А.В. Зайда, Я.А. Согуляк
Национальный исследовательский
Томский политехнический университет

Discovery of latent conference topics

This study applies natural language processing methods to infer new information about past conferences. Specifically, normalization algorithms, latent semantic analysis and clustering analysis was used on titles of articles in an attempt to uncover similarities between them. Thus, new sections can be discovered for repeating conferences, or new metrics may be developed.

Keywords: cluster analysis; normalization; dimensionality reduction; language processing; article similarity.

Conferences are an important aspect of scientific progress, as they allow researchers to hold discussions and share results in their fields of study. However, there is always a problem of finding the exact conference, fitting the scientist's work, to attend. Topics, published by conference organizers, may help combat this obstacle, but they are oftentimes rather broad. It is possible, that articles, published in different sections of the same conference, still form a coherent topic, such that organizers could not predict. This article aims to provide a method of identifying latent themes based on past conference materials. To achieve this goal, various natural language processing (NLP) techniques will be used, making use of informational technologies. The NLP pipeline is based on the one described in «A clustering approach for topic filtering within systematic literature reviews» [1]. As for the software realization, python programming language was chosen as well as scikit-learn [2] module, for the ease of development and availability. In the end, uncovered themes will provide more narrow topics, that were actually brought up in discussion.

Materials from the «Linguistical and culturological aspects of modern engineering education» conference of 2021 were chosen for validation and demonstration of the method. In particular, titles of published articles were of interest. Data was acquired through a single query on elibrary.ru, where this information is available publicly. Then article titles were represented in a more structured XML format for ease of computer processing.

First and foremost, some special procedures must be performed on the data to greatly enhance the quality of further analysis. But for any computer algorithms to work, data must be presented in a single language first. As such all the text had to be translated into English. Natural languages are characterized by high degree of polymorphism, polysemy and redundancy, making any kind of computer analysis harder. These aspects are partially compensated by normalization of text. Stop-word filtering is used to remove any insignificant

terms, prepositions for example, and focus on truly relevant parts of sentences. The technique is easy to implement and quite handy to use, so it was added in normalization sequence. More complicated algorithms for combating polysemy include stemming and lemmatization. Stemming tries to return the word to its base form, allowing the computer to make matches between verbs in different tenses, for example. Lemmatization is more powerful, as it regresses the word to the basic idea behind the term. However, lemmatization requires additional information on the part of speech, which cannot be accurately inferred by machine. As such, stemming was chosen next as easier and more reliable option.

Once dataset is normalized, actual work can begin. At this point, information must be transformed into machine-friendly form. Text is great for human understanding but is rather expensive in memory usage and processing difficulty for computers. Vectorization is a process of representing textual information with some fixed number of values. This set of numbers is called a vector, and all the titles in the dataset must be transformed in this way. All of the vectors together form another mathematical structure – a matrix. This matrix is akin to a table and is often called document-term matrix. Every row represents a single document, which is an article title in this case, and every column describes a single term, which appears in at least one title.

There are multiple ways of building vectors for sentences. The easiest one utilizes simple count of how many times particular word was encountered. This approach has significant downside – frequent and uninformative words, such as auxiliary verbs, will have high scores, while relevant terms will be lagging behind. For this reason, a more powerful technique was devised, called TF-IDF vectorization. TF-IDF stands for term frequency and inverse document frequency. Therefore, a high score of a term in a single document is achieved when there are many occurrences of the term in particular document and few in all the rest documents. All in all, TF-IDF vectorization is a better algorithm and increased performance justifies the computational difficulty, so it was chosen over the others.

Moving forward, latent semantic analysis (LSA) will be applied. LSA is technique in natural language processing, which reduces dimensionality of dataset and discovers related terms. The principle behind LSA is special matrix operation on document term matrix that decreases the number of columns but does not significantly change similarity of rows. Similarity of rows can be defined in several ways through vector algebra using vector values. After decomposition, columns no longer represent single words, but rather small groups of terms, that are frequently used together. As a result, LSA makes computation easier and helps discover latent similarities. The smaller the number of columns is, the less is computational difficulty, but the more information is lost,

therefore, it is necessary to pick the lowest possible number that still explains most data variance. The degree of reduction is determined on a case-by-case basis through experimentation, although some sources suggest that reduced matrix should account for at least thirty percent of original information. However, through trial and error, it was found that reduction to three columns works best for this dataset. This has added benefit of helping interpretation, because it is possible to plot all the vectors in 3D space.

Next thing to be done in order to achieve the goal, is to group articles, using the vectors that were obtained earlier, into clusters, in which all objects are somehow similar to each other. All clustering algorithms require some parameters to be defined, almost exclusively these parameters are found through experiments. Two different functions have been tested: affinity propagation and KMeans. The former was chosen due to its feature to determine the optimal number of clusters that we need. Then, the latter was used to get the final results of clustering and was chosen for its ease of interpretation. Affinity propagation looks for exemplars of clusters amongst existing dataset and so, this algorithm cannot be interpreted good enough. KMeans defines groups of vectors through mean values, allowing to discover which components contribute to association with this cluster. Moreover, KMeans has internal evaluation of quality of clustering, which helps with search for best parameters. This metric is called inertia and should come closer to zero for a good cluster partition.

