The LSCD Benchmark - A testbed for diachronic word meaning tasks

Dominik Schlechtweg
Universität Stuttgart

Lexical Semantic Change Detection (LSCD) is a field of NLP that studies methods automating the analysis of changes in word meanings over time. In recent years, this field has seen much development in terms of models, datasets and tasks (Schlechtweg et al., 2020). This has made it hard to keep a good overview of the field. Additionally, with the multitude of possible options for preprocessing, data cleaning, dataset versions, model parameter choice or tuning, clustering algorithms, and change measures a shared testbed with common evaluation setup is needed in order to precisely reproduce experimental results. Hence, we present a benchmark repository implementing evaluation procedures for models on most available LSCD datasets. We hope that the resulting benchmark by standardizing the evaluation of LSCD models and providing models with near-SOTA performance can serve as a starting point for researchers to develop and improve models. The benchmark allows for a wide application and testing of models by focusing on multilingual models and their evaluation on several languages.

Models solving the LSCD task often employ sub-models solving other related lexical semantic tasks like Word Sense Induction (WSI, Navigli, 2009) or Word-in-Context (WiC, Pilehvar & Camacho- Collados, 2020). Performance on these tasks can be evaluated separately contributing to optimization of individual model components and to facilitation of error analysis. However, existing data sets for the latter two tasks are usually synchronic, which makes it hard to compare different sub-models and select optimal ones for the LSCD task that requires good performance on diachronic data. Hence, we exploit existing, richly annotated LSCD datasets as evaluation data for WSI and WiC in a diachronic setting. Using the same data sets for evaluation of WSI, WiC and LSCD has the additional advantage that performance on the meta task LSCD can be directly related to performance on the subtasks WSI and WiC, as it can be assumed that performance on the subtasks directly determines performance on the meta task. We aim to stimulate transfer between the fields of WSI, WiC and LSCD by providing a repository allowing for evaluation on all these tasks with shared model components.

References

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. Proceedings of the 14th International Workshop on Semantic Evaluation.

Roberto Navigli. 2009. Word sense disambiguation: a survey. ACM Computing Surveys.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.