Tagging Portuguese with a Spanish Tagger Using Cognates

Jirka Hana
Department of Linguistics
The Ohio State University
hana.1@osu.edu

Anna Feldman
Department of Linguistics
The Ohio State University
afeldman@ling.osu.edu

Chris Brew
Department of Linguistics
The Ohio State University
cbrew@acm.org

Luiz Amaral
Department of Spanish and Portuguese
The Ohio State University
amaral.1@osu.edu

Abstract

We describe a knowledge and resource light system for an automatic morphological analysis and tagging of Brazilian Portuguese.1 We avoid the use of labor intensive resources, particularly, large annotated corpora and lexicons. Instead, we use (i) an annotated corpus of Peninsular Spanish, a language related to Portuguese, (ii) an unannotated corpus of Portuguese, (iii) a description of Portuguese morphology on the level of a basic grammar book. We extend the similar work that we have done (Hana et al., 2004; Feldman et al., 2006) by proposing an alternative algorithm for cognate transfer that effectively projects the Spanish emission probabilities into Portuguese. Our experiments use minimal new human effort and show 21% error reduction over even emissions on a fine-grained tagset.

1 Introduction

Part of speech (POS) tagging is an important step in natural language processing. Corpora that have been POS-tagged are very useful both for linguistic research, e.g. finding instances or frequencies of particular constructions (Meurers, 2004) and for further computational processing, such as syntactic parsing, speech recognition, stemming, word sense disambiguation. Morphological tagging is the process of assigning POS, case, number, gender and other morphological information to each word in a corpus. Despite the importance of morphological tagging, there are many languages that lack annotated resources of this kind, mainly due to the lack of training corpora which are usually required for applying standard statistical taggers.

Applications of taggers include syntactic parsing, stemming, text-to-speech synthesis, word sense disambiguation, information extraction. For some of these getting all the tags right is irrelevant, e.g. the input to noun phrase chunking does not necessarily require high accuracy fine-grained tag resolution.

Cross-language information transfer is not new; however, most of the existing work relies on parallel corpora (e.g. Hwa et al., 2004; Yarowsky and Ngai, 2001) which are difficult to find, especially for lesser studied languages. In this paper, we describe a cross-language method that requires neither training data of the target language nor bilingual lexicons or parallel corpora. We report the results of the experiments done on Brazilian Portuguese and Peninsular Spanish, however, our system is not tied to these particular languages. The method is easily portable to other (related) languages. Our method assumes that an annotated corpus exists for the source language (here, Spanish) and that a text book with basic linguistic facts about the source language is available (here, Portuguese). We want to test the generality and specificity of the method. Can the systematic commonalities and differences between two genetically related languages be exploited for cross-language applications? Is the processing of Portuguese via Spanish different from the processing of Russian via Czech (Hana et al., 2004; Feldman et al., 2006)?

1We thank the anonymous reviewers for their constructive comments on an earlier version of the paper.

Table 1: Verb conjugation present indicative: -ar regular verb: cantar ‘to sing’

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. sg. canto</td>
<td>canto</td>
</tr>
<tr>
<td>2. sg. cantas</td>
<td>cantas</td>
</tr>
<tr>
<td>3. sg. canta</td>
<td>canta</td>
</tr>
<tr>
<td>1. pl. cantamos</td>
<td>cantamos</td>
</tr>
<tr>
<td>2. pl. cantais</td>
<td>cantais</td>
</tr>
<tr>
<td>3. pl. cantam</td>
<td>cantam</td>
</tr>
</tbody>
</table>

B: Ei dei para Maria.
1 gave to Mary.

B: ‘I gave it to Mary.’

b. A: ¿Qué hiciste con el libro? [PS]
What did you do with the book?

A: ‘What did you do with the book?’

B: Se lo di a Maria.
Herat il discrete to Mary.

B: ‘I gave it to Mary.’

Notice also that in the Spanish example (2b) the dative pronoun se ‘her’ is obligatory even when the prepositional phrase a María ‘to Mary’ is present.

