Tagging a Morphologically Rich Language

The Construction of the Tycho Brahe Parsed Corpus of Historical Portuguese

Marcelo Finger

Department of Computer Science
Instituto de Matemática e Estatística
University of Sao Paulo – Brazil
mfinger@ime.usp.br

Abstract. Building large annotated corpora, such as is the case of the Tycho Brahe Corpus of Historical Portuguese, is only feasible if we use automatic methods for such tasks as part of speech tagging. The best automatic tools for part of speech tagging described in the literature were developed and tested for English.

However, the morphological richness of Portuguese forces us to use a number of tags several times larger than that used for English. An analysis of the complexity of the algorithm shows a prohibitive inefficiency resulting from the adoption of a much larger number of tags.

In this work, we propose a new, two-step approach for tagging texts of morphologically rich languages. We describe how the design of tags is affected by this method, and how the existing techniques must be adapted to deal with the greater number of tags found in morphologically rich languages.

1 Introduction

The Tycho Brahe Corpus of Historical Portuguese is being constructed as a part of a larger project, that aims at modelling the relationship between prosody and syntax in the process of language change which led from Classical Portuguese to Modern European Portuguese [1]. More specifically, we want to investigate the hypothesis that the syntactic change in the placement of clitics which occurred in Portuguese at the beginning of the 19th century was driven by a previous prosodic change in the 18th century.

For this end, two corpora are being built. First, we are building a Comparative Tagged Corpus of Spoken Modern European Portuguese and Brazilian Portuguese, consisting of categorized recorded registers from speakers of both dialects. It is our working hypothesis that, for the purposes of this research, the prosody of Classical Portuguese is identical to the prosody of Brazilian Portuguese.

The second corpus in construction, whose description is the aim of this paper, is the *Tycho Brahe Parsed Corpus of Historical Portuguese*. It consists of texts written by Portuguese authors born between 1550 and 1850 and it is named

after the astronomer who compiled the first systematic corpus of astronomical observations. This corpus will be available on the web.

As we wish to reach 1,000,000 words in the corpus, automated methods for morphological tagging and syntactic parsing have to be developed for Portuguese. Our approach is thus inspired by the Penn-Helsinki Corpus of Middle English [2]. However, due to the richness of the morphology of Portuguese, the number of tags to be used in a Portuguese corpus is about five times that of the English one, which brings us a severe limitation in the usability of existing tagging methods for the Portuguese corpus.

In the following we analyse the nature of such complexity, comparing the tagging of English and Portuguese. We then propose a method to overcome this limitation. Its basic idea is not to consider tags as "atomic" but to add some "structure" to them. We show how this enables the construction of a two-step tagging process that, in theory, is considerably more efficient for morphologically rich languages. We finalise by proposing new uses of taggers in the study of language evolution.

2 The Complexity of Tagging

One of the most successful part of speech taggers in the literature is that of Brill [4], which reportedly tags correctly 97% of the words in English. (One should not be over-impressed with such a figure, for the probability of a 100-word text being tagged correctly is smaller than 5%, as $0.97^{100} = 0.0475$.) Brill's results were actually measured against the *Penn Treebank Wall Street Journal Corpus* [5].

According to Brill's method, the tagger has first to be "trained" and then it can be applied to texts. The training part lasts orders of magnitude longer than the application part, so it deserves more attention. Brill's own complexity analysis of his algorithm, as presented in [3], showed that it has worst case complexity of

$$O(|n| \times |op| \times |c|)$$

where $|n|$ is the size of the training corpus, in number of words; $|op|$ is the number of possible swap operations like "change tag X by tag Y"; and $|c|$ is the number of conditions that may trigger the swapping, which is dependent on the swapping *window*, ie the number of words before and after the current one the algorithm scans to decide whether to change or not its tag; more specifically, $|c|$ is dependent on the number of words actually considered inside that window.

It is clear that, in terms of the number of tags t, the number of swap operations is $|op| = t^2$. The number of rules with a window of size 3 and inspection of at most 2 words inside the window was estimated as $3t^2 + 7t$, so the complexity[1] of the training algorithm is, in terms of tags:

$$O(t^2 \times (3t^2 + 7t) \times |n|)$$

[1] The detailing of this complexity analysis is due to Carlos Daniel Chacur Alves.

which is clearly a dependence on the number of tags of the fourth power.

