Build your own Semantle-clone: Introduction to Distributional Semantics and Word Embeddings for Linguists*

Luuk Suurmeijer

Institute for Language and Information, Heinrich Heine University Düsseldorf

An innocent letter-guessing-game called Wordle went viral in the beginning of 2022. One particular clone of this game, called Semantle[1], is particularly interesting from a linguistic perspective. The user is tasked with guessing a secret word based on its semantic similarity to other words. Semantle is premised on The Distributional Hypothesis (Firth, 1957), which states that the meaning of a lexical item can be approximated by knowing the linguistic contexts in which it is used. Distributional Semantics concerns itself with building accurate distributional lexical representations from language corpora. A very intuitive and successful way of representing words distributionally is using vector spaces (Clark, 2015; Turney & Pantel, 2010). Today, due to their attractive mathematical properties, vector-based representations of lexical items (word embeddings) are an indispensable tool for virtually all NLP tasks, for example question answering (Karpukhin et al., 2020) and co-reference resolution (Lee et al., 2017). In this workshop, aimed at linguists, participants will be familiarized with the basic concepts of count-based distributional models of lexical meaning (largely following the structure of Turney and Pantel, 2010) and will become acquainted with the more recent implementations of word-embeddings such as Word2Vec (Mikolov et al., 2013) and Fasttext (Bojanowski et al., 2017), as well as the basics of contextualized embeddings that are retrieved from transformer models like GPT-3 (Peters et al., 2018; Radford et al., 2019). In this workshop, we will specifically look at and replicate an intuitive and fun application of distributional semantics, namely Semantle. By trying to understand and implement what Semantle does and what makes it fun, we will try to gain a solid grasp of the representations that power most NLP technologies in 2022.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

[*] Participants need a laptop to do this workshop

[1] https://semantle.com/

Clark, S. (2015). Vector Space Models of Lexical Meaning.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *1952-59*, 1–32.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L. Y., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *ArXiv*, *abs/2004.04906*.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 188–197. https://doi.org/10.18653/v1/D17-1018

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Intell. Res.*, *37*, 141–188.