BYU BRIGHAM YOUNG UNIVERSITY

Journal of Undergraduate Research

Morphological Parsing and Segmentation

JUNE 24, 2019 BY ADMIN

Kyle Roth and Deryle Lonsdale, Linguistics & English Language

Morphological parsing is a task where a computer recognizes the meaning that parts of a word contribute to its overall meaning and role in the sentence. Morphological segmentation simply means splitting words up into their component parts, and is simpler than parsing.

The result of my research is two-fold: I applied a VoCRF to morphologically parse a new Basque corpus, and demonstrated the effectiveness of a paradigm-based approach to morphological segmentation. Initially, I set out to improve upon the VoCRF algorithm to account for previously-known information; unfortunately, the expected improvements to the VoCRF algorithm could not be made because I was unable to determine a way to change the output of the algorithm into a finite state automaton. Due to this circumstance, my interest shifted to exploring morphological segmentation, and I improved a recent paradigm-based approach to segmentation.

Parsing

I applied existing VoCRF code (Vieira et al. 2016) to parse the Basque corpus from Universal Dependencies (UD) (Aranzabe et al. 2015) as well as to the Zientzia eta Teknologiaren Corpusa (ZTC), a morphologically tagged corpus of science and technology articles in Basque (Areta et al. 2006). To my knowledge, a VoCRF implementation has not yet been applied to this corpus. I achieved an accuracy of 80.45% on the UD corpus, and 71.28% on the ZTC.

VoCRF implementations generally reach state-of-the-art performance on various language tasks. Morphological parsing is a relatively difficult task, because the number of possible tags is high compared to other tasks like part-of-speech (POS) tagging or constituency parsing. This is why scores on morphological parsing tasks are generally lower than on other tasks. The UD parsing result can be considered a baseline when evaluating the results from other corpora.

The most difficult problem I solved was converting the ZTC corpus from the custom TEI format to the CoNNL-U format to work with the implementation written by Vieira et al. (2016). This was more difficult than expected for the following reasons:

1. POS tags and morphological tags were included on the same line with no separation.

2. Most tags had a description, but some were undecipherable.

3. Many attributes were described in a hierarchical format which was at odds with the "attribute=value" format of CoNNL-U.

The ZTC is much larger than the UD corpus, but I encountered overflow errors when trying to use even a majority of the data. I decided to massively reduce the size of the corpus. The smaller size resulted in some overfitting, evidenced by the fact that the model scored 27 percentage points lower on the test set than the training set. Because of the large tagset, running the algorithm on these corpora required more than 64 gigabytes of RAM. Training on a supercomputer node with 128 GB RAM at 2.3 Ghz, the UD dataset took 38 hours, and the selected subset of the ZTC took 11 hours.

Segmentation

After learning about morphological paradigms in my class on morphology and syntax, I found a recent paper (Xu et al. 2018) that described a paradigm-based approach to morphological segmentation. Their code marked an F1-score of 79.8% on the Morpho-Challenge, setting the current state of the art and demonstrating the robustness of paradigms in recognizing morpheme boundaries. While their implementation only searches for suffixes, I improved it to handle both prefixes and suffixes, but I was narrowly unable to get results from my implementation because of an unresolved bug in my code.

Future work

One of the downsides to the VoCRF is that it is not easily multi-threaded; it would be beneficial to write it to allow multiple processes to increase speed.I would also like to fix the overflow error that causes the Vieira implementation to fail on the entire ZTC. With regard to the segmentation task, it remains to be seen whether allowing for other sorts of morphemes (circumfixes, infixes, vowel harmonies, etc.) would further increase accuracy, especially on languages other than English.

I would like to thank those who sponsored my ORCA grant for giving me the opportunity to dive into these interesting problems. This research has helped guide my interests in linguistics and machine learning, and I look forward to working on similar problems in the future.

References

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, & Larraitz Uria. 2015. Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In 14th International Workshop on Treebanks and Linguistic Theories (TLT) 2015.

Nerea Areta, Antton Gurrutxaga, Igor Leturia, Ziortza Polin, R. Saiz, Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza, Aitor Sologaistoa & Aitor Soroa. 2006. Structure, annotation and tools in the basque ZT corpus. In International Conference on Language Resources and Evaluations (LREC) 2006. 1406-1411.

Tim Vieira, Ryan Cotterell, and Jason Eisner. 2016. Speed-Accuracy Tradeoffs in Tagging with Variable-Order CRFs and Structured Sparsity. In Empirical Methods in Natural Language Processing (EMNLP) 2016. 1973-1978.

Hongzhi Xu, Mitch Marcus, Charles Yang & Lyle Ungar. 2018. Unsupervised morphology learning with statistical paradigms. In International Conference on Computational Linguistics (COLING) 27. 44-54.

FILED UNDER: COLLEGE OF HUMANITIES, LINGUISTICS AND ENGLISH LANGUAGE, ORCA-2018