Categorization of Unorganized Text Corpora for better Domain-Specific Language Modeling

Jan STAS, Daniel ZLACKY, Daniel HLADEK, Jozef JUHAR

Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics, Technical University of Kosice,
Park Komenskeho 13, 042 00 Kosice, Slovak Republic

jan.stas@tuke.sk, daniel.zlacky@tuke.sk, daniel.hladek@tuke.sk, jozef.juhar@tuke.sk

Abstract. *This paper describes the process of categorization of unorganized text data gathered from the Internet to the in-domain and out-of-domain data for better domain-specific language modeling and speech recognition. An algorithm for text categorization and topic detection based on the most frequent key phrases is presented. In this scheme, each document entered into the process of text categorization is represented by a vector space model with term weighting based on computing the term frequency and inverse document frequency. Text documents are then classified to the in-domain and out-of-domain data automatically with predefined threshold using one of the selected distance/similarity measures comparing to the list of key phrases. The experimental results of the language modeling and adaptation to the judicial domain show significant improvement in the model perplexity about 19 % and decreasing of the word error rate of the Slovak transcription and dictation system about 5,54 %, relatively.*

Keywords

Language modeling, large vocabulary continuous speech recognition, similarity measure, term weighting, text categorization, topic detection.

1. Introduction

One of the key problems of the text data gathered from the Internet is their thematic heterogeneity. In the case of domain-specific speech recognition and statistical language modeling, these unorganized text data bring into the process of training language models many ambiguities caused by the overestimating such n-gram probabilities that are typically unrelated with the area, in which the speech recognition is performed. Therefore, it is necessary to divide the text data into the predefined domains in the best way as it is possible and adjust the parameters of language modeling for effective and robust task-oriented speech recognition.

With an increasing number of the text documents gathered from the Internet and growing need for more accurate and robust models of the Slovak language [1] for the transcription and dictation system from the judicial domain [2], a question how to categorize the text data according their content arises, considering the fact that one document may contain more than one theme within. This question is getting on importance especially with using unorganized text corpora without any knowledge about the document boundaries in the process of training domain-specific models. Therefore, we were looking for a way of categorizing the text data in unorganized text corpora to the in-domain and out-of-domain data for better language modeling.

Contemporary text categorization is usually based on topic detection with key word identification for categorization of text data into predefined domains [3] or text document clustering based on measuring similarity between two or more documents [4], [5] with using iterative or hierarchical clustering algorithms [6]. Based on this knowledge, we propose an algorithm for text categorization, which classifies short segments (blocks of texts or paragraphs) from unorganized text corpora to the in-domain and out-of-domain data. These data are then used in statistical language modeling for enhancing the quality and robustness of the large vocabulary continuous speech recognition (LVCSR) in Slovak. By combining of several principles, methods and algorithms widely used in text categorization, we propose an effective and unsupervised algorithm that brings a significant improvement of the quality the domain-specific modeling of the Slovak language.

This paper is organized as follows. In the Sec. 2, the text corpora used either for text categorization and statistical language modeling is mentioned. Our proposed approach for text categorization based on key

phrases identification, term weighting, measuring similarity and automatic thresholding is presented in Sec. 3. Sec. 4 describes the speech recognition setup used in experiments that are discussed in the Sec. 5. Main contributions and future directions are summarized at the end of this paper in Sec. 6.

2. Text Corpora

The text data used in the process of text categorization and statistical language modeling was collected using an automatic system for text gathering and processing called webAgent [1], [7]. This system retrieves the text data from various web pages and electronic resources that are written in Slovak. The text data are then filtered from a large amount of grammatically incorrect words, symbols or numerals and normalized into their pronounced form. Finally, the processed text corpora were divided into smaller domain-specific subcorpora, ready for the training language models. Statistics of the number of tokens and sentences for particular text subcorpus is summarized in the Tab. 1.

It is important to say that the judicial corpus was obtained from the Ministry of Justice of the Slovak Republic, in order to develop the automatic transcription and dictation system for their internal purpose [2]. The corpus of fiction was created from a number of electronic books freely available on the Internet.

For morphological analysis we have used Dagger [8], the Slovak morphological classifier based on a hidden Markov model and suffix-based word clustering function. And as it was mentioned before, the preprocessed and morphologically annotated corpora were divided into five domain-specific subcorpora, oriented to the domain of fiction, justice, broadcast news, remaining web and other heterogeneous text.

