LING 413 Project 1: Supervised Model On POS Tagging

Hao Qi (haoqi3@illinois.edu)

Abstract

Part-of-Speech (POS) tagging remains a important task in NLP due to its crucial nature the branches of the field. In English, the task of POS tagging involves parsing through common stopwords: "the," "that," "which," etc. This works seeks to find if determiners or a lack thereof has any implications on the task of POS tagging. The experiment makes use of the UD English GUM Corpus and bigram taggers from NLTK. We show that for a simple bigram model tasked with POS tagging, the performance of a model trained without determiners consistently performs at or below a model trained with determiners.

1 Introduction

Part-of-Speech (POS) tagging serves as a cornerstone task in Natural Language Processing (NLP), facilitating various downstream applications such as parsing, sentiment analysis, and machine translation. It involves assigning grammatical categories to individual words in a sentence, enabling computers to comprehend and analyze natural language text more effectively. In English, POS tagging necessitates parsing through common stopwords like "the," "is," and "and," which play pivotal roles in sentence structure and meaning.

Despite its apparent simplicity, POS tagging remains intricate due to nuanced variations and contextual dependencies inherent in natural language. The role of determiners – words that precede nouns and specify their reference within a sentence - significantly contribute to sentence structure and meaning, prompting questions about their impact on the abilities of POS tagging algorithms. In this paper, we delve into the exploration of determiners and their implications for POS tagging in English, aiming to discern whether their presence or absence influences POS tagging model performance.We seek to utilize portions of the Universal Dependencies English Georgetown University Multilayer corpus to conduct experiments to elucidate the relationship between determiners and POS tagging accuracy. Through analysis of the weighted precision, accuracy and recall of these models, we seek if it is plausible to develop robust POS taggers without that minimize the need of stopwords.

2 Corpus

The UD English GUM corpus was chosen as the primary dataset for this project due to several key attributes that make it particularly suitable for POS tagging research. Firstly, the corpus provides pre-annotated POS tags, which significantly streamlines the process of training and evaluating POS tagging models, eliminating the need for manual annotation efforts. Additionally, the corpus offers pre-divided datasets for testing and training, allowing for standardized evaluation procedures and ensuring consistency in model assessment. These features enhanced the efficiency and reliability of experimentation as well as provide a streamlined and reconstructable method for easy changes.

3 Classification and Data

The classification problem at hand involves Part-of-Speech (POS) tagging used in classifying natural language, where this case had a predominantly written form. The objective of our tagging models is to accurately label each word with its corresponding part of speech, such as noun, verb, adjective, adverb, etc according the the Universal Dependencies UPOS set. Data was taken from the UD English GUM Corpus and processed to be in `.conll` format. For both the training and testing data, an additional dataset was made without the presence of determiners as a way to measure the performance of the Non-Determiner tagger model. The `.conll` files were further processed to isolate only needed information: sentences were made into lists of tu-

ples, with each tuple representing its respective word and POS tag.

4 Models

For the experiment, we set up 2 classifiers, identical except for the data they were trained on. The model relies on a bigram classifier for general cases of tagging, falling back on the default tag being `NOUN`, since it is the most common part of speech. They both assign grammatical categories to individual words in a sentence: the POS tags provide information about the word's syntactic function and its role within the sentence structure. The POS tagger is trained using annotated data in the .conllu format, which contains sentences with corresponding POS tags. Each token (word) in the training data is associated with its respective POS tag. The NLTK library is employed for training the tagger which is then exported into a `.pkl` minimizing the need to retrain. Finally, we export performance metrics: precision, recall, accuracy and the confusion matrix into an `.txt`.

5 Data

Metric	**Det**	**Non-Det**
Precision (Weighted)	89.51%	86.43%
Precision (Unweighted)	89.37%	81.76%
Recall (Weighted)	87.77%	79.52%
Recall (Unweighted)	81.88%	76.43%
Accuracy	87.77%	79.52%

Table 1: Scores for Data with Determiners

From **Table 1** we see that in all the cases using a testing set with determiners still present, our **Det** classifier significantly outperforms our **Non-Det**. However, this is case is to be expected since the **Non-Det** model is not aware of this word class at all. Nevertheless, the performance of **Det** establishes a baseline performance for consecutive tests on testing data with determiners removed.

Metric	**Det**	**Non-Det**
Precision (Weighted)	88.45%	88.44%
Precision (Unweighted)	82.73%	87.54%
Recall (Weighted)	86.21%	86.43%
Recall (Unweighted)	81.04%	79.95%
Accuracy	86.21%	86.43%

Table 2: Scores for Data without Determiners

Interestingly, we notice that in data without determiners, the **Non-Det** model improves its performance in unweighted precision. While this may be of interest, it still underperforms the baseline set in **Table 1**. Furthermore, this improvement in unweighted precision can be attributed to the lack of a **DET** row and columns in the confusion matrix[1]. When accounting for the weights, due to the number of determiners used, we see that this discrepancy is minimized. [1]

6 Conclusion

Our experiment has provided compelling evidence that training a POS tagger without considering determiners leads to lower performance across various metrics. The consistent trend observed across different evaluation criteria underscores the importance of accounting for determiners in POS tagging tasks. Our findings suggest that determiners play a crucial role in providing contextual cues in NLP algorithms, thereby enhancing the accuracy and robustness of POS tagging models.

However, given how human can recognize parts of speech even without determiners, it is noteworthy to emphasize that our study primarily focused on simple POS tagging models, specifically employing bigram taggers. While our results demonstrate a clear performance gap between models trained with and without determiners, it is important to acknowledge the potential limitations of our experimental setup. The scope of our investigation did not encompass more advanced or sophisticated POS tagging architectures, which may yield different outcomes.

Acknowledgements

We would like to express gratitude towards Dr. Dunn for his help and insights.

[1] I have no idea how to put a confusion matrix this big into LaTeX, but it is in `results.txt`