Automatically segmenting words into morphemes: A detailed comparison of unsupervised approaches applied to monolingual wordlists from different languages

Arne Rubehn
(University of Passau)

Keywords: morpheme segmentation, computational linguistics, morphology, review, unsupervised learning

Learning and understanding the morphology of a language is crucial for the successful processing of language data. For this reason, the task of unsupervised morpheme segmentation has received a fair share of attention in the Natural Language Processing community over the last decades. Often based on preliminary ideas by Harris (1955), a number of techniques for splitting words into their individual morphemes has been proposed. Even in the days of powerful large language models, scholars report that morphological preprocessing enhances the performance of several downstream tasks, especially in low resource and morphologically complex languages (Mager et al., 2022).
Segmenting words into morphemes would certainly also enhance computational methods for language comparison, enabling the identification of partial cognates and the reconstruction of complex etymologies that involve morphological processes like derivation or compounding. However, almost all methods proposed for automated morpheme segmentation require a large amount of training data (Eskander et al. (2020) being a notable exception), whereas multilingual datasets for historical language comparison are naturally small (List, 2019). In a pilot experiment, List (2019) shows that well-established methods like Morfessor (Creutz and Lagus, 2005) fail graciously when exposed to small-scale data, as it is usually found in the domain of mulitlingual computational linguistics. Since the success of techniques is highly limited by the availability of data, as well as by the morphological complexity of a language, Manova et al. (2020) conclude that the unsupervised learning of morphology is still a mostly unsolved problem.
In the talk, we will present our efforts to provide a unified implementation of several methods for unsupervised morpheme segmentation that have been proposed in the past. By testing these methods on monolingual word lists from different languages with up to 2,000 words (using a small preliminary dataset of 10 languages from different language families), we try to identify major strengths and major weaknesses of these methods and provide for the first time a detailed comparison of the performance of unsupervised morpheme segmentation methods on small wordlists.

References

Creutz, M. and Lagus, K. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0.* Helsinki University of Technology, Helsinki.

Eskander, R., Callejas, F., Nichols, E., Klavans, J. L., and Muresan, S. (2020). MorphAGram: Evaluation and Framework for Unsupervised Morphological Segmentation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 7112–7122.

List, J.-M. (2019). Open Problems in Computational Historical Linguistics. Invited talk presented at the 24th International Conference of Historical Linguistics (2019-07-01/05, Canberra, Australian National University).

Mager, M., Oncevay, A., Mager, E., Kann, K., and Vu, N. T. (2022). BPE vs. Morphological Segmentation: A Case Study on Machine Translation of Four Polysynthetic Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961-971.

Manova, S., Hammarström, H., Kastner, I., and Nie, Y. (2020). What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure*, 13(1):1–21.