Tab. 1: Statistics on the text corpora.

Text corpus	# Tokens	# Sentences
corpus of fiction	101 234 475	8 039 739
judicial corpus	565 140 401	18 524 094
broadcast news	554 593 113	36 326 920
web corpus	748 854 697	50 694 708
other text	55 711 674	4 071 165
annotations	4 434 217	485 800
development set	55 163 941	1 782 333
Total	2 085 132 518	119 924 759

3. Proposed Approach

We propose an approach for text categorization of unorganized text corpora (without knowledge about the document boundaries) to the in-domain and out-of-domain data, where the text corpora were segmented into blocks of at least 300 words. This value was determined empirically from statistical observation. In the process of text categorization, we have not considered removal of stop-words because we use key phrases that contain them in the step of key phrase identification (or topic detection). Also stemming or lemmatization would cause a high time and memory requirements, therefore it has not been introduced into the process of text categorization.

3.1. Key Phrases Identification

Based on morphologically annotated corpora, we also proposed a scheme for automatic extraction of multi-word units (key phrases) from judicial domain [9]. Using this approach, we created a list of 5 210 key phrases length from 1 to 4 words. These key phrases were later used in computing frequency of their occurrence in mentioned blocks of 300+ words – text documents. Documents that did not contain any key phrases were automatically classified as out-of-domain text data.

3.2. Vector Space Model

One of the simplest way how to represent the occurrence of words in text documents is to use a multidimensional vector space model, where each $i-th$ document $\vec{d_i}$ is represented by a vector of terms t_j (key words or key phrases) as follows [4]

$$\vec{d_i} = (t_{i,1}, t_{i,2}, \ldots, t_{i,N}). \quad (1)$$

In our case, each document was represented by a vector of 5 210 key phrases. Average number of words in one document after segmentation into blocks of 300+ words was 332. Then, the total number of text documents was 6 908 655. Main disadvantage of such representation is very high dimension and redundancy which result in high requirements on disc space.

3.3. Term Weighting

Key phrases in vector space model are represented by theirs occurrence in text documents. This value is often normalized considering its occurrence in the set of all examined text documents. Based on previous research [6], the term frequency and inverse document frequency (TF-IDF) are usually used for term weighting. TF-IDF weighting function is computed as [4]

$$w_{i,j} = tf_{i,j} \cdot idf_i = \frac{f_{i,j}}{\sum_k f_{k,j}} \cdot \log \frac{N}{|\{j : t; \in d_j\}|}, \quad (2)$$

where $f_{i,j}$ is the frequency of occurrence of term t_i in document d_j. The sum in the denominator of $tf_{i,j}$ component expresses the frequency of occurrence of all

terms in document d_j, N is the total number of documents and denominator $|.|$ of idf_i component expresses the total number of documents that contain term t_i.

Beside standard TF-IDF, there are many other term weighting schemes such as TF-ICF, in which ICF component is computed on limited data set, or ATC, LTU and Okapi weighting schemes, which can use additional attributes giving information about the document, for example of its length [10].

3.4. Measuring Similarity

The next step includes measuring distance between two documents. In our case, we measure similarity between reference text represented by a list of 5 210 key phrases and examined text document (hypothesis). Both text documents are transformed to the vector space model and weighted by TF-IDF scheme, so they could be compared. The weight of each key phrase in reference text was computed on development data set, which contain about 10 % of texts from the judicial corpus that were not used in training language model. By comparative study of distance/similarity metrics described in [11], we have chosen three measures, which satisfy a condition of: a. non-negativity; b. symmetricity; c. triangle inequality; and d. identity, when distance is equal to zero; namely the Bhattacharyya coefficient, Jaccard index and Jensen-Shannon divergence.

