INTEGRATING MORPHOLOGICAL ANALYSIS INTO MACHINE LEARNING MODELS FOR LANGUAGE PROCESSING

Sultanbayeva Oltinoy Omonbay kizi
Independent researcher

Annotation: This thesis explores the integration of linguistic morphological analysis into machine learning models for natural language processing (NLP). It focuses on how the inclusion of explicit morphological features, such as roots, affixes, and grammatical tags, can improve tasks like lemmatization. The study targets morphologically rich languages and uses both theoretical frameworks and experimental evaluation to support the findings.

Keywords: Computational linguistics, morphological analysis, machine learning, lemmatization, natural language processing, low-resource languages

Introduction

This study investigates how incorporating morphological information into machine learning models can enhance performance in language processing tasks, with a focus on lemmatization. Morphology, as a core component of linguistic theory, provides valuable structural information that many statistical models tend to overlook. By combining linguistic insights with computational techniques, the research aims to bridge the gap between theory and practice in NLP.

Literature analysis and methodology

The theoretical foundations of morphology, as presented by Aronoff (1976) and Booij (2005), highlight the structural role of affixation, root identification, and inflection in language. Cotterell and Heigold (2017) showed that morphological tagging benefits from character-level modeling across languages. Vania and Lopez (2017) explored how different input representations capture morphology in neural models. Jurafsky and Martin (2023) emphasize that subword modeling strategies like byte-pair encoding (Sennrich et al., 2016) are particularly helpful in handling morphological variation, especially in low-resource settings. Overall, the literature supports the claim that morphological awareness contributes to model robustness and generalization. This study used a comparative design involving two lemmatization models: a baseline (without morphology) and a morphology-aware version. Both were trained on parallel datasets in Arabic, Russian, and Finnish using annotated corpora.

Results

Across all languages tested, the morphology-aware model outperformed the baseline in lemmatization accuracy. In Arabic, it effectively handled clitics and

irregular verbs. In Russian, it showed stronger disambiguation across case-marked forms. In Finnish, it correctly processed long, agglutinative word forms. The model was more accurate, especially with rare or unseen word types, and better at generalizing in morphologically complex contexts.

Discussion

The findings affirm that morphological features significantly enhance machine learning performance in lemmatization. Integrating linguistic knowledge allows for better word structure modeling and improved results across different languages. This validates the relevance of linguistic theory in computational practice. However, limitations include reliance on annotated data and increased model complexity. Still, the benefits of linguistic integration outweigh these challenges, especially for under-resourced languages.

Conclusion

This thesis demonstrates that integrating morphological analysis into NLP models significantly improves lemmatization, especially for morphologically rich languages. It supports a hybrid approach that combines linguistic insight with computational methods, achieving more interpretable and accurate models. Future research may extend this framework to other NLP tasks and explore unsupervised morphology learning in low-resource settings.

References:

Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press.
Booij, G. (2005). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.
Cotterell, R., & Heigold, G. (2017). Cross-lingual character-level neural morphological tagging. In *Proceedings of EMNLP* (pp. 759–769). https://doi.org/10.18653/v1/D17-1079
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805
Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed., draft). https://web.stanford.edu/~jurafsky/slp3/
Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of ACL* (pp. 1715–1725). https://aclanthology.org/P16-1162
Vania, C., & Lopez, A. (2017). From characters to words to in between: Do we capture morphology? In *Proceedings of EACL* (pp. 751–761). https://aclanthology.org/E17-1071