Analysis of Czech Web 1T 5-Gram Corpus and Its Comparison with Czech National Corpus Data

Václav Procházka and Petr Pollák

Dept. of Circuit Theory, Czech Technical University,
Technická 2, 166 27 Prague, Czech Republic
prochva1@fel.cvut.cz, pollak@fel.cvut.cz
http://noel.feld.cvut.cz/speechlab/

Abstract. In this paper, newly issued Czech Web 1T 5-grams corpus created by Google and LDC is analysed and compared with reference n-gram corpus obtained from Czech National Corpus. Original 5-grams from both corpora were post-processed and statistical trigram language models of various vocabulary sizes and parameters were created. The comparison of various corpus statistics such as unique and total word and n-gram counts before and after post-processing is presented and discussed, especially with the focus on clearing Web 1T data from invalid tokens. The tools from HTK Toolkit were used for the evaluation and accuracy, OOV rates and perplexity were measured using sentence transcriptions from Czech SPEECON database.

Keywords: statistical language model, text corpora, Czech Web 1T 5-gram, Czech National Corpus, HTK Toolkit.

1 Introduction

The research in the field of Large Vocabulary Continuous Speech Recognition (LVCSR) has undergone intense development over the past few decades for many world languages, including languages and dialects spoken by rather small population, as e.g. Czech. Especially on the basis of increasing power of IT systems, more sophisticated speech technology applications can be currently seen in many systems used in daily human life. The first recognizers of singular commands or simple dialogue systems are joined by dictation machines converting voice input into written form, automated transcribers of audio-video records [1], moreover, on-line sub-titles generation in live TV broadcasts [2] was developed.

Current LVCSR systems typically use statistical language modelling, which describes in the simplest case the probability of single word occurrence in language (unigram) or in more general case appearance probability of n-word sequence (n-gram). Language modelling has significant importance for the accuracy of a constructed LVCSR system, and the increasing power of such systems is frequently based on the continuously increasing size of vocabulary and language model. It is very typical for Czech as well as for other languages of inflective nature that the vocabulary and language model must be very large to enable the recognition of natural continuous speech.

The creation of proper language models for Czech has also been done by other authors within the construction of the above described LVCSR systems. This article

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 181–188, 2010.

describes the newly publicly available Web based resource for language modelling in Czech [3] commonly with its comparison with reference Czech National Corpus data [4]. Using this data should prevent hard and time consuming work of collecting large text corpus for the generation of statistical n-gram model and thus it enables quicker realization of LVCSR experiments.

2 Analysed n-Grams Corpora

The Web 1T 5-gram corpora contain the set of n-grams obtained from Web sources by Google in cooperation with Linguistic Data Consortium [5]. The World Wide Web is frequently used as a source of various text corpora and the data from Web resources has usually dominant contribution within currently used text corpora in speech technology generally. Gathering text from such resource inherently means dealing with problems like collecting data only for the requested language, recognising and unifying encodings or filtering out markup tags. Consequently, assembling sufficient amount of texts for specific language is not straightforward. Crawling the Web and collecting text data is time consuming work and, finally, a high number of invalid words such as misspellings, other language words or random chunks of characters still appears in a collected text.

Web 1T 5-gram corpora offer already prepared sets of n-grams from the order of one up to the order of five. The first Web 1T 5-grams corpus was issued for English in 2006 [6] and within the last year this collection was extended by 10 other languages, including Czech [3].

2.1 Google Czech Web 1T n-Grams

The statistics of Czech Web 1T n-grams reports a very large amount of unique unigrams, almost 10 million. Initial analysis showed that it is due to the case sensitive manner of particular tokens and due to frequent appearance of semi-random character chunks as serial numbers, product codes, personal or company names, Internet nicknames, or punctuation marks. While contribution of punctuation marks to unique unigram count is insignificant, they are not necessary for creating basic bigram and trigram language models. It ultimately means that available n-grams must be filtered before further usage. As this corpus was built using texts which were rather variable and which were not carefully edited as a whole such as other collected corpora for language modelling purposes.