3 Resources

3.1 Tagset

For both Spanish and Portuguese, we used positional tagsets developed on the basis of Spanish CLiC-TALP tagged corpus (Torruella, 2002). Every tag is a string of 13 symbols each corresponding to one morphological category. For example, the Portuguese word partires ‘you leave’ is assigned the tag VM0S---2PI- because it is a verb (V), main category (M), gender is not applicable to this verb form (0), number is singular (1), case is nominative (S), person is 2nd person (2), present (P), indicative (I) and participle-type is not applicable (O). A comparison of the two tagsets is in Table 2.

When possible the Spanish and Portuguese tagsets use the same values, however some differences are unavoidable. For instance, the pluperfect is a compound verb tense in Spanish, but a separate word in Portuguese. In addition, we added a tag for ‘treatment’ Portuguese pronouns.

The Spanish tagset has 282 tags, while that for Portuguese has 259 tags.

3.2 Training corpora

Spanish training corpus. The Spanish corpus we use for training the translation probabilities as well as for obtaining Spanish-Portuguese cognate pairs is a fragment (106,124 tokens, 18,629 types) of the Spanish section of CLiC-TALP (Torruella,

Notice that we have 6 possible values for the gender position: 0 (masculine), 1 (feminine), 2 (neuter, for certain pronouns), C (common, either 0 or 2), O (unspecified for this form within the category), the category does not distinguish gender.

Jirka Hana
Department of Linguistics
The Ohio State University
hana.1@osu.edu

Anna Feldman
Department of Linguistics
The Ohio State University
afeldman@ling.osu.edu

Chris Brew
Department of Linguistics
The Ohio State University
cbrew@acm.org

Luiz Amaral
Department of Spanish and Portuguese
The Ohio State University
amaral.1@osu.edu
automatically acquired. In the experiments below, we used the following modules – lookup in a list of (mainly) closely-class words, a paradigm-based guesser and an automatically acquired lexicon.

4.1 Portuguese closed class words

We created a list of the most common prepositions, conjunctions, and pronouns, and a number of the most common irregular verbs. The list contains about 460 items and it required about 6 hours of work. In general, the closed class words can be derived either from a reference grammar book, or can be elicited from a native speaker. This does not require native-speaker expertise or intensive linguistic training. The reason why the creation of such a list took 6 hours is that the words were annotated with detailed morphological tags used by our system.

4.2 Portuguese paradigms

We also created a list of morphological paradigms. For evaluation purposes, we selected and manually annotated a small portion (1,800 tokens) of NILC corpus, containing 1.2M tokens.

Table 2: Overview and comparison of the tagsets

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>No. of values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp</td>
<td></td>
<td>Po</td>
</tr>
<tr>
<td>1</td>
<td>POS</td>
<td>14 11</td>
</tr>
<tr>
<td>2</td>
<td>SubPOS – detailed POS</td>
<td>30 29</td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>6 6</td>
</tr>
<tr>
<td>4</td>
<td>Number</td>
<td>5 5</td>
</tr>
<tr>
<td>5</td>
<td>Case</td>
<td>6 6</td>
</tr>
<tr>
<td>6</td>
<td>Possessor’s Number</td>
<td>4 4</td>
</tr>
<tr>
<td>7</td>
<td>Form</td>
<td>3 3</td>
</tr>
<tr>
<td>8</td>
<td>Person</td>
<td>5 5</td>
</tr>
<tr>
<td>9</td>
<td>Temse</td>
<td>7 9</td>
</tr>
<tr>
<td>10</td>
<td>Mood</td>
<td>8 9</td>
</tr>
<tr>
<td>11</td>
<td>Participle</td>
<td>3 3</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of Morphological analysis

Consider for example, the form funciona ‘functions’ of the feminine noun função. The analyzer without a lexicon provides 11 analyses (6 lemmas, each with 1 to 3 tags), only one of them is correct. In contrast, the analyzer with an automatically acquired lexicon provides only two analyses: the correct one (noun fem. pl.) and an incorrect one (noun masc. pl. note that POS and number are still correct). Of course, not all cases are so persuasive.

