Explicit Semantic Analysis as a Means for Topic Labelling

Anna Kriukova[1(✉)], Aliia Erofeeva[2], Olga Mitrofanova[1], and Kirill Sukharev[3]

[1] St. Petersburg State University, St. Petersburg, Russia
krukova.ann@gmail.com, oa-mitrofanova@yandex.ru
[2] University of Trento, Trento, Italy
amirzagitova@gmail.com
[3] St. Petersburg Electrotechnical University, St. Petersburg, Russia
sukharevkirill@gmail.com

Abstract. This paper deals with a method for topic labelling that makes use of Explicit Semantic Analysis (ESA). Top words of a topic are given to ESA as an input, and the algorithm yields titles of Wikipedia articles that are considered most relevant to the input. An alternative approach that serves as a strong baseline employs titles of first outputs in a search engine, given topic words as a query. In both methods, obtained titles are then automatically analysed and phrases characterizing the topic are constructed from them with the use of a graph algorithm and are assigned with weights. Within the proposed method based on ESA, post-processing is then performed to sort candidate labels according to empirically formulated rules. Experiments were conducted on a corpus of Russian encyclopaedic texts on linguistics. The results justify applying ESA for this task, and we state that though it works a little inferior to the method based on a search engine in terms of labels' quality, it can be used as a reasonable alternative because it exhibits two advantages that the baseline method lacks.

Keywords: Topic labels · Topic modelling
Explicit Semantic Analysis · Russian

1 Introduction

One of the most claimed approaches in contemporary computational semantics is topic modelling which describes a corpus in terms of latent topics and reveals the distribution of documents over topics. Being a variety of fuzzy clustering, such representation characterises text semantics and effectively depicts a structure of large document collection. Researchers have developed various types of topic models, preferences being given to Probabilistic Latent Semantic Analysis (pLSA) [5] and Latent Dirichlet Allocation (LDA) [3]. In this study we focus our attention to LDA which is a generative model consisting of two stages: (1) distribution θ_d of documents d over topics t in a collection D is defined; (2) topic ϕ_t

D. Ustalov et al. (Eds.): AINL 2018, CCIS 930, pp. 110–116, 2018.
https://doi.org/10.1007/978-3-030-01204-5_11

for a word w in a document d is chosen in accordance with distribution θ_d. It is supposed that θ_d and ϕ_t conform with distributions $Dir(\alpha)$ and $Dir(\beta)$ where α and β are considered as hyperparameters of Dirichlet allocation. In practice, the number of topics and their size are defined by users in course of experiments.

Generated topics are standardly presented to end-users as a list of the top n terms from the multinomial distribution of words ranked by the probability $Pr(w|\phi_t)$. However, this often hinders the understanding of a topic by human readers, especially when the selected number of topics is large. Aimed to reduce the cognitive load of interpreting such topics, the task of automatic topic labelling emerged [10], i.e. the task of finding a concise and salient label that describes the content of a given topic. The problem has been studied extensively for English, and there have been proposed numerous methods of topic labelling, varying in label modality (words [9] and phrases [2,8,10], or images [1], or both [16]), label generation (relying only on the content of the modelled corpus [6,9,10] or involving external resources [2,8,16]), and algorithms employed (broadly, supervised [1,8,16] or unsupervised [2,6,9,10]).

This line of research is still actively developing in Russian NLP: a label that generalises words of a topic would make its interpretation substantially easier. Our project continues experiments in this field. Therefore, the main task of our paper is to perform comparative analysis of two algorithms adjusted for topic labelling: Explicit Semantic Analysis which relies upon external knowledge from Wikipedia, and Graph-based topic labelling which implies label extraction from a search engine output.

2 Graph-Based Topic Labelling

As a strong baseline topic labelling algorithm, we take the unsupervised graph-based method first introduced for English [2] and proved to be applicable to Russian [12,13]. In the following, we briefly describe the procedure.

At the initial stage of candidate generation, the first k topic words are used as a single query to a search engine. The titles of the top n search results are stripped from stopwords and concatenated into a continuous synthetic text which is then lemmatised and fed into the TextRank [11] ranking algorithm. The text is transformed into an oriented graph $G = \{V, E\}$, where V is a set of nodes representing lemmata, E is a set of weighted edges defined by some similarity metric, e.g. the co-occurrence frequency within the input text. Next, the TextRank value is recursively computed for each node based on the in- and out-degrees [11]. Nodes (words) having higher scores are assumed to be more salient, while edges with larger weights indicate a stronger semantic association between the corresponding word pairs.