In the original corpus all tokens which were considered invalid (contains malformed, non-european or other invalid characters or are too long [3]) were replaced by the special token <UNK> (unknown word). Similarly all tokens with occurrence lower than 40 were also replaced by special token <UNK>. Subsequently all n-grams with occurrence count lower than 40 were discarded. Cutoff 40 is a common simplified notation for this operation. In addition to the original corpus, more tokens were mapped to special token <UNK>. These tokens do not form a proper Czech word, e.g. mixed strings of alphabetical and numeric characters or URLs. For this purpose the following Czech alphabet letters were considered valid: ’aábcčdďeéěfghiíjklmnňoópqrřsštťuúůvwxyýzž’.

Table 1. Additional class mappings done during 5-gram filtering of Czech Web 1T 5-gram and SYN2006PUB n-gram corpora

token content	token	example of string
alphanumeric characters and numbers	<UNK>	abc123
words with letter out of Czech alphabet	<UNK>	sjöberg
numbers 0–9 +-,.	<NUM>	123

Table 2. CNC SYN2006PUB n-gram corpus; unique n-gram counts

	original corpus	cleared corpus
unigram	9,786,424	1,804,682
trigrams	117,264,988	47,376,975
5-grams	103,280,138	51,747,413

Table 3. Czech Web 1T 5-gram corpus statistic; unique n-gram counts

	original corpus	after postprocessing
unigram	2,554,028	1,779,006
trigrams	189,152,100	170,243,851

During further processing new special token <NUM> was introduced to represent number expressions. Overview of all additional mappings introduced during post-procesing is presented in table 1. The meaning of special tokens for sentence start (<S>) and sentence end (</S>) was left unchanged. The count of unique unigrams in the corpus after this processing is approximately 1.8 million.

Discarding a token containing a punctuation mark from n-gram is not a straightforward operation. When such token is discarded, cross punctuation mark context is preserved but the order of n-gram is lowered. To preserve as much cross punctuation mark context as possible, this reduction was done using maximal available n-gram order, i.e. 5-grams reduced to the desired trigrams afterwards and n-grams of resulting order two or lower were discarded. Table 3 shows counts of selected n-gram orders before and after described processing.

Analysis performed on postprocessed trigrams also revealed that a lot of problematic tokens have remained in the corpus, e.g. foreign words (English, Slovak, German and other European languages) or variants of Czech words written without diacritical marks which is frequent practice to avoid problems with character encoding of Czech text over Internet. These tokens has been left untouched in the current level of the post-processing.

2.2 Czech National Corpus SYN2006PUB n-Grams

The above mentioned Web 1T 5-grams corpus was compared with the data obtained from Czech National Corpus [7] (CNC). CNC is a large corpus of written Czech collected within an academic project focused on the building and continuous extension of electronic resources of Czech texts and is supposed to be used as reference corpus in this work.

Limited access to this corpus is available via interactive Web, however, this interface does not provide the level of access needed for multigram generation. Multigrams used in this article as the reference for the analysis of Web 1T 5-grams corpus were obtained on the basis of bilateral agreement. We have the n-gram corpus generated from SYN2006PUB [4], i.e. a synchronic corpus of written journalism of 300 million of words (tokens). This corpus contains exclusively journalist texts from November 1989 to the end of 2004 which were not covered by other similar corpora SYN2000 and SYN2005 [7]. This n-gram corpus contains n-grams of the order one up to five, no cutoff for token or n-gram occurrence counts is used, and it also does not contain punctuation marks. Just one special token for the end of a sentence (</s>) is included. The number of unique unigrams in this corpus is approximately 2.5 million and tokens are case-sensitive. Unique n-gram counts are summarised in Table 2.

