Measuring semantic similarity of words

Eszter Iklódi
Dept of Automation and Applied
Informatics Budapest U of Technology
and Economics H-1117 Budapest,
Magyar tudosok krt. 2
eszter.iklodi@gmail.com

ABSTRACT

We propose a procedure for creating universal (language-independent) word embeddings using pretrained word vectors and a lexical database, for use in natural language processing (NLP) applications. In our work we used the fastText [1] word vectors as pretrained embeedings and the Swadesh list [2] of the PanLex lexical database [3].

KEYWORDS

computational linguistics, natural language processing, semantic representations, word embedding, universal semantic representation

1 CONTENT

Computational linguistics (a.k.a. natural language processing, NLP) is a dynamically evolving research field within computer science. A more specific area concentrates on creating semantic representations of language data, which can be used not only in solving tasks in the field of computational semantics (e.g. question answering, chatbots), but also other standard NLP tasks such as machine translation or syntactic parsing.

One way to build a semantic representation is to use a distributional model, commonly referred to as *word embedding*. Embeddings map each word of a language to a real-valued vector of some fixed dimension. They are trained on large corpora (collections of texts) with the objective that the more similar two words are the bigger their cosine similarity shall be. . This approach, which has proved highly practical in nearly all areas of NLP, is based on the *distributional hypothesis*, i.e. that words occurring in similar contexts have similar meanings

Given the need for robust representations for many languages, the question of whether human conceptual structure is universal has recently gained interest among computational linguists. Youn et al. [4] has shown that human conceptual structure is independent of certain non-linguistic factors such as geography, climate, topology or literary traditions. Based on such findings we propose a procedure to construct a universal semantic representation in form of a universal embedding along with translation matrices that serve to map each language to the universal space. We use the pretrained fastText word embeddings [1], which are available for 294 languages, and the Swadesh list of the Panlex parallel lexical database [3]. We construct the universal word embedding by minimizing the following loss function:

$$\sum_{i \in L} \|W_i \cdot T_i - A\|_2^2$$

where $W_i$denotes the embedding and $T_i$the translation matrix of the ith language, while A denotes the common universal embedding.

Training was performed on 73 languages using the 110-entry-long Swadesh list [2] which is a parallel dictionary of basic concepts (e.g. sun, water, earth, stone etc.) that exists for thousands of languages. We evaluated our results on words of the 207-entry-long Swadesh list that were not used during the training, i.e. are not part of the 110-entry-long Swadesh list. Early results show a correlation of 0.6-0.9 measured between the cosine similarities of the original embeddings on the one hand and the cosine similarities of the embeddings obtained through mapping the original ones to the universal space on the other hand.

REFERENCES

[1] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

[2] Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. International journal of American linguistics, 21(2), pp.121-137.

[3] Kamholz, D., Pool, J. and Colowick, S.M., 2014. PanLex: Building a Resource for Panlingual Lexical Translation. In LREC (pp. 3145-3150).

[4] Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J.F., Maddieson, I., Croft, W. and Bhattacharya, T., 2016. On the universal structure of human lexical semantics. Proceedings of the National Academy of Sciences, 113(7), pp.1766-1771.