The evaluation of the system is in Table 3. The 98.1% recall is equivalent to the upper bound for the task. It is calculated assuming an oracle-Portuguese tagger that is always able to select the correct POS tag if it is in the set of options given by the morphological analyzer. Notice also that for the tagging accuracy, the drop of recall is less important than the drop of ambiguity.

5 Tagging

We used the TnT tagger (Brants, 2000), an implementation of the Viterbi algorithm for second-order Markov model. In the traditional approach, we would train the tagger’s transitional and emission probabilities on a large annotated corpus of Portuguese. However, our resource-light approach means that modules with manually created resources are used before modules with resources automatically acquired. In the experiments below, we used the following modules – lookup in a list of (mainly) closely-class words, a paradigm-based guesser and an automatically acquired lexicon. This does not require native-speaker expertise or intensive linguistic training. The reason why the creation of such a list took 6 hours is that the words were annotated with detailed morphological tags used by our system.

4.3 Lexicon Acquisition

The morphological analyzer supports a module or modules employing a lexicon containing information about lemmas, stems and paradigms. There is always the possibility to provide this information manually. That, however, is very costly. Instead, we created such a lexicon automatically.

We automatically acquired lexic and similar systems are used as a backup for large high-precision high-cost manually created lexicons (e.g. Mikeljev, 1997; Havelco, 2001). Such systems extrapolate the information about the words known by the lexicon (e.g. distributional properties of endings) to unknown words. Since our approach is resource light, we do not have any such large lexicon to extrapolate from. The general idea of our system is very simple. The paradigm-based Guesser, provides all the possible analyses of a word consistent with Portuguese paradigms. Obviously, this approach massively overgenerates. Part of the ambiguity is usually real, but most of it is spurious. We use a large corpus to weed the spurious analyses out of the real ones. In such corpus, open-class lemmas are likely to occur in more than one form. Therefore, if a lemma+paradigm candidate suggested by the Guesser occurs in other forms in other parts of the corpus, it increases the likelihood that the candidate is real and vice versa. If we encounter the word cantamos ‘we sing’ in a Portuguese corpus, using the information about the paradigms we can analyze it in two ways, either as being a noun in the plural with the ending -s, or as being a verb in the 3rd person plural with the ending -amôs. Based on this single form we cannot say more. However, if we also encounter the forms canto, cantas, cantam the verb analysis becomes much more probable, and therefore, it will be chosen for the lexicon. If the only forms that we encounter in our Portuguese corpus were cantamos and (the non-existing) cantam (such as the existing word namo and namus) then we would analyze it as a noun and not as a verb.

With such an approach, and assuming that the corpus contains the forms of the verb matar ‘to kill’, matam², matas³, matam², matas³, etc., we would not discover that there is also a noun matar ‘Trees’ with a plural form matas = the set of the 2 noun forms is a proper subset of the verb forms. A simple solution is to consider not the number of forms covered in a corpus, but the coverage of the possible forms of the particular paradigm. However this brings other problems (e.g. it penalizes paradigms with large number of forms, paradigms with some obsolete forms, etc.). We combine both of these heuristics in Hana (2005).

Lexicon Acquisition consists of three steps:

1. A large raw corpus is analyzed with a lexicon-less MA (an MA using a list of mainly closely-class words and a paradigm based guesser);
2. All possible hypothetical lexical entries over these analyses are created.
3. Hypothetical entries are filtered with aim to discard as many nonexistent entries as possible, without discarding real entries.

Obviously, morphological analysis based on such a lexicon still overgenerates, but it overgenerates much less than if based on the endings alone.
the results of our morphological analyzer.

3. they are similar: we map the Spanish emissions onto the result of morphological analysis using automatically acquired cognates.

5.1 Tagging – Baseline

Our lower bound measurement consists of training the TnT tagger on the Spanish corpus and applying this model directly to Portuguese. The overall performance of such a tagger is 56.8% (see the min column in Table 4). That means that half of the information needed for tagging of Portuguese is already provided by the Spanish model. This tagger has seen no Portuguese whatsoever, and is still much better than nothing.