As punctuation marks are not present in this corpus, the usage of 5-grams does not bring any improvement in terms of preserving cross punctuation mark context, so trigrams were used for further analyses. Similary to processing Web 1T 5-grams, numeric expressions were also replaced by token <NUM>. In addition, end of sentence token </s> was expanded with pair of tokens for sentence start (<S>) and sentence end (</S>). Resulting 4-gram was splitted in two trigrams which were inserted back. It was done because the tools from HTK Toolkit [8] used in further steps require both these special tokens. Finally, these tokens were uppercased to keep them identical in both n-grams corpora.

3 Comparison of Analysed n-Gram Corpora

For the evaluation of a n-gram corpus various metrics may be used. Unique counts indicate how much diversified the original data are. Apart from comparing previously mentioned n-gram corpora to each other, basic comparison to Czech LC-StarII lexicon [9,10] is also shown.

3.1 Unigram Comparison

Firstly, the statistics of unique n-grams in particular corpora were computed and compared Statistics for SYN2006PUB n-gram corpus were counted with no cutoff, for Czech Web 1T 5-gram with original unchanged cutoff 40. Raw n-gram count in LC-StarII corpus corresponds to n-gram counts after postprocessing (cleaned) of other two n-gram corpora. Count of cleaned unigrams (proper names and abbreviations were deleted) is showed only for reference (Table 5). Results showed that the very high unigram counts in original Czech Web 1T n-gram corpus were reduced approximately 5 times by postprocessing. It yielded to the reduction of unique 5-gram counts to one half. The number of unigram in reference SYN2006PUB n-gram corpus was lowered by one quarter, the number of trigram decreased by 10%.

Comparison of unigram intersections between Web 1T 5-gram, SYN2006PUB 5-gram and LC-StarII corpora is shown in table 6. More than 1/3 of unigrams in Web 1T 5-gram corpus are also present in SYN2006PUB corpus. Almost all of the most common words represented by LC-StarII lexicon are found also in Web 1T 5-gram corpus. This result was expected as LC-StarII lexicon was created with the aim of covering 95% of words from representative Czech texts.

Similar comparison was also done for subsets of most frequent unigrams from Czech Web 1T 5-gram corpus and SYN2006PUB corpus. Subsets of size 60K, 120K,

Table 4. Intersections of unigram in Web 1T 5-gram and CNC SYN2006PUB 5-gram for most frequent unigrams in several limited vocabularies

	Web 1T 5-gram		SYN2006PUB 5-gram		LC-StarII	
	original	cleaned	original	cleaned	raw	filtered
unigram	9,786,424	1,804,682	2,554,028	1,799,005	132,574	84,724
trigrams	117,264,988	47,376,975	189,152,100	170,243,851	-	-
5-grams	103,280,138	51,747,413	302,836,997	302,770,408	-	-

Table 5. Summary of unique n-gram counts in original and cleared n-gram corpus

combination of corpora	count of common unigrams
CNC SYN2006PUB, Web 1T 5-gram, LC-StarII	83,856
CNC SYN2006PUB and Web 1T 5-gram	700,806
CNC SYN2006PUB and LC-StarII	84,587
LC-StarII and Web 1T	83,941

Table 6. Intersection of unigrams between analyzed corpora

vocabulary size	count of common unigrams
60,000	30,273
120,000	62,458
180,000	93,949
240,000	125,173

180K and 240K of most frequent unigrams were compared and counts of common unigrams are collected in Table 4. These subsets are later (Section 3.2) used also to limit vocabulary for language models.

3.2 Perplexity of Bigram and Trigram Models

Quality of language model (LM) is best measured by using LM together with the acoustic model in LVCSR and by measuring the achieved accuracy of recognized text. Another method, based on perplexity computation, quantifies LM quality without application in LVCSR. Perplexity is defined as $PP = 2^{LP}$ where

$$LP = -\frac{1}{N}\sum_{i=1}^{N}\log_2 q(x_i) \tag{1}$$

and it is often explained as mean log probability of each word for a piece q of previously unseen text of N pieces not used in building the language model. We use this measure for our first analysis, as it requires only LM and testing text corpus, however the perplexity does not necessarily tell exactly how well will analyzed LM perform in speech recognition.

