A Semi-Supervised Interactive Algorithm for Word Sense Clustering

Dominik Schlechtweg[1] , Sabine Schulte im Walde[1]

[1] *University of Stuttgart*

dominik.schlechtweg@ims.uni-stuttgart.de, schulte@ims.uni-stuttgart.de

Traditionally, word sense annotation in lexical semantics relied on a fixed inventory of senses assigning a single best sense for a given use (Navigli, 2009). Nowadays, studies often take a graded view on word meaning where a use may be assigned to multiple senses on a graded scale (Erk, McCarthy, & Gaylord, 2013) or use pairs may be annotated for their semantic proximity and then clustered (McCarthy, Apidianaki, & Erk, 2016). While the latter approach avoids the definition of a word sense inventory, and thus by itself gives no information on the quality of a sense cluster (what sense a cluster represents), the approach allows to measure important lexical properties such as polysemy or vagueness.

We propose an online interface with an underlying algorithm requiring only two manual inputs: (i) a sample of uses for a target word for each of the different corpora that should be compared, and (ii) consecutive judgments of use pairs from these samples. The algorithm will pass through several steps presenting use pairs to annotators and using their judgments to infer word sense clusters in an efficient way. It may be applied to create data sets for different fields such as lexical semantic change detection (Schlechtweg, Hätty, del Tredici, & Schulte im Walde, 2019), term extraction (Hätty et al., 2019), graded word similarity inference or lexicography.

Use pair annotation is attractive, because it requires no manual preparation except for the sampling of uses from a corpus. We extend this approach to annotate use pairs sampled from different corpora allowing us to measure differences in a word's corpus-specific sense distributions. The resulting approach is largely automatized, efficient, language-independent and yields high inter-annotator agreement (Erk et al., 2013).

References: Hätty, A., Schlechtweg, D., & Schulte im Walde, S. (2019): SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*. Minneapolis, MN, USA, pp. 1–8. • Erk, K., McCarthy, D. and Gaylord, N. (2013): Measuring word meaning in context. *Computational Linguistics* 39(3), 511–554. • McCarthy, D., Apidianaki, M., & Erk, K. (2016): Word sense clustering and clusterability. *Computational Linguistics*, 42(2), 245-275. • Navigli, R. (2009): Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2), 1–69. • Schlechtweg, D., Schulte im Walde, S. and Eckmann, S. (2018): Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* New Orleans, Louisiana, pp. 169–174. • Schlechtweg, D., Hätty, A., del Tredici, M., & Schulte im Walde, S. (2019): A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting Association for Computational Linguistics* (Volume 1: Long papers). Florence, Italy, pp. 732–746.