Utilizing Vector Models for Processing Text on the Web

Ladislav GALLAY *

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
ladislav.gallay@lentil.sk

Unifying various word forms is the first step in understanding the natural language. Lemmatization is the process of transforming the word into its root form – lemma. Understanding of each individual word is followed by understanding the whole sentence and the context of given document. Current approaches in this area are limited to knowing either full grammar rules or building the translation matrix from the word to its basic form [2].

While the former approach is impossible to fully support the fusional languages such as Slavic ones, the latter is hard to maintain and needs to be manually updated every time new word is introduced into the language. It requires a lot of human input and is error prone.

We have realized that these linguistic rules must be are already captured in the natural text. Vector models are capable of extracting these linguistic regularities from the text into mathematical vectors. The operation such as vector('king') - vector('man') + vector('woman') results in vector that is close to vector('queen'). Recently the word2vec has been introduced [1]. It is capable of training the model using continuous bag-of-words and skip-gram architectures even more effective. The input is some meaningful text in natural language and the output is list of words and their latent vectors.

Based on the results of word2vec tool we have discovered that not only semantic but also morphologic relations are stored in vectors. This allows us to make a query ‘dogs’ to ‘dog’ is the same as ‘cats’ to which results in the word cat. These regularities are shown in Figure 1.

We propose new algorithm that utilizes vector model of words. The above example shows how our algorithm works. We take two reference words, for example ‘vodníkom’ and ‘vodník’ and try to find similar shift from the input word ‘rybníkom’. As the result we are given several words around the word ‘rybník’ which is the root form.

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

Spring 2015 PeWe Workshop, April 11, 2015, pp. 91–92.

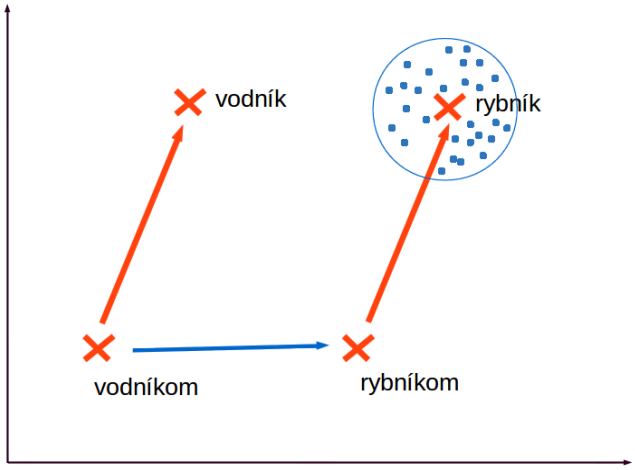

Figure 1. Relationship between vodníkom and vodník is the same as between rybníkom and rybník.

Our algorithm expects to have small set of those reference pairs to choose from. Then we propose several methods of choosing the correct pair. The longest suffix seems to be the most reasonable. However we will test also random choosing or choosing the semantic related words first. Doing several iterations we end up with many words probably around the root form.

Then we need to filter the correct word. The words usually differ only in suffix and the prefix should remain the same. Although this is not necessary true for all languages, it can be applied in many cases for Slovak language. We calculate the common prefix divided by average length of the input and each output word. The result is multiplied by the distance. This number should represent output accuracy.

We have tested the algorithm on corpus from SME.sk and Wikipedia. The results were not impressive and are heavily dependent on how well the corpus is trained. Recently we were given the access to much larger Slovak national corpus which outputs much better results. Our next work will include final testing and comparison of different methods of evaluation that we propose.

Amended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 89-90.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0752/14.

References

[1] Mikolov T. et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, (2013).

[2] Garabík, R., Giantisová, L., Horák, A., Šimková, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. (2004).