Exploring Multidimensional Continuous Feature Space to Extract Relevant Words

Márius ŠAJGALÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
sajgalik@fiit.stuba.sk

With growing amounts of text data the descriptive metadata become more crucial in efficient processing of it. One kind of such metadata are keywords, which we can encounter e.g. in everyday browsing of webpages. Such metadata can have various purposes, like usage in web search or content-based recommendation.

In our work we focus on vector representation of words to simulate the understanding of word semantics. Each word is thus represented as a vector in N-dimensional space, which includes the advantages of using various vector operations like easy similarity measuring between pairs of words, or vector addition and subtraction to compose meaning of longer phrases. We can also calculate what words are the most similar by finding the closest vectors, or vector that encodes a relationship between pair of words, e.g. vector transforming singular into plural, etc. With word vectors, we can encode many semantic and also syntactic relations [3].

Such representation has a big potential in NLP and we can only anticipate that in a few years it will completely supersede all those manually crafted taxonomies, ontologies, thesauri and various dictionaries, which are often rather imprecise and erroneous. Moreover, there is no means of measuring similarity directly between pairs of words in this hand-crafted data. Most relations are just qualitative and described by their type (e.g. approach in [1] reveals only a relation type, but it cannot determine the relation quantitatively) and thus, all existing methods for measuring (semantic) similarity are limited to achieving only rather imprecise results.

We research the computation of keywords in vector space. This perspective on the keyword extraction problem also brings another new interesting challenges and there are lots of unsolved open problems. So far, we developed a method of extracting relevant words in clusters of words with similar meaning (see Fig. 1).

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 37–38.

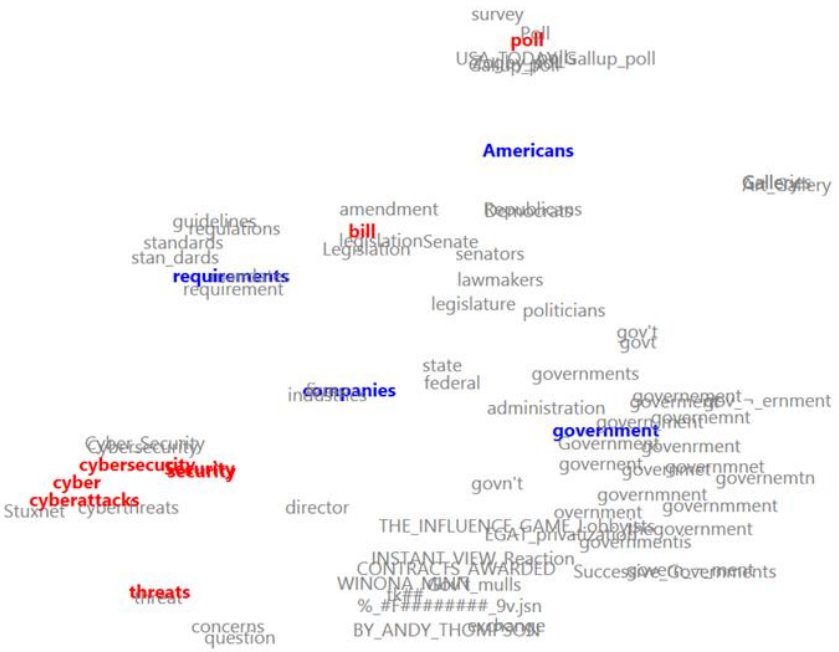

Figure 1: Top 100 most relevant words extracted from website titled "Cybersecurity poll: Americans divided over government requirements on companies" and visualised by t-SNE [2]. Red words are keywords selected by website editors and blue words are keywords that we would also manually and subjectively choose as the most relevant for this article.

We can see that each cluster contains a keyword, but is cluttered with other similar words that often correspond to less common synonyms or their misspelled alternatives, so that computed data is still noisy and needs to be cleaned. This could be achieved by using frequency statistics, which is our next task to complete.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 93-100.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Barla, M., Bieliková, M.: On Deriving Tagsonomies: Keyword Relations Coming from Crowd. In: Proc. of the 1st Int. Conf. on Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. pp. 309–320 Springer-Verlag (2009).

[2] Van der Maaten, L.J.P.: Barnes-Hut-SNE. In: Proceedings of the International Conference on Learning Representations, 2013.

[3] Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of NAACL HLT, 2013.