The crucial element in the complexity of the number of rules is actually the maximal number of words inspected in a window; each iteration of the algorithm generates 10 rules, 3 of them inspecting 2 words in a window (responsible for the dominating term $3t^2$) and the remaining 7 frames inspecting only one word (giving the term $7t$). In general, for a set of rule frames in which at most m words are inspected in a window, the complexity of the algorithm would be $O(t^{2+m} \times |n|)$.

Due to the morphological richness of Portuguese, the Tycho Brahe corpus ended up with using 154 different tags, as opposed to the 36 different tags reportedly used by Brill in [4]. If we use the complexity calculus above as a basis for estimation, supposing we had Brill's algorithm being trained in Portuguese and English on the same computer with a corpus of the same size, the time complexity would compare:

$$\frac{\text{Time(Port)}}{\text{Time(English)}} = \frac{154^2 \times (3 \times 154^2 + 7 \times 154) \times |n|}{36^2 \times (3 \times 36^2 + 7 \times 36) \times |n|} = \frac{1712911816}{5365440} \approx 319.$$

That is, the Portuguese program runs 319 times slower than the English one. If, as reported in the documentation of the Penn Treebank Corpus [5], the training of his tagger on a manually tagged 500,000-word corpus took 1 day, the Portuguese tagger running in a similar machine would take 319 days, more than 10 months!

Since this is unacceptable, we devised an alternative approach, developing the Tycho Brahe Corpus through the following steps:

3 Taming the Complexity of Tagging Morphologically Rich Languages

The basic idea is to stop considering tags as basic atomic entities and to start considering some "internal structure" in the tag. What we do is to separate in a tag the *basic* component from its *complementary* part. Roughly speaking, the basic component can be thought as a category and the complement as a set of features. However, we tag with a distinct basic tag four special verbs, even though they are not grammatical categories on their own; similar creation of special basic tags happens with clitics, conjunctions, determiners, etc.

With the division between basic component/suffix of tags, we can divide the training part of the tagger in two phases: the *learning* of the basic components and the *refinement* of the suffixes.

3.1 The Design of Appropriate Tags

Designing "appropriate" part of speech tags has to take in consideration both the rich morphology of Portuguese and the desirable "economy" that we want to apply at representing such morphology .

As an example of this morphological richness, we see that for each verb in Portuguese, we have to tag 19 different inflected forms as opposed to 7 in English; we tag 15 types of determiners as opposed to 5 in English.

The economy side is the heart of our solution. We propose the use of tags with "internal structure" to separate the basic component from its inflections and features. The tag structure rule is the following

> *Each tag has just one basic component, and one or more complementary components.*

So, the tag of word *umas* is

D-UM-F-P

representing a determiner D as the basic tag, with the details showing indefiniteness (UM), feminine gender (F) and plural (P). Also within this spirit of economy, we have default tags. For example the masculine singular version of *umas* is *um* which receives the tag

D-UM

such that the gender, by default, is masculine and the number is singular. This default system is not so important for computational reasons as is the separation between basic and complementary tags.

Verbs are tagged with basic tag VB, but four special verbs are tagged separately, namely *ser*, *estar*, *haver* and *ter*. This decision was taken due to the frequency of occurrence of this verbs and the prominent distinct grammatical roles that these verbs can have. That was a very expensive decision for this four verbs and their inflected forms (not counting the addition of clitics) contribute with 50 tags (but just 4 basic tags, of course).

Due to the main motivation for the construction of the Tycho Brahe corpus, special care was taken with regards to the tagging of clitic pronouns and the way they attach to verbs. The following table illustrates the possibilities:

lhe/CL dei/VB-D	a separate basic tag before the verb
dei-lhe/VB-D+CL	attached after the verb
lhedei/CL+VB-D	attached before the verb
dar-lhe-ia/VB-R!CL	at the interior of the verb

Without counting separately the possibilities of clitic attachment, we ended up with 154 different. Luckily, we have only 36 distinct basic components, the same number as all English tags. This should enable us to reduce the training of the corpus to a complexity comparable to that of English.

