6
Conclusions

The whole scenario of Natural Language Processing can be thought as a pipeline composed of a sequence of tasks. In this pipeline, more fundamental problems are solved first as preprocessing steps, and their results are used as input to enhance the performance of more complex applications. Among the former group is text chunking, whose output has been shown to be relevant to many problems in several languages.

Because text chunking aims at identifying non-overlapping phrase structures in sentences, it exposes straightforward syntactic relations that Machine Learning models can take advantage of as a feature. Initially, Machine Learning approaches for noun phrase chunking have been studied more deeply because of the more explicit benefits of this subtask. Research focus on other types of phrases started appearing progressively, until it culminated in the CoNLL-2000 shared task. It proposed the creation of extractors trained with an English corpus containing an exhaustive set of phrase types. To this date, as far as our knowledge extends, no similar study has been done for the Portuguese language.

We report the impact of different chunk definitions, i.e., sets of chunk types, for Portuguese. These definitions are not intended to be linguistically precise, nor to comprehensively cover all possible types. Rather, they are meant to provide highly informative syntactic data to Machine Learning models. To generate the chunks corresponding to a given definition for a corpus with full syntactic parsing information, we apply a heuristic that uses the phrase structure of its sentences. Using the *Bosque* corpus from the *Floresta Sintá(c)tica* project, plus its derived chunks, for the clause identification and dependency parsing tasks, we attest the value of those chunk definitions as a feature.

We also propose two Machine Learning approaches for the text chunking task in Portuguese, both based on the Entropy Guided Transformation Learning algorithm. The first approach is a direct ETL classifier with self-contained training and extraction phases. The second is a sequence of three subtasks: identification of chunk starting tokens, identification of chunk ending tokens,

and complete chunk identification. For this second approach, we solve the first two subtasks through the creation of ETL models for each, and use the results from these previous subtasks to tackle the third one using a heuristic. The created models are fine-tuned by varying specific ETL parameters and testing the influence of different derived features. We employ the *Bosque* corpus once more in order to evaluate these approaches. Additionally, we apply them for three of the aforementioned definitions.

After determining the effectiveness of both approaches, we verify that the subtasks approach outperforms the direct one. Moreover, we compare the performance of our extractors with another Machine Learning model based on maximum entropy. Using the same chunk definitions and training and test corpora, we determine that our approaches yield better results. Besides that, since our proposed models are not overly sophisticated, their results are comparable to the results of their counterparts built for the English language and can be improved upon.

Since we establish our chunk definitions mainly for engineering purposes, a natural extension of this work, from a linguistic point of view, would be to define a complete set of chunk types similar to the one made available for the CoNLL-2000 shared task. It is possible that other advanced tasks may benefit from those additional chunk classes, even if these classes do not occur as frequently as the ones we consider. In addition, the impact of chunks in several other NLP tasks in Portuguese is yet to be analyzed, but the results we present here seem to indicate a good potential overall.

We determine that NP chunks are the most frequent ones for any definition, and that they result in one of the greatest error rates when using our extractors. Therefore, one possibility to boost their effectiveness in the future is to focus on strategies to better recognize NP chunks, like adding relevant derived features. Finally, since it is known that techniques such as Support Vector Machines have top performance on this task for other languages, another reasonable future work would be to investigate their behavior on Portuguese corpora.