In order to move from single tokens to higher level n-grams, the algorithm has been tailored for Russian by applying a set of manually crafted morphological patterns to extract grammatically valid key phrases [12]. At the stage of candidate selection, having each lemma assigned a TextRank score allows ranking the phrases according to the sums of weights of the constituent words.

3 Explicit Semantic Analysis

The approach we concentrate on in this paper makes use of the algorithm for constructing topic labels described in Sect. 2, except that we employ Explicit Semantic Analysis (ESA) [4] rather than a search engine as a way for using external knowledge sources. ESA is a way of representing words and texts in a vector space. ESA makes use of a large collection of documents as a knowledge source: the authors of the initial paper carried out experiments on Wikipedia[1] and described how ESA can be used for both mono- and cross-lingual tasks. Wikipedia is an open and constantly growing source of Russian texts (the Russian Wikipedia contains now almost 1.5 mln articles). Wikipedia articles in ESA are treated as *concepts*, because each article is supposed to describe in detail a single topic. The algorithm deploys the "bag-of-words" approach for representation of concepts, which is often used for NLP problems. Though being a simplification of real-world intertextual relations, it is justified here by the fact that we can usually describe any concept by means of separate words associated with it. Each concept therefore is described by a vector that contains words co-occurring in the corresponding article. Words are assigned with TF-IDF [15] weights that reflect association strength. ESA thus represents text meaning in terms of a weighted concept vector, sorted according to the relevance of concepts to the text. At this point an inverted index is created, that bounds a word with concepts where it occurs. If a concept's weight for a given word is too small, the concept is deleted from the interpretation vector, which allows us to eliminate insignificant links between words and concepts. The intuition behind such representations of text semantics is that in this way we get the most important concepts related to a text and can represent its meaning with their help.

Summing up, ESA represents meaning of a text in a high-dimensional space of concepts derived from Wikipedia. A vector for topic words contains TF-IDF values, i.e. figures; however, as we know which number refers to which article, we can now get a list of articles' titles sorted in the descending order by their weights. The titles are then processed in the same way as search engine results in the previously described algorithm.

4 Experiments

Experiments were performed on topics obtained from a corpus of Russian encyclopaedic texts on linguistics [12]. Size of the corpus, containing more than 1.3 million tokens before pre-processing, reduced to 934,855 tokens after lemmatisation and removing of stop words, digits, and punctuation. We extracted 20 topics using the LDA model from *scikit-learn* [14], with default settings.

Top 10 words from each topic were used as an input for ESA and Yandex[2] search engine, and given the titles of 30 most relevant Wikipedia articles from ESA and search results from Yandex, we then applied the procedure described

[1] https://www.wikipedia.org.
[2] https://yandex.ru.

Table 1. Example of ESA output, titles discarded in post-processing are shown in red.

Topic: время лексема реконструкция том часы число средний антоним фигура использоваться ("time lexeme reconstruction volume hours number neutral antonym figure being_used")	Лексема, Реконструкция, 48 часов, Часы (значения), GiST, Токен, Мариус (значения), T-15 ("Lexeme, Reconstruction, 48 hours, Hours (disambiguation), GiST, Token, Marius (disambiguation), T-15")

in Sect. 2: we created a graph and weighted the candidate labels using the TextRank algorithm. The methods based on the search engine and on ESA will be referred to below as Labels-Yandex and Labels-ESA, respectively, for the sake of convenience.

The initial results provided by Labels-ESA turned out, however, to be rather noisy. First of all, after analyzing intermediate ESA outputs (when it is presented with topic words), we decided to exclude the following article titles (ref. Table 1):

1. dedicated to people, as names would not likely make a meaningful label;
2. containing numbers, for numbers are not supposed to serve as a topic label;
3. containing words not in Russian, since they are later deleted as stop-words;
4. whose length is less than 3 symbols (e.g. articles about alphabet letters);
5. containing the mark "(значения)" ("disambiguation"), as such titles refer to pages with links to other articles.

