The evaluation of semantic modelling techniques for the topic detection task on social media content

Barnabás Nagy, László Grad-Gyenge

Eötvös Loránd University, Faculty of Informatics

barna@inf.elte.hu, laszlograd@inf.elte.hu

In this paper, we present the evaluation of various semantic representation techniques for the topic detection task on social media content for Hungarian language. Topic detection is one of the core problems of machine learning based text analytics. In a nutshell, the technical problem can be described as applying distance metrics to the semantic representation of various instances of text of natural language. The processing pipeline typically consists of specific stages as text extraction, natural language processing and semantic modelling. At first, text extraction is more of a technical task. At second, the natural language processing toolchain is well established for the Hungarian language. This is the reason why we focus on the semantic representation problem on the mentioned domain.

Regarding the textual content, the content on social media has specific attributes as short messages, specific vocabulary and the less strict use of the grammar rules. This is the reason why semantic modelling techniques showing reasonable performance on traditional content can show a different spectrum of evaluation on this domain.

Our goal is to identify the most adequate semantic representation method for topic detection for social media messages. To describe our corpus, the work has been conducted on a previously agreed list of Facebook pages in a specific time frame. The social media messages are extracted utilizing the Facebook API.

To conduct topic detection, a list of topics is defined by a human expert. Each topic is represented by a manually assembled list of messages referred as prototypes. This technique basically identifies each topic as a set of message. The automatic topic assignment for a specific message is then defined as the topic of the closest prototype in the semantic space.

The performance of the methods are evaluated by human experts. To measure the performance of the methods, a particular user interface has been developed. During the evaluation, each topic assignment is presented to one expert. The expert can reinforce the topic assignment or suggest an alternate topic. The judgement of the experts is then summarized and presented. The primary contribution of our work is the comparison of the performance of involved methods as LDA [1], LSI [2], W2V and WW2V [3].

References

[1] Blei D. M, Ng A. Y, Jordan M. I. *Latent Dirichlet Allocation.* Journal of Machine Learning Research 3, pp 993-1022. 2003.

[2] Morie T, Uchimura K, Amemiya, Y. *Analog LSI implementation of self-learning neural networks.* Computers & Electrical Engineering, Volume 25, Issue 5, September pp 339-355. 1999.

[3] Sanjeev Arora, Yingyu Liang, Tengyu Ma *A Simple but Tough-to-Beat Baseline for Sentence Embeddings.* ICLR 2017 conference, Princeton University, 2017.