Finally, comes the validation of the method. A program was written in python programming language, utilizing the algorithms described. For the dataset, consisting of articles from «Linguistical and culturological aspects of modern engineering education» conference of 2021, 6 clusters were identified through affinity propagation and minimalization of inertia (Fig).

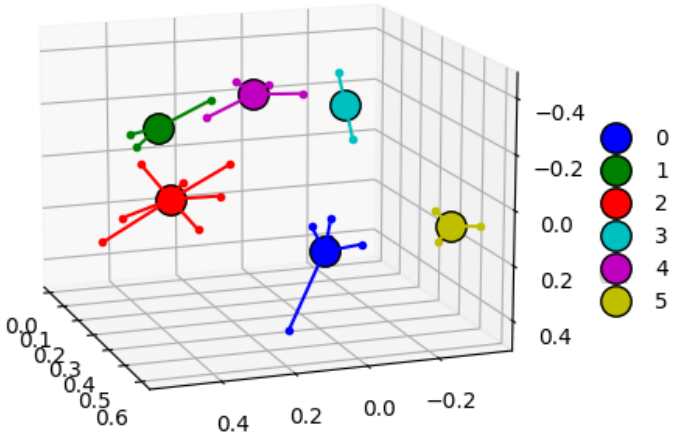

Fig. 3D plot of clustering results

The most interesting cluster based on its contents is the red one. It includes following articles: «Method of project education», «On compatibility

of English borrowings in Norwegian internet-texts about anti-crisis management», «Company motto in international business: translation or transcreation?», «German borrowings, which appeared in Russia after perestroika (in economics and politics)», «Startup projects: large public technical university experience», «Strategy and tactics of self-presentation based on political speeches of Donald Trump», «Evaluation of quality of artistic translation of poetic text from English into Russian», «Rhetorical portrait of a student from faculty of international business communication». Several prevalent themes are immediately noticeable in this cluster, such as «loanwords», «politics» and «business», which could have been viable topics for discussion. Moving on, blue cluster provides less useful information. It includes: «Linguistic features of electronic discourse», «Comparison of language learning applications», «The problems of localization in videogames based on «Uncharted» game series», «Training of translators based on Lederer's theory of meaning», «Special aspects of translating military vocabulary in Warhammer 40,000 – related literature». Characteristic terms for this group appear to be «linguistics» and «translation», which is not entirely unexpected, given the nature of conference.

On the other side of the spectrum, there are clusters, that are barely of interest, because their themes are already covered by conference sections. Yellow cluster builds around the word «engineer», consisting of: «Comparison of exams in terms of benefits for the future engineer», «Monge Gaspard. Founder of the language of engineers», «The importance of English in the IT engineering profession». Next, purple cluster seemingly lacks common terms: «Intercultural communication as interpersonal interaction: history of emergence», «Concept of product through prism of two social and language groups», «Modern techniques for memorizing foreign words», «Ways to develop specialists in engineering activities: professional development and retraining». Then, cyan cluster grabs at «foreign languages» as terms: «Communication skills in foreign languages in engineering», «Traditions and innovations in the methodology of teaching foreign languages». And finally, green cluster focuses on plain «language»: «Computer games as a way to learn foreign languages», «National and cultural specifics of zoomorphism in Russian and Chinese proverbs», «Effectiveness of using authentic sources for foreign language studying (English language, master's degree)».

In conclusion, the method described does uncover some latent topics of discussion from the past conferences, but inconsequential results are also possible. It is likely that algorithm performance may improve with more data, accumulated through years. Even though results at the moment do not allow uncovering new big conference sections, some interesting topics of discussion still were discovered.

Литература

1. A clustering approach for topic filtering within systematic literature reviews / T. Weißer, T. Saßmannshausen, D. Ohrndorf, et al. // MethodsX. – 2020. – Volume 7. – URL: https://www.sciencedirect.com/science/article/pii/S2215016120300510?via%3Dihub (date of access 27.09.22). – Text: electronic.

2. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort, et al. // Journal of Machine Learning Research. – 2011. – Volume 12. – P. 2825–2830. – URL: https://scikit-learn.org/stable/about.html#citing-scikit-learn (date of access 27.09.22). – Text: electronic.

Науч. рук.: Аксёнова Н.В., к-т филол. н., доц.

Н.Н. Зяблова [1], В.А Зяблов[2]

[1]Национальный исследовательский
Томский политехнический университет
[2] Томский государственный университет
систем управления и радиоэлектроники

Terminology in IT sphere in modern English: structural aspect

The present article presents the results of the study of the terminological system of the sphere of information technology in modern English. The analysis of the grammatical structure of terms and terminological units in the abovementioned sphere has been carried out. Most common, least common and not common ways of formation of terminological units of the specified subject area have been identified.

Key words: terms; terminological units; English; grammatical structure; information technology.

In the current article the results of the linguistic analysis of the grammatical structure (parts of speech) of terms and terminological units in the topical and developing scientific sphere of information technologies in modern English are presented. Most common, least common and not common ways of formation of the specified terminology have been identified which will enable us to predict the most and least popular ways of nominating special notions and objects of the abovementioned scientific sphere.

A large number of publications are written annually in scientific and technical journals in English by scientists from different countries. English is used for international communication to exchange information coded in terms and terminological units throughout the world.