We also revealed some characteristic features of ESA, probably because of which good labels did not end up at top positions in labels lists and the algorithm needed some enhancement. Firstly, top Wikipedia articles that are delivered by ESA sometimes seem to characterise only few topic words out of ten. For example, if a topic contains the word "диалект" ("dialect"), the first articles describe only different kinds of dialects. That is connected with the manner how a text vector is formed within the ESA approach. ESA combines vectors for each word in a text, i.e. ten vectors for topic words in our case. Thus, if a particular topic word has high TF-IDF values for its articles, they tend to outweigh other words articles in the text vector resulting in its being almost the same as for this only word.

Secondly, ESA finds hyponyms of words more likely than hyperonyms. For example, for a word "гласный" ("vowel") ESA would find articles like "гласныйпереднего ряда верхнего подъёма" ("high front vowel"), rather than articles like "звуки" ("sounds") or "фонетика" ("phonetics"). It takes place as in specific articles, words we are trying to characterise are normally mentioned more often than in general articles, whereas for making topic labels, hyperonyms are more likely required. Thirdly, it is homonymy and polysemy. ESA does not take it into account when searching for most relevant articles because words in the Wikipedia dump are not provided with such information. Thus, some articles in the output can actually be connected with other domains.

Taking all these into account, we decided to manually write rules that would rearrange lists with 20 first labels by Labels-ESA so that the relevant ones would be drawn up. The following post-processing rules were generated empirically:

Table 2. Top-3 labels assigned to some of the topics by Labels-ESA and Labels-Yandex.

Topic	Labels-ESA	Labels-Yandex
лингвистика язык наука лингвистический теория метод исследование анализ идея год ("linguistics language science linguistic theory research analysis idea year")	история лингвистики, морфологический анализ, словари лингвистических терминов ("history of linguistics, morphological analysis, dictionaries of linguistic terms")	методы лингвистического анализа, методология лингвистического анализа, лингвистический анализ ("methods of linguistic analysis, methodology of linguistic analysis, linguistic analysis")
форма глагол язык вид значение время действие наклонение наречие иметь ("form verb language aspect meaning tense act modality adverb have")	спряжение глаголов, наклонение, лингвистика ("verb conjugation, modality, linguistics")	категория наклонения глагола, категория наклонения, наречие слова категории ("category of verb modality, category of modality, adverb category words")

1. If a two-word label is a part of a longer label, the latter is excluded and the former is moved to the first place (e.g. "statistical machine translation" → "machine translation").
2. (a) If more than five labels contain the same noun, all of them are deleted and the noun in plural form is placed at the first position (e.g. the word "dialects" replaces different kinds of dialects).
   (b) If labels contain adjectives from the corresponding topic, we add the most frequent adjective to the noun from the previous step and also move the resulting label to the first place in the labels list.
3. If a label contains more than three words, it is moved back by two positions.
4. If a label contains more than one word and an adjective from the corresponding topic, it is placed at the first position.

The rules proved to considerably improve output of Labels-ESA. Some topics and their top three labels can be seen in Table 2.

5 Evaluation and Analysis

First of all, in the task of topic labelling there is a problem with evaluation, because no gold standard is usually available. Therefore, to evaluate the results we asked six experts to rate obtained labels manually. The experts, students at the department of mathematical linguistics at the St Petersburg State University, were to look through the first ten words for each topic and choose which of the top label from Labels-ESA and one the label from Labels-Yandex matches the corresponding topic better. It was also allowed to mark both or none of the labels. A part of the assessment can be seen in Table 3. We computed the mean value for each method (considering each plus as 1 and each minus as 0), which turned out to be 0.47 for Labels-ESA and 0.54 for Labels-Yandex.

Table 3. Evaluation examples of Labels-ESA and Labels-Yandex, in respective order.

