Between Digital Text and Language Model: Role of Context in Language Sampling, Segmentation, and Representation

Summary

This thesis investigates the interface between natural language and computational language models through three themes: corpus as language environment, tokenization, and role of context.

The first theme builds on the word frequency effect, the observation that more frequent words are processed faster by humans, and treats a text corpus as a proxy for a person's language environment. Chapter 2 proposes a method to move beyond treating a corpus as a monolithic collection of documents by decomposing it into latent thematic components using latent Dirichlet allocation (LDA). Word probabilities are then re-estimated as a topic mixture tuned to fit lexical decision reaction times, enabling personalisation at the group or individual level. Chapter 3 transfers these ideas into NLP, where language data are Zipfian and low-frequency expressions behave unreliably across contexts. The chapter introduces an evaluation method for tokenizers that divides test data into LDA-derived topical sections and measures performance degradation as a function of distributional distance from the training corpus, making evaluation less sensitive to the choice of evaluation data. It also proposes a vocabulary-building regime that uses per-document contextual diversity statistics to favour broadly reusable subword expressions over domain-specific ones, yielding tokenizers expected to generalize better across contexts.

Tokenization, a low-level mechanism through which a language model adapts to its language environment, is the central theme of this thesis. Chapter 4 compares five tokenization algorithms using downstream performance of LSTM-based classifiers on sentiment analysis and natural language inference, finding no clear winner, suggesting that choice of tokenization strategy is not crucial for discrimination tasks. Chapter 5 investigates the role of context in tokenization in two ways. Framing tokenization as a shortest-path search over a segmentation graph, the first experiment shows that adding contextual order affects the ranking of optimal segmentation strategies in a language-specific way, most notably for Turkish and Chinese. The second experiment on statistical word segmentation confirms that training data size dominates: bigram context yields small but non-negligible accuracy gains over well-trained unigram models, while trigrams add little. This supports the practical use of unigram-decoder tokenizers, which are computationally far more practical than higher-order decoders.

The final chapter departs from tokenization to examine whether surrounding sentence context improves unsupervised sentence representations. Using the Switchboard Dialog Act Corpus as a testbed, content-based representations (topic models, word-vector

averaging, LSTM auto-encoders) are compared against context-based Skip-Thought-style variants via a linear probe classifying dialog act tags. Context-based training offers only a small advantage overall. We speculate that the best-performing model likely benefits from implicit data augmentation rather than from the colocation principle itself, consistent with the subsequent dominance of content-based sentence representation methods in the field.

Taken together, the thesis demonstrates that representing a corpus as a mixture of latent components is useful both for modelling word-processing behaviour and for evaluating tokenizer generalisation. Context plays a real but modest role at both the token and sentence levels. More broadly, the work reframes tokenization as a low-level language modeling task which provides a convenient language-interfacing solution, raising further questions about fidelity, efficiency, and equity that point toward future research directions.

References

Hronský, R., & Keuleers, E. (2021). Word Probability Re-Estimation Using Topic Modeling and Lexical Decision Data. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*(43). https://escholarship.org/uc/item/2mm461qs

Hronský, R., & Keuleers, E. (2023). Role of Context in Unsupervised Sentence Representation Learning: The Case of Dialog Act Modeling. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8784–8792. https://aclanthology.org/2023.findings-emnlp.588/

Hronský, R., & Keuleers, E. (2024). Tokenization via Language Modeling: The Role of Preceding Text. *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024*, 23–35. https://aclanthology.org/2024.cawl-1.4/

Hronský, R., & Keuleers, E. (2022). Does the Choice of a Segmentation Algorithm Affect the Performance of Text Classifiers? *Proceeding of BNAIC/BeNeLearn 2022*.

Hronský, R. & Keuleers, E. A. (under review). Addressing Sample Bias of Language Data in Vocabulary Creation and Evaluation.