Part-of-Speech Tagging for Portuguese Texts

Aline Villavicencio[†]
alinev@inf.ufrgs.br
CPGCC - Federal University of Rio
Grande do Sul

Nuno M. C. Marques[‡]
nmm@fct.unl.pt
FCT- CRIA/UNINOVA - New
University of Lisbon

José Gabriel P. Lopes[‡]
gpl@fct.unl.pt
FCT- CRIA/UNINOVA - New
University of Lisbon

Fabio Villavicencio[†]
fabiov@inf.ufrgs.br
CPGCC-Federal University of Rio
Grande do Sul

Content Areas: Natural Language Processing, Part-of-Speech Taggers, Probabilistic Models

1. Abstract

In this paper we will describe the work that is being cooperatively done by Portugal and Brazil. It uses Statistical Methods for Natural Language Processing. Namely, we will focus on the problem of Part-of-Speech (POS) Tagging. POS Tagging is a recent and successful technique for assigning each word in a sentence its correct POS tag. This technique can achieve more than 96% of accuracy, even with unseen untagged texts. All steps involved in this process will be described as well as the problems faced. Besides, we will present the stochastic approach to POS Tagging, which treats the generation of tag alignments as a probabilistic problem. Finally, we will report the results achieved by using these kinds of techniques for Portuguese texts.

2. Introduction

A Corpus is defined as a collection of texts written in a given language. When it has some linguistic features associated with its constituents, for instance, their part-of-speech tags, it is called a Tagged Corpus. This kind of annotation can help to show the patterns that occur in that corpus.

In the project being developed, we are using two Portuguese Corpora: the Lusa Corpus, which contains news from Lusa Agency from Portugal and the Radiobras Corpus, that contains news from the Radiobras Agency from Brazil. But the results presented in this paper were obtained using only the Radiobras Corpus.

One of the uses of a Tagged Corpus is to train a Part-of-Speech Tagger. A Part-of-Speech Tagger is a program that "learns" automatically the linguistic patterns from a given Corpus. It is done through the computation of the probabilities of

[†] Curso de Pós Graduação em Ciência da Computação- Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500 - Porto Alegre - Brasil - 91501-970 Phone/Fax: +55-51-3428687
[‡] Faculdade de Ciências e Tecnologia- CRIA/UNINOVA - Universidade Nova de Lisboa
Quinta da Torre,2825 - Monte da Caparica - Portugal Phone:+351-1-350-0281

Work partially suported by a PhD Scholarship by JNICT-PRAXIS XXI/BD/2909/94

word-tag alignments. After finishing the learning process, the Tagger is ready to use its acquired knowledge to tag any unseen untagged Corpus. Thus the work of the tagger is, given an input word sequence, to assign each word its corresponding part-of-speech tag. Hence its output is a sequence of tags.

These programs have been presenting very good results (see [5], [7] and [14 among others). They can tag at least 96% of the words correctly, with minimal restrictions on the input text. And they achieve this performance using only modest resources of space and time.

In this paper we will briefly explain the process of Part-of-Speech Tagging using stochastic methods. After that, we will comment on some tag sets commonly used and mainly on the tag set we used. We will also explain the architecture of the system developed. Finally, we will present some preliminary results of this work.

3. Part-of-Speech Tagging

Given a sentence *W*, which can be defined as a string of words, composed of $w_1,w_2,...,w_n$. Part-of-Speech (POS) Tagging can be described as the process of assigning each word w_i of *W* a corresponding tag t_i from the *T* set of tags. These tags must be previously defined by the user. For each sentence *W* we get an alignment:

$$(W,T)= w_1t_1,w_2t_2,...,w_nt_n.$$

However, because there might be more than one tag for each word, the sentence can have more than one alignment [2], [14]. But, from all possible alignments for a sentence only one should be considered correct. And the tagging process should select this correct alignment (the most probable one).

For example, consider the following English sentence[1]:

He can	can	a	can.	
ProN	Mod	Mod	Det	Mod
	Nn	Nn		Nn
	Vb	Vb		Vb

Figure 1 - POS tag alignments

There are 27 possible alignments, but the correct one is:

He can	can	a	can.	
ProN	Mod	Vb	Det	Nn

Figure 2 - Correct alignment

The performance of a tagging procedure can be measured by using two functions, following the ideas described by [14]:

- the first function computes the percentage of sentences correctly tagged in a given Corpus;
- the other one determines the percentage of words correctly tagged in the Corpus.

The first measure will always produce a lower result than the second one, because a sentence will only be tagged correctly if all the words in this sentence have the correct tag.

[1] ProN - Personal Pronoun, Mod - Modal Verb, Det - Determiner, Nn - Noun and Vb - Verb.