The Bhattacharyya coefficient is often used for clustering phonemes in the training acoustic models in LVCSR. This coefficient expresses the relative accuracy of the estimate the probability between two density functions and comes from the sum of the geometric mean of these probabilities. It specifies the separability of two classes x and y and is used as classification criterion as follows

$$d_{Bha} = -\ln \sum_{i=1}^{N} \sqrt{x_i y_i}, \tag{3}$$

The Jaccard correlation index expresses scalar sum of two vectors and is usually used for measuring similarity between two probability density functions. It comes from harmonic mean and the equation for computing cosine similarity, normalized by absolute deviation of two probability distributions according formula

$$d_{Jac} = \frac{\sum_{i=1}^{N}(x_i + y_i)^2}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} x_i y_i}. \tag{4}$$

And finally, the Jensen-Shannon divergence comes from the principle of uncertainty. This measure is a special case of averaged Kullback-Leibler divergence (also relative entropy), which satisfies the symmetry in the entire range of values. It is often used in information theory and natural language processing. Jensen-Shannon divergence is computed as

$$d_{JS} = \frac{1}{2}\left[\sum_{i=1}^{N} x_i \ln\left(\frac{2x_i}{x_i + y_i}\right) + \sum_{i=1}^{N} y_i \ln\left(\frac{2y_i}{x_i + y_i}\right)\right]. \tag{5}$$

Note that each measure is represented as a distance for better implementation in our algorithm for text categorization. There exists a number of other distance/similarity measures used in text categorization, such as cosine similarity or Pearson correlation coefficient [5], which were omitted from observation because of similarity with Jaccard index or its asymmetricity.

3.5. Automatic Thresholding

The last step in text categorization is appropriately setting the threshold, when text data appertain to the in-domain or out-of-domain area. This value is usually determined empirically from long-term observation or automatically from examined statistic values. We used automatic thresholding based on the calculation of median of a sequence of coefficients derived from computing one of the similarity measure described in the previous section. Threshold values were calculated on development set and used in the algorithm for text categorization to the in-domain and out-of-domain data.

4. LVCSR Setup

Trigram language models were created using SRILM Toolkit [12] with vocabulary size of 325 555 of unique words. All models have been trained on the text corpora size of about 2 billion of tokens (see Sec. 2) and smoothed by the Witten-Bell back-off algorithm. Particular in-domain and out-of-domain language models were combined and adapted into the judicial domain using linear interpolation with computing interpolation weights based on minimization of perplexity on a development data set using our proposed algorithm [1].

The triphone context-dependent acoustic model (AM) based on hidden Markov models (HMMs) have been used. Each of 4 states of the AM had been modeled by 32 Gaussian probability density functions. The model has been generated from feature vectors containing 39 mel-frequency cepstral (MFC) coefficients using HTK Toolkit [13]. It has been trained on 120 h of readings of real adjudgments from the court, 130 h of read phonetically rich sentences, newspaper articles, and spelled items, recorded in offices and conference rooms and 100 h of spontaneous speech recorded at council hall. The acoustic database is characterized by gender-balanced speakers and contains read and spontaneous speech [2]. For modeling rare triphones the effective triphone mapping algorithm has been used [14].

For decoding, we have used the LVCSR engine Julius based on two-pass strategy, where input data are processed in the first pass with bigram LM and the final search for reverse n-gram is performed again using the result of the first pass to narrow the search space [15].

Test data were represented by 315 min. of recordings obtained from randomly selected speech segments from the acoustic database of judicial proceedings. These segments contain 41 820 words in 3 462 sentences that were not used in the training AM.

For evaluation, the word error rate (WER) and model perplexity (PPL) has beeen used. WER is the standard extrinsic measure of performance the LVCSR system, computed by comparing the reference text read by a speaker against the recognized results and it takes into account insertion, deletion and substitution errors. For intrinsic evaluation of the LMs the model perplexity has been used. PPL is defined as the reciprocal of the weighted (geometric) probability assigned by the LM to each word in the test set.

5. Experimental Results

Experiments have been oriented on the evaluation of model perplexity and word error rate of the LVCSR system after text categorization to the in-domain and out-of-domain data and statistical modeling of the Slovak language from the judicial domain.

Statistics of the number of text documents after term weighting, measuring the distance/similarity between examined documents and weighted list of key phrases and categorization of the text to the in-domain and out-of-domain data we can see in the Tab. 2.

We focused on two types of experiments. In the first experiment, proposed text categorization has been performed within particular text subcorpora apart, when the text data after categorization were merged to the two subcorpora: in-domain and out-of-domain corpora.

In the second experiment, particular text subcorpora described in the Sec. 2 were merged into one text corpus and text categorization has been performed on the entire corpus without any knowledge about the topic in every text document. The result of model perplexity after adaptation to the judicial domain and combination of in-domain and out-of-domain trigram models to the final domain-specific model of the Slovak language and word error rate of the Slovak transcription and dictation system is summarized in the Tab. 3.

