THE IMPLEMENTATION PROCESS OF A STATISTICAL PARSER FOR BRAZILIAN PORTUGUESE

Andréia Gentil Bonfante and Maria das Graças Volpe Nunes
Instituto de Ciências Matemáticas e de Computação
USP São Carlos, Brazil
andreia@nilc.icmc.sc.usp.br, mdgvnune@icmc.sc.usp.br

Abstract

This paper presents the project of a statistical parser for brasilian portuguese, which is based on the model proposed in [6], to a generative model of lexicalised dependency grammar rules proposed in [3].

1 Introduction

Although statistical parsers have been produced with very satisfactory results for english, for portuguese this hasn't been verified yet.

In this paper we report the initial steps for building a generative parsing system to handle portuguese tagged sentences following the annotation system based on the dependency grammar paradigm proposed by [3]. Although it is a flat-annotation system, we have used some Perl scripts for transforming the annotated sentences to a constituent tree structure, producing a suitable hierarchy for the parser.

The treebank used in this project has 130.000 authentic portuguese sentences with an average of 20 words per sentence, which were extracted from NILC corpora [1]. Ten percent of the selected sentences will be used for tests. The gathered probabilities from sentences are stored in a relational database, where each element which is part of a rule is stored together with its morphologic and syntactic tags, and, still, the head features to which the element is associated, as its parent in the hierarchy.

In the pre-processing step of a new sentence, before parsing, the words are tagged by a brazilian portuguese part-of-speech (POS) tagger [2], eliminating the first step of probability estimates.

Following, we describe some features of the parser generative model and the sentence annotation system.

2 Probabilistic generative model

The Collins generative parser model [5] maximises the probability of a sentence S has a parsing tree T ($P(T, S)$), by attaching probabilities to a top-down derivation of a tree T of a sentence S. By using a lexicalised probabilistic context free grammar (PCFG), with words w and parts-of-speech (POS) tags t attached to non-terminals $X(x =< w, t >)$ in the tree, each rule has the form, where H is the head-child of the phrase, which inherits the head-word h from its parent P. $L_1..L_n$ and $R_1..R_m$ are left and right modifiers of H:

$$P(h) \rightarrow L_n(l_n)...L_1(l_1)H(h)R_1(r_1)...R_m(r_m) \tag{1}$$

The generation of the right hand side (RHS) of a rule, giving the left hand side (LHS), is made by three steps, first generating the head, then making independence assumptions for the left and for the right modifiers which are generated by 0th-order Markov processes:

- Generate the head constituent label of the phrase, with probability $P_h(H|P,h)$
- Generate modifiers to the right of the head with the probability

$$\prod_{i=1...m+1} P_R(R_i(r_i)|P,h,H) \tag{2}$$

  where $R_{m+1}(r_{m+1})$ is defined as STOP - the STOP symbol is added to the vocabulary of non-terminals, and the model stops generating right modifiers when it is generated.

- Generate modifiers to the left of the head with probability

$$\prod_{i=1...n+1} P_L(L_i(l_i)|P,h,H) \tag{3}$$

  where $L_{n+1}(l_{l+1}) = STOP$.

This kind of decomposition helps avoiding the enormous number of potential rules and the sparse data problem generated by direct estimation of $P(RHS|LHS)$.

2.1 Unknown words

There are various levels of back-off for each type of parameter in the model. The modifiers are decomposed and smoothed separately [5]. For example, $P_L(L_i(lw_i lt_i)|P,H,w,t)$ is decomposed in the product $P_{L1}(L_i(lt_i)|P,H,w,t) \times P_{L2}(lw_i|L_i,lt_i,P,H,w,t)$, where lw_i and lt_i are the word and the POS tag generated with non-terminal L_i. In each case, the final estimate is:

$$e = \lambda_1 e_1 + (1-\lambda_1)(\lambda_2 e_2 + (1-\lambda_2)e_3) \tag{4}$$

where e_1 (for example, $P_H(H|P,w,t)$), e_2 (for example, $P_H(H|P,t)$) and e_3 (for example, $P_H(H|P)$) are maximum likelihood estimates with the context at different levels (three, in the exemplified case). All words occuring less than 5 times in training data, and words in test data that have never occured during the training, are replaced by the token *"uknown"*. This allows the model to handle the statistics for rare or new words in a robust way.

A best-first chart parser was built [4] and used to find the maximum probability tree for each sentence. The parser has been trained and tested on NILC treebank [1] (aproximately 100.000 and 30.000 sentences respectively).

3 Portuguese treebank

The portuguese corpus annotation proposed by Bick [3] is based on the Dependency Grammar paradigm, where morphology, syntax and semantics are treated. The flat dependency syntax is enriched by attaching direction markers and form and function tags to subclauses, allowing its transformation into constituent trees. With some Perl scripts, the group and the boundaries of the flat description are identified, the constituents are delimited, then constituent boundary brackets are marked in a complex form (np - noun phrase, pp - prepositional phrase, etc.), and finally every complex constituent is assigned to a function tag derived from the syntactic tag of its head. Table 1 shows an example of annotation system for the sentence "Erros abalam credibilidade da imprensa"(*Errors damage press credibility*).

Table 1: An example of Bick's annotation system

$SUBJ : n(MP)Erros$
$P : v - fin(PR \vert 3P \vert IND)abalam$
$ACC : np$
$= H : n(FS)credibilidade$
$= N <: pp$
$== H : prp(< sam- >)de$
$== P <: np$
$===> N : art(< -sam > FS)a$
$=== H : n(FS)imprensa$

where SUBJ = subject; n = noun; M = male; P = plural; $P :$ = predicate; v-fin = finite verb; PR = present tense; 3P = third person; IND = indicative; ACC = accusative (direct) object; np = noun phrase; H = head; F = female; S = singular; N = (pos/pre) adject; pp = prepositional phrase; prp = prepositional adverbs; $< sam- >$ = first part of morphologicaly fused word pair (de); art = article.

4 A relational database to store probabilities

The probabilities estimates have been stored in a traditional relational database. We begin by storing the word and its morphological and syntactic tags; then binding it with the headword's features in that context, as its parent's features and its position (left or right) in relation to the head.

The option for the database to store probabilities came from the idea that, in a initial step, we want a easier way to update and to add new informations.

5 Future work

We have introduced the initial steps for implementing a generative statistical parser in order to apply it to tagged sentences which follows the dependency portuguese grammar. We hope to achieve results as satisfatory as the ones for english (88.1/87.5% constituent precision/recall on Wall Street Journal corpus [6]), and to produce a very significant and representative treebank from a newspaper corpus.

Acknowledgements

This work is supported by CNPQ.

References

[1] http://www.nilc.icmc.sc.usp.br/tools.html.

[2] R. V. X. Aires; S. M. Alusio; D. C. S. Kuhn; M. L. B. Andreeta and O. N. Oliveira Jr. Study for brazilian portuguese. In *Proceedings of the 2000 Brazilian Artificial Intelligence Symposium*, 2000.

[3] E. Bick. *The Parsing System "Palavras" - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[4] A. G. Bonfante and M. G. V. Nunes. Um chart parser para o português do brasil. Technical Report (In portuguese), To be published.

[5] M. J. Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computatational Linguistics*, 1997.

[6] M. J. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.