Exploring sentence embeddings: a suitable method for a lexicography-oriented analysis of argument structures?

Fritz Kliche (Hildesheim University) &
Laura Giacomini (Hildesheim University/Heidelberg University)

kliche@uni-hildesheim.de
laura.giacomini@uni-hildesheim.de/laura.giacomini@iued.uni-heidelberg.de

According to the approach at the core of the PhraseNet project in learner's lexicography (DiMuccio-Failla & Giacomini, 2017), each sense of a verb is identified by a specific argument structure filled with specific semantic types and roles. Experiments in the analysis of argument structures of verbs through sentence embeddings are presently being carried out as an alternative or complementary solution to the current collocation-based method for finding word senses (Giacomini & DiMuccio-Failla, 2019). Our presentation discusses the application of this method to the lexicographic process by illustrating the example of the English verb *follow*.

We use SBERT (Reimers & Gurevych, 2019), which was developed for measuring similarities of sentences, for testing different approaches that might provide information on semantic types and roles for a given argument structure. On the one hand, for instance, we cluster instances of syntactic patterns of *follow*, while, on the other hand, we sort occurrences of *follow* by their similarities to instances which are representative for a given argument structure. All input data are retrieved from the British National Corpus by means of the Sketch Engine (Kilgarriff et al., 2014).

A number of crucial issues will be addressed by the presentation, such as the choice of the initial input for the sentence transformer, the analysis of implicit semantic arguments, the reliability of the method in terms of lexicographic data quality as well as the amount of required manual work in comparison to the current method. The potential role of word embeddings (cf. Domínguez, 2021, Sørensen & Nimb, 2018) and sentence embeddings in lexicography will be also discussed.

References

DiMuccio-Failla, P. & Giacomini, L. (2017). In M. Mitkov (ed.) Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017, LNAI 10596. Springer, 290-305.

Domínguez Vázquez, M. J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über word embeddings bis hin zum automatischen Wörterbuch. *Lexikos*, *31*(1).

Giacomini, L. & DiMuccio-Failla, P. (2019). Investigating Semi-Automatic Procedures in Pattern-Based Lexicography. In: Proceedings of the eLex 2019 conference. Electronic lexicography in the 21st century. Sintra, Portugal.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1, 7-36.

Reimers, N, & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of EMNLP 2019.

Sørensen, N. H., & Nimb, S. (2018). Word2dict – lemma selection and dictionary editing assisted by word embeddings. In: Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts, 819-827.