As we can see from these results, the best class separation was achieved with Bhattacharyya coefficient as similarity measure. On the contrary, the worst class separation in both experiments was noticed for Jensen-Shannon divergence. Better class separation was observed when text categorization has been performed in particular text corpora that had a positive impact on the model PPL or WER of the LVCSR system. This fact is caused by using TF-IDF weighting, or its IDF component, computed for the entire set of documents, while the first experiment has been evaluated for not equal numbers of documents in particular subcorpora. Therefore, text categorization is more accurate when all text documents are concentrated in one corpus.

The results of model perplexity and WER of the LVCSR system for these two types of experiments are moderately different. However, the best distance/similarity measure in connection with TF-IDF weighting in proposed text categorization appears Jaccard index. This approach brought significant decrease of the model perplexity in modeling of the Slovak language approximately 19 % and WER of the Slovak transcription and dictation system about 5,54 %, relatively against language modeling on unorganized text corpora and approximately 11 % in model perplexity and 3,47 % in WER, relatively against previous categorization of text data based on rough categorization using the URL of downloaded text document or other additional information about it [1], [2].

6. Conclusion

This article was focused on design of the algorithm for categorization of unorganized text corpora to the in-domain and out-of-domain data with the aim of increasing the quality and robustness of language modeling in large vocabulary continuous speech recognition. By combining effective methods for topic detection based on most frequented key phrases, term weighting, measuring similarity between hypothesis and reference text and automatic thresholding, we have achieved significant improvement in modeling of the Slovak language, both in the model perplexity and in the word error rate values. Further research should be focused on application of more effective term weighting schemes such as ATC, LTU, Okapi, or TF-ICF that would not be evaluated through the entire text corpora to reduce the time and memory requirements of the algorithm for the text categorization.

Acknowledgment

The research presented in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project MS SR 3928/2010-11 (25 %) and Research and Development Operational Program funded by the ERDF under the projects ITMS-26220220155 (25 %) and ITMS-26220220141 (50 %).

Tab. 2: Statistics of the number of text blocks after text categorization.

Measure	Partially		Together	
	In-domain	**Out-of-domain**	**In-domain**	**Out-of-domain**
Bhattacharyya coefficient	684 598	6 224 057	1 166 805	5 741 850
Jaccard correlation index	969 853	5 938 802	1 258 169	5 650 486
Jensen-Shannon divergence	1 811 093	5 097 562	2 305 230	4 603 425

Tab. 3: Model perplexity and word error rate of the Slovak transcription and dictation system for the judicial domain.

Measure	Partially		Together	
	PPL	**WER** [%]	**PPL**	**WER** [%]
without text categorization	40,4302	5,48	44,3262	5,60
Bhattacharyya coefficient	42,0395	5,57	36,0428	5,32
Jaccard correlation index	41,2654	5,53	**35,9444**	**5,28**
Jensen-Shannon divergence	42,7054	5,66	38,1756	5,50

References

[1] JUHAR, J., J. STAS and D. HLADEK. Recent Progress in Development Language Model for Slovak Large Vocabulary Continuous Speech Recognition. In: *New Technologies – Trends, Innovations and Research*. Rijeka: InTech, 2012, pp. 261–276. ISBN 978-953-51-0480-3. DOI: 10.5772/32623.

[2] RUSKO, M., J. JUHAR, M. TRNKA, J. STAS, S. DARJAA, D. HLADEK, M. CERNAK, M. PAPCO, R. SABO, M. PLEVA, M. RITOMSKY and M. LOJKA. Slovak Automatic Transcription and Dictation System for the Judicial Domain. In: *Human Language Technologies as a Challenge for Computer Science and Linguistics: 5th Language & Technology Conference, 2011*. Poznan: Fundacja Uniwersytetu im A. Mickiewicza, 2011, pp. 365–369. ISBN 978-83-932640-1-8.

[3] YUE, L., S. XIAO, X. LV and T. WANG. Topic Detection based on Keyword. In: *Proc. of 2011 International Conference on Mechatronic Science, Electric Engineering and Computer*. Jilin: IEEE, 2011, pp. 464–467. ISBN 978-1-61284-719-1. DOI: 10.1109/MEC.2011.6025502.

