The Use of Semantic Word Classes in Document Classification

Stevan Ostrogonac,*† Branislav Popović,*† Milan Sečujski*†

* Faculty of Technical Sciences, University of Novi Sad
Trg Dositeja Obradovića 6, 21000 Novi Sad

†AlfaNum – Speech Technologies Ltd,
Polgar Andraša 38a/61, 21000 Novi Sad
ostrogonac.stevan@uns.ac.rs
bpopovic@uns.ac.rs
secujski@uns.ac.rs

1. Introduction

Document classification and topic modeling represent some of the biggest challenges in the fields of natural language processing and information retrieval. Many of the techniques developed for these purposes are language-independent (Sanderson and Bruce Croft, 2012). However, language resources are needed for each language, along with domain-specific data sets for particular applications, and every new language introduces a specific set of problems. In this paper, a method for addressing the problem of data sparsity in document classification for under-resourced, highly inflective languages, is proposed. The case of Serbian is considered, but the method is applicable to other languages as well. The approach includes training a language model on a large textual corpus, using it to create semantic word classes and using the extracted semantic information to obtain a more robust document classifier. As a topic model, Latent Dirichlet Allocation can be used, as well as its variants or other types of topic models.

2. Semantic Information Extraction

The method for semantic word class derivation has been described in a previous research (Ostrogonac et. al, 2015) and here will be described briefly. A textual corpus for Serbian, which contains over 20 million word tokens, which correspond to around 360 thousand word types, 180 thousand lemmas and around 1000 morphologic classes (Ostrogonac et. al, 2012) was used to train a language model (LM). The LM was lemma-based, since morphologic information was available for Serbian (Sečujski, 2002) and could be restored after semantic lemma classes were derived. The semantic classes were created by applying a greedy clustering algorithm (Mikolov, 2012) to the lemmatized textual corpus, which was based on lemma colocation as a basis for determining semantic similarity measure. The clustering algorithm leans on the probabilities obtained from the LM for hypotheses created by replacing a lemma with other lemmas from the dictionary. The lemmas for which the replacement causes the smallest change in probabilities are likely to be semantically similar to the original word. After the entire corpus is processed, and morphologic information is restored to derive words from lemmas, semantic word classes are created. The parameters for clustering should be fine-tuned by iteratively observing the results and adjusting the values so that the classes are optimized for a particular application. A semantic class can, therefore, represent only synonyms, but it can also represent all the words that can be placed in certain positions within sentences and result in semantically correct sentences, or it can represent something in between.

3. Semantic Word Classes in LDA

An LDA is a generative model which can be used for document classification (Blei et. al, 2003). One of the most popular document classification tasks is e-mail classification into regular messages and spam, which will be used in the following text in order to illustrate the effect of semantic word clustering. In LDA, a document is considered to be a mixture of a number of topics, which is similar to the bag of words (Mikolov, 2012) concept. Each word may belong to many topics, to each with a certain probability. In order to define those probabilities and the topics themselves, a great amount of data is needed. The main problem is that two spam messages can contain similar or the same topics, but consist of very different sets of words. For example, two spam messages containing the same advertisement may contain corresponding sentences such as “Buy now at lower price and enjoy the trip!” and “Purchase immediately, experience an exciting travel with our discount!”. This problem is emphasized in highly inflective languages. The lack of data results in poor classifiers. However, even though textual data of specific content may not be enough to train highly accurate classifiers, other textual resources can be used to obtain additional information. Semantic classes derived from

a large textual corpus which contains many different types of documents can be used to make a document classifier more robust. By using semantic class IDs instead of words, an LDA can model topics quite well even with a small amount of application-specific data, since for each word that is observed within the training data set, an entire semantic class is included in the modeling process. Semantic word clustering described in Section 2 insures that words with the same meaning but different morphological features are grouped together and therefore eliminates morphology as a cause of data sparsity in topic modeling. However, a morphologic dictionary is not available for all inflective languages. In those cases, other methods may be used to deal with this problem. Semantic classes may be grouped manually, or by applying a rule-based approach including word-stem derivation, for which the implementation is language-dependent. Other suboptimal solutions may be used as well. Furthermore, semantic classes include words with similar meaning, which reduces the number of topics to be modeled, resulting in more accurate topic representations. Spam detection is a fine example of how a classification process can benefit from semantic information extracted from an external source, but the application of the described approach is far more broad and includes other information retrieval tasks.

4. Further Research and Application

Semantic word clustering itself can be improved by implementation of a probabilistic approach, meaning that words would belong to more than one semantic class with specific corresponding probabilities. Furthermore, even though semantic classes obtained in the described way may contain words with similar meaning, no information about the correlation between the classes is extracted. For example, semantic class A = {malaria, flu, meningitis, AIDS, cancer…} and class B = {drug, medicine, therapy, cure, pill…} are highly semantically correlated, but this information is not extracted. Obtaining this higher-level semantic information requires wider context analysis, which will be the main topic of further research.

The applications of the extracted semantic information are numerous and represent the basis for creation of advanced dialogue systems, which would be able to mimic natural dialogue. The most important pursuit in this area would be to develop the possibility of determining the menaning of a word that a dialogue system has not seen before.

5. Acknowledgements

The work described in this paper was supported in part by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the project TR32035: “Development of Dialogue Systems for Serbian and Other South Slavic Languages”.

6. References

David M. Blei, Andrew Y. Ng, Michael I. Jordan, “*Latent Dirichlet Allocation*”, Journal of Machine Learning Research 3 (4–5): pp. 993–1022, January 2003. doi:10.1162/jmlr.2003.3.4-5.993

Mark Girolami, A. Kaban, “*On an Equivalence between PLSI and LDA*”, Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3

Mark Sanderson, W. Bruce Croft, “*The History of Information Retrieval Research*”, Proceedings of the IEEE 100: 1444–1451, 2012. doi:10.1109/jproc.2012.2189916

Milan Sečujski, “*Accentuation Dictionary for Serbian Intended for Text-to-Speech Technology*”, Proceedings of DOGS, pp.17-20, Novi Sad, Serbia 2002.

Stevan Ostrogonac, Branislav Popović, Robert Mak, Milan Sečujski: “*Automatic Word Clustering Based on Semantics - an Approach for Serbian*”, 3rd International Acoustics and Audio Engineering Conference, TAKTONS 2015, Novi Sad, Srbija: Radio-televizija Vojvodine, Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Srpska sekcija AES (Audio Engineering Society), Dirigent Acoustics, Beograd, 18-21. novembar 2015, pp. 36-37, ISBN: 978-86-7892-758-4.

Stevan Ostrogonac, Dragiša Mišković, Milan Sečujski, Darko Pekar, Vlado Delić: “*A Language Model for Highly Inflective Non-Agglutinative Languages*”, 10. SISY, International Symposium on Intelligent systems and Informatics, Subotica: IEEEXplore, 20-22.09.2012, ISBN: 978-1-4673-4749-5, pp. 177-181.

Tomáš Mikolov, “*Statistical language models based on neural networks*”, in PhD Thesis, Brno University of Technology, 2012.

Xiaogang Wang, Eric Grimson, “*Spatial Latent Dirichlet Allocation*”, Proceedings of Neural Information Processing Systems Conference (NIPS), 2007.