Keyword Extraction in Slovak

Adam RAFAJDUS*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
adam.rafajdus@gmail.sk

Keywords often represent important role in text documents, not only used for categorization, but also as a placeholder, since human can imagine the theme, the category and the content of the document based on the well-chosen keywords.

The process of keyword extraction used to be primarily manually evaluated, but as it is not very efficient method, more capable methods like TF-IDF with relative positions or distance-based metrics were successfully applied to extract keywords [2]. The main disadvantage of these new methods was that they were heavily dependent on special pre-processing of the text dataset, such as Part-of-Speech tags and/or using lemmas, which require additional time to process, since Slovak language is morphologically rich. This problem can be solved using the new methods in NLP, which are changing the way of processing words by changing them into word vectors, which are much better thanks to their vector properties. In vector space, we can perform various vector operations while keeping semantically close words together in the vector space.

Our approach is to implement a recurrent neural network based on LSTM – Long Short-Term Memory module, which thanks to its structure is working similarly to a person reading text and learning text's keywords and category by understanding similarities of texts. The proposed model is not only taking advantage of coherency of the text by processing words one after another, but also of word vectors that we use, which offer us another layer of information about text since it captures many semantic and syntactic regularities [3].

Our proposed architecture is created based on categorization architecture of LSTM and has 4 layers:

1. Latent feature vectors of input words
2. Main LSTM module
3. Keywords vectors processing
4. Softmax classifier

* Supervisor: Márius Šajgalík, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 69–70.

The first layer is using well pre-trained word vectors to initialize the lookup table, which is fed into the LSTM module. We use Word2vec tool to create our latent feature vectors. For initializing the lookup table we have corpus of pre-trained word embeddings with 3 types of word vectors - 80, 200 and 300 dimensional. Each offers different level of complexity and efficiency.

The LSTM module is currently very popular architecture of recurrent neural network in literature, referred to as vanilla LSTM, which have emerged as an efficient and scalable model for several problems related to sequential data [1]. The main idea behind using the LSTM architecture is a memory cell which can maintain its state over time, and non-linear gating units, which can regulate the information flow into and out of the cell. Although the initial version of LSTM block was different and struggled with the same problems as RNN, this version already includes many changes, mainly based on forget gate and output activation function, which are the critical components of LSTM block. It is performing reasonably well on various datasets [1].

Keyword processing is performed by transforming output of the LSTM module into form of n-keyword feature vectors and comparing them to n-word feature vectors. This should end up by transforming output into real word vectors with characteristic of processed text – keywords.

At the end, the softmax function is applied as a classifier to predict the probability distribution over the category set. Comparing prediction results with real categories of text, we can use regular backpropagation to apply supervised learning on network to train categorization over the training set, while extracting keywords from third layer.

As our next step, we plan to apply this method on Slovak Wikipedia - our corpus, thanks to its large database of articles with assigned categories, first testing efficiency of categorization and eventually of keyword extraction.

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

Acknowledgement: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Greff, Klaus, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink a Jürgen Schmidhuber, 2015. LSTM: A Search Space Odyssey. arXiv [online]. 2015, p. 10.

[2] Hulth, Anette, 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. EMNLP'03: Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing. 2003, p. 216-223.

[3] Mikolov, Tomas, Greg Corrado, Kai Chen, Jeffrey Dean, 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013). 2013, p. 1-12.