1
Introduction

The huge increase in the volume of textual data during the last decades started to reveal our necessity for automatic information processing, categorizing and filtering. It is no coincidence that, along with the exponential growth of the World Wide Web, tools such as search engines became an indispensable part of our daily lives while we look for appropriate information during our work and leisure times.

More recently, online social networks started to gain prominence as one of the people's main activities on the Web. On Facebook[1], as an example, an astounding amount of textual information was generated every 20 minutes along 2010 (4):

- 1,587,000 wall posts;
- 1,851,000 status updates;
- 4,632,000 personal messages;
- 10,208,000 comments.

Applications like advertising and personal product recommendations are currently exploiting this massive conglomerate of unstructured data, supported by the recent advances in Natural Language Processing (NLP). By revealing the syntactic and semantic structures of sentences in these corpora, among other tasks, NLP provides access to otherwise hidden knowledge contained in them. However, NLP research is still in full progress, and there is enormous untapped potential behind all that data.

The field of NLP comprises many different problems, and two easily recognizable examples are: part-of-speech (POS) tagging, which is the process of assigning a POS or another lexical class marker to each word in a text (33); and named entity recognition (NER), the problem of finding all proper nouns in a text and to classify them among several given categories of interest (45). A good analogy to help visualize the multitude of NLP tasks and how they relate to each other is a pipeline: the ones with comparatively low complexity are tackled first in order to help solving the more sophisticated ones. Thus, for

[1] `http://www.facebook.com`

example, POS tagging is typically solved first as a preprocessing step to the NER task.

The initial approaches to NLP problems tried to determine formal rules that described the formation of linguistic structures. However, we now assume that it is just not possible to provide an exact and complete characterization of well-formed utterances of a language that cleanly divides them from all other sequences of words, which are regarded as ill-formed utterances (40). Presently, Machine Learning techniques are regarded as an excellent fit for NLP. Machine Learning is programming computers to optimize a performance criterion using example data or past experience (6). Since they inspect great quantities of data, those techniques are able to discover patterns in this data and generalize them to future examples without relying on static rules.

One remarkable advantage of Machine Learning models for NLP tasks is that they are considerably language independent, and so a model fit for a specific language generally performs well for the same problem in another language. But that comes with a cost: they need annotated corpora with enough examples to work adequately, and the availability of those corpora for each language varies greatly. This annotation has been almost exclusively made manually in the past, but we have reached a point where such analysis of data can no longer be done by people, both because the amount of data is huge and because people who can do such analysis are rare and manual analysis is costly (6).

The yearly Conference on Computational Natural Language Learning (CoNLL)[2] has been calling researchers to work on Machine Learning solutions for tasks based on typical NLP problems, and making annotated corpora on English and some other languages available for that purpose. These shared tasks have intensely contributed to the advance of the current research state on problems such as clause identification (62), named entity recognition (58, 61), semantic role labeling (14, 15) and dependency parsing (10).

In the context of Machine Learning applied to NLP, previously solved tasks provide the input to subsequent ones in the aforementioned metaphorical pipeline in the form of features. A good feature is one that exposes patterns to the learning algorithm that would otherwise not be easily identified. Text chunking is one of the tasks that contribute greatly to the performance of more complex ones. Its goal is to divide a sentence into disjoint groups of syntactically correlated words, called chunks. These groups are directly related to the linguistic concept of phrases, but they are made non-recursive to simplify their representation. That way, we can classify them as noun phrase chunks,

[2] http://www.ifarm.nl/signll/conll/

verb phrase chunks and prepositional phrase chunks, for instance. Since it generates plain structures, its output may be directly used as a feature by other models.

Text chunking was the main theme of the CoNLL-2000 shared task, which provided an English corpus (60). In addition, many following shared tasks have provided chunk data as a feature. Since then, several Machine Learning approaches have been proposed for this task. Among the most effective models are those that utilize Support Vector Machines, Hidden Markov Models and Winnow algorithm-based approaches.

Although text chunking has been thoroughly researched for languages such as English, the Portuguese language has not yet received the same attention. To the best of our knowledge, no linguistically supported comprehensive definition for chunks, similar to the one provided with the CoNLL-2000 English corpus, has been established for Portuguese. Most studies for phrase identification in Portuguese corpora limit their scope to noun phrases. The present work presents a heuristic to extract phrase chunks from corpora that contain previously annotated full syntactic parsing information. Our goal is to analyze the impact of different chunk definitions on other tasks. These definitions depend on which phrase types from the original full parsing we choose to consider or ignore. Since each task may take advantage of phrasal information in different ways, this study allows us to verify which definition is most relevant to the tasks taken into account.

Additionally, we propose two Machine Learning models for the automatic extraction of chunks for Portuguese, using the chunks yielded by our heuristic as golden values. One of the models is a direct classifier, and the other divides the chunking task into three subtasks. These models are based on the Entropy Guided Transformation Learning (ETL) algorithm, which has been previously applied with success on the text chunking task for other languages. Finally, we evaluate their results and compare them to the ones of another Machine Learning model operating under the same conditions, verifying that our approach shows significantly better performance.

Using the chunk values extracted by our proposed heuristic directly as a feature, we are able to increase the $F_{\beta=1}$ score of a clause identification system from 67.28 to 74.13, and the score of a dependency parsing system from 87.50 to 89.04. Furthermore, our best model, using automatic POS tags and based on a chunk definition with five distinct chunk types, achieves an $F_{\beta=1}$ score of 87.95.

This work is structured as follows. In chapter 2, we describe the text chunking problem in more details, discussing its motivation, the best known approaches that solve it and the state of its related research in the Portuguese language. Chapter 3 gives more background about our research, describing the corpus used and the tagging styles for the chunking task, and later presenting our chunk derivation heuristic for Portuguese, the chunk definitions tested and the subsequent tasks for which we provide a chunk feature. In chapter 4, we give more information about the Machine Learning techniques used in our extractors, and we also dissect these classifiers, explaining how they approach the chunking task and what is done to optimize their results. Chapter 5 demonstrates our experiments and the results achieved. Finally, in chapter 6, we present our final remarks and future directions to improve upon this work.