Matthew Tang

Technology Review on possible improvements to POS tagging

Introduction:

POS tagging is an important part of almost every natural language processing application. The process itself provides a lot of important information for the models to correctly understand and analyze the text. Therefore, it is important to make POS tagging as accurate as possible. Through our class lectures, we learned that the accuracy for POS is relatively high for each token and it is almost impossible to reach 100 percent accuracy for POS tagging.

Even with this relatively high accuracy for each token at around 96% [1], it does not tell the whole story because of how many easy to classify parts there are. Also, some words are more important to classify, so a seemingly high accuracy could still fail to correctly tag the important part of a sentence to understand the text. Therefore, this review looks into some papers which look into the problems with current POS methods and what can be improved to make the accuracy as close to 100 percent as possible. Another important thing to note is that while token accuracy is high, sentence accuracy is often lacking in comparison making it a somewhat misleading statistic.

Body:

To understand the solutions to improving POS tagging, we first need to understand what is flawed with the current approaches to POS tagging. From the papers, the difficulties that affect POS tagging are unknown words, lack of context, and corpus quality/size [1] [2]. For unknown words, this means words that do not show up in training data, making it difficult to identify the tags as it has not seen them before. Lack of context occurs when a word can have multiple tags depending on the context and the training can make it select the wrong option. The corpus quality can affect POS because there are bound to be human annotated errors/inconsistencies in a large training set which can affect the models greatly. Corpus size can also be a problem, because a large training set could take much longer to train and converge.

Now that we have identified the problems with current POS approaches, let us examine some of the proposed methods for improving the accuracy. In one paper, they based their research off of the Penn treebank, a large annotated corpus of English [4]. While examining the errors of the models on this corpus, they found that many of the errors/inconsistencies could be solved through deterministic linguistic rules. One example of this is when words that are not nouns are combined in a proper noun, such as “United” in “United States”. This case normally makes it difficult to determine whether “United” is an adjective or part of the proper noun. Normally, the model would typically label them as NNP due to the verbal agreement, but adding a rule would favor the correct label of NNPS [2]. This rule, along with many other rules similar to it was the basis of this paper. The results are shown in this table:

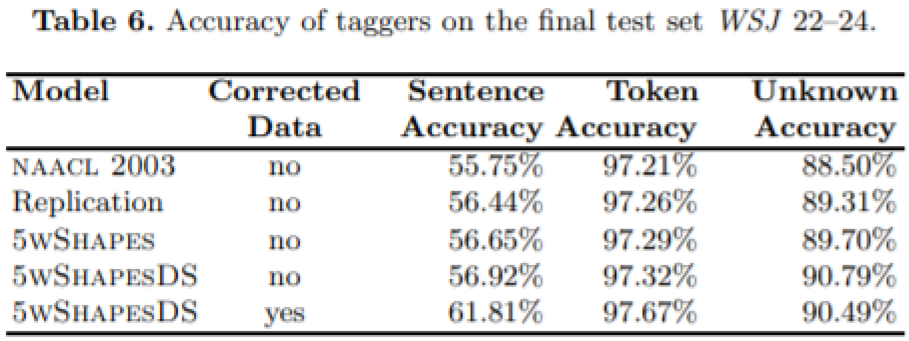

Table 6. Accuracy of taggers on the final test set *WSJ* 22–24.

Model	Corrected Data	Sentence Accuracy	Token Accuracy	Unknown Accuracy
NAACL 2003	no	55.75%	97.21%	88.50%
Replication	no	56.44%	97.26%	89.31%
5wShapes	no	56.65%	97.29%	89.70%
5wShapesDS	no	56.92%	97.32%	90.79%
5wShapesDS	yes	61.81%	97.67%	90.49%

Table from reference 2

In the table, the 5wShapes are variations that use the suggested additional rules to tag POS. As we can see from this table, by adding these additional rules to the model, it was able to improve on the accuracy at all levels, and an especially significant improvement on sentence level accuracy. This seems to be a successful improvement based on linguistic rules, and there are still many more possibilities mentioned in the paper to possibly improve on.

In another paper, they proposed using RAGE (Rapid Application Generation Engine) AI Hybrid POS Tagger with an additional step of applying linguistic based rules before processing. This approach, similar to the other paper, makes use of linguistic rules to better tag the POS in addition to grammar rules to address grammatical inconsistencies. To evaluate the effectiveness of the RAGE-AI Hybrid POS Tagger, they tested it on three different datasets, and found that the token accuracy improved slightly while sentence accuracy increased a lot more for all three [1]. The three datasets they used were the Penn Tree bank, also used in the other paper, RAGE Reuters-110, a corpus based on news articles from Reuters, and RAGE PubMed-110, a corpus based on abstracts from PubMed. Based on these results, it seems like linguistic rules can play a major role in improving the results of POS tagging and is something we should look at more to improve POS tagging.

Conclusion:

POS tagging is a vital part of natural language processing, and improving it would propagate through the process. Even though token accuracy may seem high, the sentence level accuracies can still use improvement. This makes improving POS tagging still a relevant problem to address. As we can see from the results of both papers mentioned above, linguistic rules can help improve the results of POS tagging. These linguistic rules provide the framework to improve on the issues faced when POS tagging.

References:

[1] Jatav, V., Teja, R., & Bharadwaj , S. (2017). *Arxiv.org*. Retrieved November 1, 2021, from https://arxiv.org/ftp/arxiv/papers/1708/1708.00253.pdf

[2] Manning, C. (2011). Manning tagging - the Stanford natural language processing ... https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf

[3] Purnawirman, P. (n.d.). *What is part-of-speech tagging and how can I use it?* super.ai Blog. Retrieved November 1, 2021, from https://super.ai/blog/what-is-part-of-speech-tagging-and-how-can-i-use-it

[4] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics 19 (1993) 313–330 [4]