For perplexity counting of created language models a subset of transcripts from Czech SPEECON [11,12] corpus was used as test data. The subset contains the total of 148,557 tokens in 14,914 sentences.

Table 7. Perplexity and n-gram count for language model trained from post-processed SYN2006PUB 3-gram corpus with cutoff 6

cut off 1			bigram model		trigram model	
vocabulary size	OOV rate	cut off	perplexity	bigram count	perplexity	trigram count
60,000	10.93%	1	932	17,587,483	928	65,783,406
120,000	6.36%	1	1,226	29,928,179	1,197	96,243,374
180,000	4.18%	1	1,432	35,987,677	1,400	108,421,863
240,000	3.06%	1	1,571	39,963,558	1,521	115,381,912

Table 8. Perplexity and n-gram count for language model trained from post-processed SYN2006PUB 3-gram corpus with cutoff 1

cut off 6			bigram model		trigram model	
vocabulary size	OOV rate	cut off	perplexity	bigram count	perplexity	trigram count
60,000	10.93%	6	939	4,156,790	817	6,733,685
120,000	6.36%	6	1,259	6,016,818	1,075	7,994,323
180,000	4.18%	6	1,483	6,739,430	1,263	8,289,070
240,000	3.06%	6	1,631	7,150,260	1,385	8,420,779

Table 9. Perplexity and n-gram count for language model trained from Czech Web 1T 5-gram corpus

			bigram model		trigram model	
vocabulary size	OOV rate	cut off	perplexity	bigram count	perplexity	trigram count
60,000	17.76%	40	49,507	11,437,940	162,509	35,166,960
120,000	12.14%	40	126,001	14,904,654	255,144	42,125,506
180,000	9.44%	40	153,682	16,522,260	664,866	43,882,665
240,000	7.63%	40	206,831	17,666,327	935,077	45,443,605

Currently, for corpus evaluation purpose, bigram and trigram models were created for several vocabulary sizes (60K, 120K, 180K, 240K). Models from SYN2006PUB 5-gram corpus were created with cutoff 1 and 6. Perplexities, out of vocabulary (OOV) rates and n-gram counts are summarised in tab. 8 and 7. The models from CNC SYN2006PUB show well known and expected pattern. Perplexities of these models are between 928 and 1,631. This is consistent with perplexity measurements from similar corpora presented in [14,15,16].

Exactly the same approach used for creation of models from SYN2006PUB n-gram corpus was used for the creation of models from Web 1T 5-grams corpus. These models were formally created with cutoff 1 (table 9) but effective cutoff is 40, according to cutoff of original corpus. Models created from this corpus have very high perplexity and significantly higher OOV counts than models from SYN2006PUB n-gram corpus (tables 8 and 7). There are several possible explanations and some of the reasons (foreign words, no diacritics in Czech words) have already been mentioned in Section 2.1. The nature of Web might be another reason, i.e. pages with parts of identical code (headers, footers, menus) or almost completely identical as Internet shops pages, where only small part of text might be different but the same common part will unproportionally increase the count of several particular n-grams. It means, that direct usage of these n-grams without further post-processing is not possible.

4 Conclusions

The analysis of recently issued and publicly available Czech Web 1T 5-grams corpus has been presented in this paper. The corpus was compared with the reference CNC SYN2006PUB 5-grams corpus in means of n-gram counts and perplexity computation on language models created from these corpora. The most important conclusions could be summarized as follows:

- The analyzed Web 1T 5-grams corpus seems to be a good source of data for language modelling. It offers large amounts of data from Web resources transformed into n-grams, with stripped (X)HTML markup, unified encoding and basic filtering. In spite of these advantages, it still contains a lot of unsuitable words so further filtering is necessary before the usage in LVCSR.
- Currently achieved perplexity with n-grams from cleaned Czech Web 1T 5-grams was still extremely high, i.e. more than 10^5 in comparison to results for reference SYN2006PUB n-grams corpus where low perplexity between 900 and 1,600 above chosen reference text corpus.