[4] MANNING, Ch. D. and H. SCHUTZE. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999. ISBN 02-621-3360-1.

[5] HUANG, A. Similarity Measures for Text Document Clustering. In: *Proc. of the 6th New Zealand Computer Science Research Student Conference*. Christchurch, New Zealand, 2008, pp. 49–56.

[6] ZLACKY, D., J. STAS, J. JUHAR and A. CIZMAR. Slovak Text Document Clustering. *Acta Electrotechnica et Informatica*. 2013, vol. 13, no. 2, will be published. ISSN 1338-3957.

[7] HLADEK, D. and J. STAS. Text Mining and Processing for Corpora Creation in Slovak Language. *Journal of Computer Science and Control Systems*. 2010, vol. 6, iss. 1, pp. 65–68, ISSN 1844-6043.

[8] HLADEK, D., J. STAS and J. JUHAR. Dagger: The Slovak Morphological Classifier. In: *Proc. of the 54th International Symposium ELMAR 2012*. Zadar: IEEE, 2012, pp. 195–198. ISSN 1334-2630. ISBN 978-1-4673-1243-1.

[9] STAS, J., D. HLADEK, J. JUHAR and M. OLOSTIAK. Automatic Extraction of Multiword Units from Slovak Text Corpora. In: *Proc. of the 7th International Conference on NLP, Corpus Linguistics, SLOVKO 2013*. Bratislava: E/Learning, 2013, will be published.

[10] REED, J. W., Y. JIAO, T. E. POTOK, B. A. KLUMP, M. T. ELMORE and A. R. HURSON. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. In: *Proc. of the 5th International Conference on Machine Learning and Applications*. Orlando: IEEE, 2006, pp. 258–263. ISBN 0-7695-2735-3. DOI: 10.1109/ICMLA.2006.50.

[11] CHA, S. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*. 2007, vol. 1, no. 4, pp. 300–307. ISSN 1998-0140.

[12] STOLCKE, A. SRILM – An Extensible Language Modeling Toolkit. In: *Proc. of ICSLP*. Denver: ISCA, 2002, pp. 901–904. ISBN 1876346450.

[13] YOUNG, S., G. EVERMANN, M. GALES, T. HAIN, D. KERSHAW, X. LIU, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV and P. WOODLAND. *The HTK Book (for HTK Version 3.4)*. Massachusetts: Cambridge University Engineering Department, 2006. ISBN 978-1-4419-7712-0.

[14] DARJAA, S., M. CERNAK, M. TRNKA, M. RUSKO and R. SABO. Effective Triphone Mapping for Acoustic Modeling in Speech Recognition. In: *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*. Florence: DBLP, 2011, pp. 1717–1720. ISBN 9781618392701.

[15] LEE, A., T. KAWAHARA and K. SHIKANO. Julius - an Open Source Real-Time Large Vocabulary Recognition Engine. In: *7th European Conference on Speech Communication and Technology, EUROSPEECH 2001*. Aalborg: ISCA, 2001, pp. 1691–1694. ISBN 978-8-7908-3410-4.

About Authors

Jan STAS was born in Bardejov, Slovakia in 1984. In 2007 he graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He received his Ph.D. degree at the same department in the field of Telecommunications in 2011. He is currently working as a post-doctoral researcher with focus on natural language processing and understanding, computational linguistics and statistical language modeling in speech recognition and is author of several conference and journal papers from this area.

Daniel ZLACKY was born in Poprad, Slovakia in 1988. He received his M.Sc. (Ing.) degree in the field of Telecommunications in 2012 at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He is currently Ph.D. student at the same department in the field of Telecommunications. His research interests include automatic word segmentation, text categorization, text document clustering and statistical language modeling in speech recognition and is author of several conference and journal papers from this area.

Daniel HLADEK was born in Kosice, Slovakia in 1982. In 2006 he graduated M.Sc. (Ing.) at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He has obtained Ph.D. degree in 2009 at the same department in the field of Computational Intelligence. He is currently working as a post-doctoral researcher at the Department of Electronics and Multimedia Communications at the Technical University of Kosice with focus on natural language processing, speech and audio processing and intelligent decision methods. He is author of several conference and journal papers from this area.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as a Full Professor at the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech and speaker identification and verification, speech synthesis and development in spoken dialogue and speech recognition systems.