Relationship Extraction using Word Embeddings

Matúš PIKULIAK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matus.pikuliak@gmail.com

The Web currently contains the largest collection of documents the humanity has ever seen. Dealing with the documents written in natural languages is subject of intensive research. The quantity and nature of this unstructured data make it difficult to process for machines. There are numerous tasks concerned with extracting knowledge from such data. In this paper we are dealing with extraction of relationships, particularly with discovering semantic relations between lexical units such as words or phrases. Example of such relation can be a relation between countries and their capital cities.

Extracting such information manually is very arduous and error-prone task. Our aim is to partially automatize this task and help engineers build knowledge bases in various expert systems. Existing methods for such extraction were usually tailored only for extracting one concrete type of relation, e.g. taxonomic one [2]. In addition, they usually involve the necessity of manually creating exact rules used to find new relations. These rules are being specifically tailored to one specific type of relation and cannot be easily modified to fit another.

We propose new method with an aim to considerably decrease time and cognitive difficulty of relationship extraction for generic semantic relations. Our method requires only a handful of examples to define a relation. I.e. we expect our method to expand given set of pairs. If we supply our system with pairs like *France-Paris, Italy-Rome, Russia-Moscow*; we expect it to return new pairs with the same semantic relation. We think this kind of interaction with system is very simple and straight-forward. The set can be assembled manually by human users or automatically by machines extracting existing knowledge from structures such as ontologies.

We are using deep learning algorithm to create vector space for lexical units [1]. This algorithm project units into high-dimensional space based on the contexts the words are being used in text corpus. Words that are being used in similar contexts have similar vectors, thus making this algorithm an application of Harris' hypothesis stating that

* Supervisor: Marián Šimko, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 59–60.

words used in similar sentences are semantically similar. This method of representation of units became popular recently and was already successfully used in variety of NLP tasks [3].

We project the pairs examples given as input into this space and we utilize machine learning algorithms to learn about the patterns they leave in this space. Existence of these patterns were already proved empirically [4]. We then apply this knowledge to rate set of selected candidates. The higher the rating these candidates have the higher the chance they should have desired relation. Technique called PU Learning looks very promising solving this task. It is essentially variant of binary classification that can work only with positive and unlabelled data. Other techniques based on similarity calculations were also evaluated.

Result of our method is a list of pairs sorted by their chance of having the same relationship as the pairs in seed set. We have evaluated results for several semantic relations with seed sets consisting of 25 examples pre relation. We judged the relevancy to seed set for first 100 results from each list. In average approximately 20% of these results are correct while our best result is 52%. We consider these results very promising as it shows that our method was able to learn about semantic relations based only on handful of examples. We believe this approach can be further researched and ultimately used in real life knowledge engineering applications.

We were surprised that different measures for scoring pairs did not differ only quantitatively but also qualitatively. It seems that each measure has its own strategy for scoring. We think that we could use this phenomenon and further improve the efficiency of our method by combining several unique scoring methods. We are also looking for ways to enrich existing space with more and better defined concepts other than words. This could simplify the knowledge extraction layer of our task.

Extended version was published in Proc. of the 12th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2016), STU Bratislava.

References

[1] Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. *In Innovations in Machine Learning*, pages 137-186. Springer, 2006.

[2] R. Fu, J. Guo, B. Qin, W. Che, H.Wang, and T. Liu. Learning semantic hierarchies via word embeddings. *In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, 2014.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746-751, 2013.