  It means that currently realized post-processing has not been sufficient yet and it should be extended by additional filtering to remove mainly words from foreign languages misspelled words, Czech words without accents or to decrease the influence of (almost) identical n-grams originating from identical page fragments, especially headers or footers.
- Next experiments will be performed with LVCSR using audio data from Czech SPEECON database to see whether the recognition accuracy will be correlated with achieved perplexity measurements.

Acknowledgements

The research was supported by the grant GAČR 102/08/0707 “Speech Recognition under Real-World Conditions” and by research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”. The creation of LC-StarII Czech and Slovak lexica was supported by Harman/Becker, Ulm, Germany which is also the current owner of the collected lexica.

References

1. Nouza, J., Ždánský, J., David, P., Červa, P., Kolorenč, J., Nejedlová, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In: Interspeech 2005, Lisboa, Portugal (September 2005)
2. Pražák, A., Muller, L., Psutka, J.: LIVE TV SUBTITLING – Fast 2-pass LVCSR System for Online Subtitling. In: SIGMAP 2007. INSTICC PRESS, Lisabon (2007)

3. Brants, T., Franz, A.: Web 1T 5-gram, 10 European Languages, version 1. Linguistic Data Consortium, Philadelphia (2009), Web page http://www.ldc.upenn.edu
4. Czech National Corpus: Český národní korpus (Czech National Corpus) – SYN2006PUB. Institute of the Czech National Corpus FF UK, Praha (2006), http://www.korpus.cz
5. Linguistic Data Consortium: Home page (2010), http://www.ldc.upenn.edu
6. Brants, T., Franz, A.: Web 1T 5-gram, version 1. Linguistic Data Consortium, Philadelphia (2006), Web page http://www.ldc.upenn.edu
7. Czech National Corpus: Home page. Institute of the Czech National Corpus FF UK, Praha (2010), http://www.korpus.cz
8. Young, S., et al.: The Hidden Markov Model Toolkit (HTK), Version 3.4.1, Cambridge (2009), http://htk.eng.cam.ac.uk
9. Moreno, A.: LC-StarII. Lexica and Corpora for Speech-to-Speech Translation Components, http://www.lc-star.org
10. Pollák, P., Černocký, J., Smrž, P.: LC-STAR CSCZ. Czech lexicon for ASR and TTS (October 2008), http://www.lc-star.org
11. Pollák, P., Černocký, J.: Czech SPEECON Adult Database (November 2003), http://www.speechdat.org/speecon
12. Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., Kiessling, A.: SPEECON – Speech Databases for Consumer Devices: Database Specification and Validation. In: Proc. of LREC 2002 (May 2002)
13. Young, S., et al.: The HTK Book, Version 3.4.1, Cambridge (2009), http://htk.eng.cam.ac.uk
14. Mikolov, T., Oparin, I., Glembek, O., Burget, L., Karafiát, M., Černocký, J.: Použití mluvených korpusů ve vývoji systému pro rozpoznávání přednášek (Use of spoken corpora in the development of system for recognition of Czéch lectures). In: Proc. of Čeština v mluveném korpusu (Czéch in Spoken Corpus), Praha (2007)
15. Mikolov, T.: Language Models for Automatic Speech Recognition of Czech Lectures. In: Proc. of STUDENT EEICT 2008, Brno (April 2008)
16. Byrne, W., Hajič, J., Ircing, P., Krbec, P., Psutka, J.: Morpheme Based Language Models for Speech Recognition in Czech. In: Proc. of Text, Speech, and Dialog 2000, Brno (2000)