Word Sense Disambiguation

P. S. Langeslag

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA
SEIT 1737

The “Most Frequent Sense” Baseline

If no other information is available, assign the most frequent sense.

The “One Sense per Discourse” Heuristic

Ambiguous words recurring in the same context will tend to share a single sense.

The “One Nearest Neighbour” Algorithm

Use contextual word embeddings, select the nearest matching embedding.

- Requires a semantically tagged corpus (e.g. SemCore) as well as contextual embeddings (e.g. BERT).
- Words not in the tagged corpus can be handled by averaging the embeddings of its synset in WordNet.

Context (“Features”)

- Part-of-speech data on tokens in a window
- Collocation information (*n*-grams)
- Syntactic relations
- Weighted averages of context embeddings.

Context POS alone awarded rather a modest place in Jurafsky and Martin, surely because the fine sense distinctions in current WSD work don’t correlate with differences in context POS.

Lesk Algorithm

Use overlap between context and dictionary definitions.

Bibliography

Jurafsky, Dan, and James H. Martin. *Speech and Language Processing*. 3rd ed. draft., 2021. http://web.stanford.edu/~jurafsky/slp3/.

Wunderlich, Martin, Alexander Fraser, and P. S. Langeslag. "'God Wat Þæt Ic Eom God': An Exploratory Investigation into Word Sense Disambiguation in Old English." In *Proceedings of GSCL 2015: International Conference of the German Society for Computational Linguistics and Language Technology*, 39–48. Munich: Gesellschaft für Sprachtechnologie und Computerlinguistik, 2015. https://konvens.org/proceedings/2015/.