5.2 Tagging – Approximating Emissions I

The opposite extreme to the baseline, is to assume that Spanish emissions are useless for tagging Portuguese. Instead we use the morphological analyzer to limit the number of possibilities, treating them all equally. The emission probabilities would then form a uniform distribution of the tags given by the analyzer. The results are summarized in Table 4 (the even column) – accuracy 77.2% on full tags, or 47% relative error reduction against the baseline.

5.3 Tagging – Approximating Emissions II

Although it is true that forms and distributions of Portuguese and Spanish words are not the same, they are also not completely unrelated. As any Spanish speaker would agree, the knowledge of Spanish words is useful when trying to understand a text in Portuguese.

Many of the corresponding Portuguese and Spanish words are cognates, i.e. historically they descend from the same ancestor root or they are mere translations. We assume two things: (i) cognate pairs have usually similar morphological and distributional properties, (ii) cognate words are similar in form.

Obviously both of these assumptions are approximations:

1. Cognates could have departed in their meanings, and thus probably also have different distributions. For example, Spanish embarazada ‘pregnant’ vs. Portuguese embarazada ‘embarrassed’.

2. Cognates could have departed in their morphological properties. For example, Spanish cerca ‘near’ adverb vs. Portuguese cerca ‘fence’ noun (from Latin circa, circle ‘circle’).

3. There are false cognates – unrelated, but similar or even identical words. For example, Spanish salada ‘salty’ adj vs. Portuguese doce ‘salty’ noun, Spanish doce ‘twelve’ numeral vs. Portuguese doce ‘candy’ noun.

Nevertheless, we believe that these examples are true exceptions from the rule and that in majority of cases, the cognates would look and behave similarly. The borrowings, counter-borrowings and parallel developments of the various Romance languages have of course been extensively studied, and we have no space for a detailed discussion.

Identifying cognates. For the present work, however, we do not assume access to philological evidence, nor to accurate Spanish-Portuguese translations or even a sentence-aligned corpus. All of these are resources that we could not expect to obtain in a resource poor setting. In the absence of this knowledge, we automatically identify cognates, using the edit distance measure (normalized by word length).

Unlike in the standard edit distance, the cost of operations is dependent on the arguments. Similarly as Yarowsky and Wicentowski (2000), we assume that, in any language, vowels are more mutable in inflection than consonants, thus for example replacing a for i is cheaper that replacing r by r. In addition, costs are refined based on some well known and common phonetic-orthographic regularities, e.g. replacing a with e is less costly than replacing e with a.

However, we do not want to do a detailed contrastive morpho-phonological analysis, since we want our system to be portable to other languages. So, some facts from a simple grammar reference book should be enough.

Using cognates. Having a list of Spanish-Portuguese cognate pairs, we use these to map the emission probabilities acquired on Spanish corpus to Portuguese.

Let’s assume Spanish word \(w_s\) and Portuguese word \(w_p\) are cognates. Let \(T_s\) denote the tags that \(w_s\) occurs within the Spanish corpus, and let \(p_s(t)\) be the emission probability of a tag \(t\) if \(\exists T_s \rightarrow p_s(t) = 0\). Let \(T_p\) denote tags assigned to the Portuguese word \(w_p\) by our morphological analyzer, and the \(p_p(t)\) is the even emission probability: \(p_p(t) = \frac{1}{|T_p|}\). Then we can assign the new emission probability \(p_p'(t)\) to every tag \(t \in T_p\) in the following way (followed by normalization):

\[
    p_p'(t) = \frac{p_s(t) + \frac{p_p(t)}{2}}{1} 
\]

Results. This method provides the best results. The full-tag accuracy is 82.1%, compared to 56.9% for baseline (58% error rate reduction) and 77.2% for even-emissions (21% reduction). The accuracy for POS is 87.6%. Detailed results are in column e-cognates of Table 4.

