1
Introduction

In the last few years, we have noticed a huge increase of textual data on the Web. The main reason for this is the popularization of the Internet. As a consequence, it is difficult to find desired information without spending some time searching for it on search engines such as Google.

Today there are applications exploring this huge amount of unstructured data in order to make users' lives easier, reducing the time wasted looking for information. Advertising systems and personal product recommendation systems are clear examples of this trend. Such applications are supported by Natural Language Processing (NLP) which, among other things, identifies syntactic and semantic structures from the text with the aim of extracting previously hidden information.

NLP research started in the 1950s. Up until the 1980s, most NLP systems were based on handwritten rule-based methods, which were incapable of expressing language richness. However, in the late 1980s, there was a breakthrough in NLP research, when Machine Learning (ML) techniques started being applied to text processing. They proved to be adaptive to different writing styles, producing more reliable results.

The ML paradigm uses general learning algorithms in order to analyze large volumes of data. Such algorithms learn patterns in data and apply them to new inputs.

ML is language independent, that is, we are able to apply to Portuguese a model originally prepared for English, making just a few changes to it.

Many important ML problems involve the prediction of complex structures which comprise interdependent variables.

Most of the best performing systems for such structured problems are complex ML systems which combine several binary classifiers. Additionally, in order to consider the natural interdependencies among output variables, the binary classifiers are trained by task-specific strategies which share information or enforce constraints among the basic classifiers.

In the last few years, ML methods that *directly* solve structured problems have emerged. They are called *structured learning* methods. They have been

successfully used to model many NLP tasks (1, 6, 4, 40, 22).

In order to approximate user to information, we propose a Quotation Extraction (34) system for Portuguese, which consists of identifying quotations from a text and associating them to their authors. Our system handles direct and mixed quotations for Portuguese.

The proposed system is very useful in several situations. A voter may want to see what his or her candidate is saying in the media. A person may want to see the statements of the PETROBRAS[1] president before acquiring its stocks. A Natalie Portman fan may want to see what she says about her last motion picture as well as about her baby. A news portal may release the last statements of every public person or company. There are many other situations our system may be applied to.

Quotation Extraction has been previously approached using different techniques and for several languages. The *NewsExplorer*[2] system, based on lexical rules, extracts quotations from multilingual news (30). The *Sapiens* system, based on syntatic rules, extracts quotations from news wires in French (5). The VERBATIM[3] system, based on speech act rules, extracts quotations for Portuguese (34). The EVRI[4] portal offers a Quotation Extraction API for English news feeds (21). Their approach is based on rules which use several linguistic features automatically provided by standard auxiliary processors.

Our proposal differs from previous work since we use ML to automatically build specialized rules instead of human-derived rules. We use two algorithms, *Entropy Guided Transformation Learning* (ETL) (9) and *Structured Perceptron* (6). ETL is used to predict a label sequence and uses information about tokens in a token neighbourhood. Structured Perceptron is used to predict complex and interdependent outputs like sequences, trees and even more general graphs.

The Structured Perceptron predictor is based on an optimization problem whose objective function is linear in the input-output feature vector. Algorithms of this kind are called structured learning methods in the literature.

Since we employ supervised ML algorithms, we need an annotated corpus to train and evaluate the system. In order to accomplish this task, we build the GLOBOQUOTES corpus, with news extracted from the GLOBO.COM portal. We generate the golden features for entities, coreferences, quotations and associations between quotations and authors. Moreover, we include the part-of-speech (POS) annotation in the sample corpus using a state-of-the-art tagger

[1] A Brazilian oil company
[2] http://press.jrc.it/NewsExplorer
[3] http://irlab.fe.up.pt/p/verbatim
[4] http://www.evri.com

(11), based on ETL as well. After producing the annotations, we divide the sample corpus into two sets, training set and test set.

In our work, we create three models to tackle the Quotation Extraction task. The first model is Structured Perceptron based on weighted interval scheduling problem (SP-WIS). In this model, we find a maximum-weight subset of non-overlapping tasks, where each task represents a combination of quotation and author candidate. The second model is ETL. In this model, we divide the original task into two subtasks, quotation identification and association between quotation and author. ETL predicts the corresponding label sequence for each subtask. The last model is a baseline system, created using a rule-based approach.

In Table 1.1, we present the quality of our models assessed in the test set. SP-WIS obtains an $F_{\beta=1}$ score of 76.80%, which is an error reduction of 31.70% compared to the baseline system. Comparing to ETL, SP-WIS reduces errors by 19.28%. The performance of our models cannot be directly compared to previous work, since the corresponding corpora are not publicly available.

Table 1.1: Quotation Extraction performance on the test set

Model	*Precision (%)*	*Recall (%)*	$F_{\beta=1}$*(%)*
SP-WIS	83.24	71.49	76.80
ETL	69.44	73.17	71.26
Baseline	64.35	67.80	66.03

The remaining of this dissertation is organized as follows. In section 2, we describe the Quotation Extraction task. Section 3 describes the Machine Learning algorithms used in this work. We report our models in section 4 and experiments in section 5. Finally, in section 6, we present our conclusions and future work.