Representation of Wikipedia articles for NLP tasks

Julian Szymański

julian.szymanski@eti.pg.edu.pl

Gdańsk University of Technology, Poland

Proper representation of the text is crucial for achieving good results of natural language processing. Because understanding of the language is still far from the machine capabilities, to process the text it is usually represented as a set of features. The features approximate the meaning that allows us to compare the sequences of the characters in the way that captures elementary semantics. The key thing is where these features come from. In general, we can divide representation methods into two main groups: based on content and based on references (or context).

The first one use statistics based on token co-occurrences analysis. In this group of representation methods, the most intuitive approach is the Bag of Words, where tokens (in the processing tasks called features) are words with weights. There are many variants for building the weights, which allows us to express different aspects of similarity. Other approaches compute the feature weights from the token contexts eg.: Hyperspace Analogue to Language. There are many variants for selecting the tokens (eg.: n-grams, keywords) as well as different methods for representation space transformation (eg.: Latent Semantic Analysis, neural embeddings). In addition, content-based methods enhanced can be by using external sources of information (eg. WordNet, Explicit Semantic Analysis [1]).

The second group of representations is based on references. In this approach, an elementary chunk of the meaning (concept), is represented using relations to the others. Here, a set of the features that characterize the concept is constructed with its neighbors. Depending on the source of the information, the references can be acquired from hyperlinks (as in the case of web pages, or bibliographic notes as in the case of scientific articles).

In our research, we focus on representation of Wikipedia articles. Each of the articles can be considered as a unique entity that represents a concept. In addition, Wikipedia provides a hierarchical structure of the categories that can be used as description of the concepts on the abstract, more general level. To process efficiently the Wikipedia data we construct an application that enables us to construct its machine-processable form, using a selected representation method. The application is available on the website of Computational Wikipedia Project.[1] We test different methods of representation and evaluate them in classification tasks aiming at reconstruction of Wikipedia categories structure [3]. In the project, we also construct mappings between Wikipedia and Wordnet.

We are working on methods for improvement of information retrieval. In our research, we test the algorithms for clustering Wikipedia search results [4]. The results using interaction with the user based on selection, a so-called conceptual direction can be refined and thus irrelevant information is filtered [2] [5]. We are developing the methods of mining relations between Wikipedia [6] categories that aiming at automatically identifying significant relations. This method can be used for creating the conceptual directions - sequences of related Wikipedia categories that allows one to differentiate the sets of similar articles.

[1] https://kask.eti.pg.gda.pl/CompWiki/

[2] https://kask.eti.pg.gda.pl/BetterSearch/en/

Human understanding of the concepts is strongly related to the already possessed knowledge. One of the aspects of the understanding is integration of the perception of objects with this what we already know. Thus, Wikipedia can be used as the representation space where the concepts identified in the text are projected. This approach has been used by Wikification [2] methods, but they were tested on a very limited set of the articles. We are working now on the methods that allow efficient processing and identifying of the Wikipedia articles in the text. Using them, the text can be represented in the space created by Wikipedia categories that provides abstract insight into relations between concepts. We construct the platform for building and evaluating Wikification algorithms[3]. Wikification combined with conceptual directions allows the capturing of different aspects of the similarities computed in the selected subspace. This expreses semantic similarities selected by the user interests. This should provide better results than computing proximity on the whole set of the features that usually leads to effect cruse of the dimensionality.

References:

[1] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

[2] D. Milne and I.H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[3] Julian Szymanski. Comparative analysis of text representation methods using classification. *Cybernetics and Systems*, 45(2):180–199, 2014.

[4] Julian Szymanski and Tomasz Dziubich. Spectral clustering wikipedia keyword-based search results. *Front. Robotics and AI*, 2017, 2017.

[5] Julian Szymanski, Henryk Krawczyk, and Marcin Deptula. Retrieval with semantic sieve. In *Intelligent Information and Database Systems - 5th Asian Conference, ACIIDS 2013, Kuala Lumpur, Malaysia, March 18-20, 2013, Proceedings, Part I*, pages 236–245, 2013.

[6] Julian Szymanski and Jacek Rzeniewicz. Identification of category associations using a multilabel classifier. *Expert Syst. Appl.*, 61:327–342, 2016.

[3] https://kask.eti.pg.gda.pl/eyt/