6 Evaluation & Comparison

The best way to evaluate our results would be to compare it against the TnT tagger used the usual way – trained on Portuguese and applied on Portuguese. We do not have access to a large Portuguese corpus annotated with detailed tags. However, we believe that Spanish and Portuguese are similar enough (see Sect. 2) to justify our assumption that the TnT tagger would be equally successful (or unsuccessful) on them. The accuracy of TnT trained on 90K tokens of the CLC-TALPCorpus is 94.2% (tested on 16K tokens). The accuracy of our best tagger is 82.1%. Thus the error rate is more than 3 times bigger (17.9% vs. 5.4%).

Brando and Silva (2003) report 97.2% tagging accuracy on 23K testing corpus. This is clearly better than our results, on the other hand they need a large trained corpus of 207K tokens.

The details of the tagset used in the experiments are not provided, so precise comparison with our results is difficult.

7 Related work

Previous research in resource-light language learning has defined resource-light in different ways. Some have assumed only partially tagged training corpora (Merialdo, 1994); some have begun with small tagged seed word lists (Cucerzan and Yarowsky, 1999) for named-entity tagging, while others have exploited the automatic transfer of an already existing annotated resource in a different genre or a different language (e.g. cross-language projection of morphological and syntactic information in (Yarowsky et al., 2001; Yarowsky and Ngai, 2001), requiring no direct supervision in the target language).

Ngai and Yarowsky (2000) observe that the total weighted human and resource costs is the most meaningful measure of the degree of supervision. Cucerzan and Yarowsky (2002) observe that other useful measure of minimal supervision is the additional cost of obtaining a desired functionality from existing commonly available knowledge sources. They note that for a remarkably wide range of languages there exist a plenty of reference grammar books and dictionaries which is an invaluable linguistic resource.

7.1 Resource-light approaches to Romance languages

Cucerzan and Yarowsky (2002) present a method for bootstrapping a fine-grained, broad coverage POS tagger in a new language using only one person-day of data acquisition effort. Similarly to us, they use a basic library reference grammar book, and access to an existing monolingual text corpus in the language, but they also use a medium-sized bilingual dictionary.

In our work, we use a paradigm-based morpho-phonology, including only the basic paradigms from a standard grammar textbook. Cucerzan and Yarowsky (2002) create a dictionary of regular inflectional affix changes and their associated POS and on the basis of it, generate hypothesized inflected forms following the regular paradigm.

Carreras et al. (2003) present work on developing low-cost Named Entity Recognizer (NER) for Romance languages. Their system is based on a statistical approach using unsupervised learning methods. They demonstrate that the performance of their system is on par with existing NER systems for English.

In our work, we focus on developing a system for identifying named entities in Romance languages using a combination of rule-based and statistical methods.

7.2 Cognates

Bilingual lexicons within language families are often created using similar techniques. The assumption is that languages within a family share a common ancestral language, and that cognates are words that have a common origin.

For example, the Latin word "pater" (father) is cognate with the English word "father". Similarly, the Spanish word "madre" (mother) is cognate with the English word "mother".

To build a cognate lexicon, we use a combination of manual and automatic processes. Manual processes involve human annotators who identify cognates and classify them according to their semantic and syntactic properties.

Automatic processes involve the use of computational methods to identify cognates. These methods include pattern matching, word similarity measures, and neural network models.

7.3 Multilingual Corpora

Multilingual corpora are an important resource for the development of cross-lingual NER systems. These corpora contain text in multiple languages, and are used to train models that can recognize named entities in different languages.

To build a multilingual corpus, we use a combination of manual and automatic processes. Manual processes involve human annotators who transcribe text from one language to another. Automatic processes involve the use of computational methods to align text across languages and to identify named entities.

Acknowledgments

We would like to thank Maria das Gracas Mihayo, Natalia Cabezas, and Maria das Gracas Mihayo for their help with the annotation tasks. We also thank the anonymous reviewers for their valuable feedback.

References


We would like to thank Maria das Gracas Mihayo for giving us access to the NILC corpus annotated with PALAVRAS and to Carlos Pinto for the Spanish part of the CLiC-TALP corpus.