4 The Two-step Learning Phase

The adoption of tags with internal structure allows us now to refine Brill's rule-based tagging method [4]. The learning of tag transformation rules will be done

in two steps, one that deals only with the basic component of tags, and other that deals with its complements.

The two phases of this method are the following:

(a) Use Brill's method to obtain a simplified tagger using the basic tags only, ignoring their internal structure. As a result of that phase, the program will have learned transformation rules that deal only with the 36 basic tags. Thus this step has exactly the same complexity of the English tagger.
(b) Refine the tags obtained in the initial step, taking into consideration features such as gender and number agreement, tense inflection, etc.

Step (b) uses explicit linguistic (morphological) knowledge, as opposed to step (a) which is basically a generate-and-test search process. In this respect, step (b) itself can be divided into two. The first one, step (b_1), is a morphological inspection of the words, together with other agreement verification; no learning is involved in step (b_1), as it uses built-in linguistic knowledge. Step (b_2) is another learning phase in which only restricted forms of rules are learned so that we allow only rules that *refine* a basic tag.

Ideally, step (b_1) could be developed to a point where step (b_2) becomes unnecessary, in the sense that the output of step (b_1) cannot be further refined. It can still contain errors resulting from step (a) that no refinement is capable of eliminating.

So, we also have to contemplate the possibility that the details produced by step (b) may lead to a revision of the basic tag obtained at the end of step (a); that is, it may be the case that we have enough information to revise and improve the output of step (a). In such a case, we will need a step (b_3). This final correction step will deal unrestricted transformation-rules. Because of that, step (b_3) is not currently being added to the learning process, for it falls back into the original problem of learning transformation-rules with 154 different tags. It is clear that error correction, if done through the learning of transformation-based rules, must incorporate some strong restriction on the types of rules it allows, otherwise it becomes unfeasible. Some further research is needed to investigate restricted ways of performing this error-correcting step (b_3).

After the tagger learned the transformation rules, it can be applied to any text.

5 Further Uses of Taggers in the Study of Language Evolution

One of the problems of studying old languages is that no speakers of these languages are available. At best we may find speakers of the modern versions of that language. This is what happens with Portuguese, where we have no living speaker of Classical 17th century Portuguese, but we can hypothesize that several features of it are present in, say, modern Brazilian Portuguese.

However, the use of taggers that "learn" are suggested here as a means of simulating a speaker with part of speech knowledge one would expect to find in

17th century Portuguese speaker. This could be obtained by training the tagger only with texts of that period; preferably, those texts should be from the same author, and written over a concentrated period.

This same process can be repeated with texts of only the 18th and only the 19th century, each one generating a tagger which has learned different transformation rules. A fourth tagger can be generated by training with all texts used by the other taggers.

A text of, say, the 19th century can be tagged by all those taggers and we can (automatically, of course) count the number of discrepancies obtained between them. Significant discrepancies may indicate a pronounced language change in that period, simulating a disagreement between speakers of the language at different times. On the other hand, very little discrepancies may indicate that no significant change on the part-of-speech level of the language has happened (even though there may have happened word meaning transformations, to which this method is totally opaque.)

This is an experiment we expect to be carrying once we have enough annotated texts to train several different taggers.

6 Conclusions

The main contribution of the current work is the refinement of the part of speech rule-based tagger for languages that require a large number of tags due to morphological richness. Its implementation is currently under way and should eventually become available at [1]. The development of a syntactical parser remains an area of active research, in Portuguese as much as in any other language.

References

[1] Rhythmic Patterns, Parameter Setting and Language Change. Please refer to the URL: `http://www.ime.usp.br/~tycho` .

[2] The Penn-Helsinki Corpus of Middle English. Please refer to the project's URL: `http://www.ling.upenn.edu/mideng` .

[3] Eric Brill. *A Corpus-Based Approch to Language Learning.* PhD thesis, University of Pennsylvania, 1993.

[4] Eric Brill. Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[5] B. Santorini. *Part of Speech Tagging Guidelines for the Penn Treebank Project*, 1990. 3rd revision, 2nd printing, updated in 1995 by R. MacIntyre; available at the Penn Treebank Project URL: `http://www.cis.upenn.edu/ treebank/home.html`.