>tri2</th>
<th>tri3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C+S</td>
<td>C-240-N</td>
<td>25</td>
<td>63.9</td>
<td>57.2</td>
<td>63.4</td>
<td>56.3</td>
</tr>
<tr>
<td>C+S</td>
<td>C-240-N-kn</td>
<td>75</td>
<td>68.9</td>
<td>62.1</td>
<td>68.2</td>
<td>61.8</td>
</tr>
<tr>
<td>C+S</td>
<td>C-240-N-wb</td>
<td>25</td>
<td>59.3</td>
<td>52.0</td>
<td>47.0</td>
<td>40.7</td>
</tr>
</tbody>
</table>

References


About Authors...

Jiri VALICEK was born in Sokolov, Czech Republic in 1990. He currently studies Master’s program at FEE, CTU in Prague.

Petr MIZERA was born in Jilemnice, Czech Republic in 1987. He graduated from the Faculty of Electrical Engineering and Informatics, University of Pardubice (2012) and the Faculty of Electrical Engineering, Czech Technical University In Prague (2012) where he currently studies the Ph.D programme and carries out research on spontaneous speech recognition based on articulatory features.