Topic: перевод словарь текст система компьютерный машинный язык прикладной русский ("translation dictionary text system computational machine language applied Russian")	
машинный перевод ("machine translation")	+ − + + ++
система компьютерного перевода ("computer translation system")	+ + + − −+
Topic: язык литературный диалект русский говор современный норма немецкий ("language literary dialect Russian accent modern norm German")	
диалекты ("dialects")	+ + + + ++
диалекты немецкого языка ("German dialects")	− − − − −+
Topic: лингвистика язык наука теория метод исследование анализ идея год ("linguistics language science theory method research analysis idea year")	
история лингвистики ("history of linguistics")	− − + − −−
методы лингвистического анализа ("methods of linguistic analysis")	+ + − + ++

Although Labels-ESA is assessed a little worse, there are several reasons why using ESA instead of a search engine may be beneficial. First of all, when trying to automatically obtain titles from a search engine, one can run into a problem that there is often a limit set up on the number of queries per minute from one IP address when addressing a search engine automatically. Consequently, it takes much more time to find titles for a topic and assign it with a label. ESA has no such limitation and processes a ten-word topic at about 1.2 s. Secondly, search systems usually use complex algorithms for ranking pages, and, what is more, can individualize results for a certain user, which is why experiments of a method relying on a search engine may be hard to reproduce. In case of the ESA approach, we make use of a certain Wikipedia dump, so that the results remain consistent.

Both these characteristics let us regard ESA as a reasonable alternative to the baseline method.

6 Conclusions

In this work we propose to use Explicit Semantic Analysis as a means for dealing with automatic topic labelling. ESA has only recently been adopted for the Russian language and has yet been used for measuring degree of texts' semantic relatedness [7], while we describe its advantages and drawbacks with regards to topic labelling.

We compared our method, based on ESA, with an alternative algorithm that uses titles of first outputs in a search engine, given topic words as a query [12]. The work of both of them was evaluated on topic models extracted from a corpus of Russian encyclopaedic texts on linguistics. The evaluation procedure showed that our method works almost as well as the alternative algorithm, whereas it has a number of significant advantages. Future work will address assessing results on other corpora and domains to prove that post-processing of Labels-ESA output provides equally good results on different kinds of data.

References

1. Aletras, N., Mittal, A.: Labeling topics with images using a neural network. In: Jose, J.M., et al. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 500–505. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_40
2. Aletras, N., Stevenson, M., Court, R.: Labelling topics using unsupervised graph-based methods. In: Proceedings of the 52nd Annual Meeting of ACL, pp. 631–636. ACL (2014). https://doi.org/10.3115/v1/P14-2103
3. Blei, D., Ng, A., Jordan, M.L.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.1162/jmlr.2003.3.4-5.993
4. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007). https://dl.acm.org/citation.cfm?id=1625535
5. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999). https://doi.org/10.1145/312624.312649
6. Kou, W., Li, F., Baldwin, T.: Automatic labelling of topic models using word vectors and letter trigram vectors. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) AIRS 2015. LNCS, vol. 9460, pp. 253–264. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28940-3_20
7. Kriukova, A., Mitrofanova, O., Sukharev, K., Roschina, N.: Using explicit semantic analysis and Word2Vec in measuring semantic relatedness of Russian paraphrases. In: 2018 Digital Transformations and Modern Society (2018)
8. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th Annual Meeting of the ACL, pp. 1536–1545. ACL, Stroudsburg (2011)
9. Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best topic word selection for topic labelling. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), No. August, pp. 605–613 ACL, Stroudsburg (2010)
10. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD Knowledge Discovery and Data Mining, KDD 2007, p. 490. ACM Press (2007). https://doi.org/10.1145/1281192.1281246
11. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of EMNLP, vol. 85, pp. 404–411 (2004). https://doi.org/10.3115/1219044.1219064
12. Mirzagitova, A., Mitrofanova, O.: Automatic assignment of labels in topic modelling for Russian corpora. In: Botinis, A. (ed.) Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics, pp. 107–110. ISCA, Saint Petersburg (2016). https://www.researchgate.net/publication/320444549
13. Panicheva, P., Mirzagitova, A., Ledovaya, Y.: Semantic feature aggregation for gender identification in Russian Facebook. In: Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) AINL 2017. CCIS, vol. 789, pp. 3–15. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-71746-3_1
14. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
15. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
16. Sorodoc, I., Lau, J.H., Aletras, N., Baldwin, T.: Multimodal topic labelling. In: Proceedings of the 15th Conference of EACL, vol. 2, pp. 701–706 (2017). https://doi.org/10.18653/v1/E17-2111