6
Conclusion

In Natural Language Processing, one of the most important tasks is *syntactic parsing*, where the structure of a sentence is inferred according to a given grammar. Syntactic parsing tells us how to determine the meaning of the sentence from the meaning of the words in it. Recently, the NLP community has drawn its attention to syntactic parsing based on *Dependency grammars*. Dependency Parsing, due to its simple but powerful structure, has been proved to be useful in a plethora of applications.

In Dependency Parsing, each word has a *direct head-dependent* relation to a word it depends on. Therefore, it generates a graph, where the nodes are the words in the sentence and the edges represent typed dependency relations between them. Thus, the Dependency based syntactic parsing task consists in identifying a head word for each word in an input sentence.

We propose a token classification approach to the Dependency Parsing problem by creating a special tagging set that helps to correctly find the head of a token. To evaluate our modeling we use three *corpora* that were made available by the occasion of CoNLL 2006 Shared Task on Dependency Parsing: Danish, Dutch and Portuguese. Although the number of possible classes can be very high, our tagging style has shown great adherence to dependency parsing, covering more than 95% of the evaluated *corpora* with only 20 classes.

Additionally, we introduce a statistically built baseline system by applying the most frequent tag in the training set, given a token's part-of-speech.

Using this tagging style, any classification algorithm can be trained to identify the syntactic head of each word in a sentence. Since our classification model treats projective and non-projective dependency graphs equally, classification algorithms can be applied straightforwardly, thus avoiding pseudo-projective approaches. As far as we know, this is the first reported study that effectively treats dependency parsing as a token classification problem.

To evaluate our approach effectiveness we apply the ETL algorithm, a transformation-based entropy guided algorithm that achieves state-of-the-art in some NLP tasks. An ETL model is trained and evaluated for a wide range of

parameters, as well as many derived features are created and tested. The One Model approach, with the best set of parameters and derived features, achieves above average accuracy in two of the three languages.

Our tagging set has the advantage of allowing us to split the dependency parsing into three subtasks, analogous to the three information built into the tags: find if the token's head comes before or after the token; identify the token's head part-of-speech; and find the distance between the token and its head. In this work, we create an ETL model for each one of these subtasks and a final ETL model to improve the results of joining the subtasks output. This subtask approach consistently improves our results, achieving above average performance in all three languages.

Furthermore, the Portuguese *corpus* provides information about clause boundaries that can be used to generate phrase chunking information. Clause and chunking information are used to generate features to improve our ETL models. Both information significantly improve our one ETL model and our subtasks approach.

Our error analysis also suggests that better results can be achieved, as further classification algorithms can be evaluated. Support Vector Machines, Hidden Markov Models, and Conditional Random Fields are known to achieve excellent results in NLP and remain yet to be tested within our modeling approach. Also, recent works show that ETL can be used as a base learner for ensemble methods, thus enhancing its performance.

Finally, our findings indicate that this is a promising modeling, achieving results comparable to the state-of-the-art, while using a much simpler modeling approach.