Statistical Parsing of Spanish and Data Driven Lemmatization
Joseph Le Roux, Benoît Sagot, Djamé Seddah

To cite this version:
Joseph Le Roux, Benoît Sagot, Djamé Seddah. Statistical Parsing of Spanish and Data Driven Lemmatization. ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012), Jul 2012, Jeju, South Korea. 6 p., 2012. <hal-00702496>

HAL Id: hal-00702496
https://hal.archives-ouvertes.fr/hal-00702496
Submitted on 30 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Statistical Parsing of Spanish and Data Driven Lemmatization

Joseph Le Roux†  Benoît Sagot⋆  Djamé Seddah⋆,⋄
† Laboratoire d’Informatique Paris Nord, Université Paris Nord, UMR CNRS 7030  
⋆Alpage, INRIA Université Paris Diderot  
⋄ISHA, Université Paris Sorbonne
leroux@univ-paris13.fr, benoit.sagot@inria.fr, djame.seddah@paris-sorbonne.fr

Abstract

Although parsing performances have greatly improved in the last years, grammar inference from treebanks for morphologically rich languages, especially from small treebanks, is still a challenging task. In this paper we investigate how state-of-the-art parsing performances can be achieved on Spanish, a language with a rich verbal morphology, with a non-lexicalized parser trained on a treebank containing only around 2,800 trees. We rely on accurate part-of-speech tagging and data-driven lemmatization in order to cope with lexical data sparseness. Providing state-of-the-art results on Spanish, our methodology is applicable to other languages.

1 Introduction

Grammar inference from treebanks has become the standard way to acquire rules and weights for parsing devices. Although tremendous progress has been achieved in this domain, exploiting small treebanks is still a challenging task, especially for languages with a rich morphology. The main difficulty is to make good generalizations from small example sets exhibiting data sparseness. This difficulty is even greater when the inference process relies on semi-supervised or unsupervised learning techniques which are known to require more training examples, as these examples do not explicitly contain all the information.

In this paper we want to explore how we can cope with this difficulty and get state-of-the-art syntactic analyses with a non-lexicalized parser that uses modern semisupervised inference techniques. We rely on accurate data-driven lemmatization and part-of-speech tagging to reduce data sparseness and ease the burden on the parser. We try to see how we can improve parsing structure predictions solely by modifying the terminals and/or the preterminals of the trees. We keep the rest of the tagset as is.

In order to validate our method, we perform experiments on the Cast3LB constituent treebank for Spanish (Castillan). This corpus is quite small, around 3,500 trees, and Spanish is known to have a rich verbal morphology, making the tag set quite complex and difficult to predict. Cowan and Collins (2005) and Chrupała (2008) already showed interesting results on this corpus that will provide us with a comparison for this work, especially on the lexical aspects as they used lexicalized frameworks while we choose PCFG-LAs.

This paper is structured as follows. In Section 2 we describe the Cast3LB corpus in details. In Section 3 we present our experimental setup and results which we discuss and compare in Section 4. Finally, Section 5 concludes the presentation.

2 Data Set

The Castillan 3LB treebank (Civit and Martí, 2004) contains 3,509 constituent trees with functional annotations. It is divided in training (2,806 trees), development (365 trees) and test (338 trees).

We applied the transformations of Chrupała (2008) to the corpus where CP and SBAR nodes are added to the subordinate and relative clauses but we did not perform any other transformations, like the coordination modification applied by Cowan and Collins (2005).

The Cast3LB tag set is rich. In particular part-of-speech (POS) tags are fine-grained and encode precise morphological information while non-terminal tags describe subcategorization and function labels.
Without taking functions into account, there are 43 non-terminal tags. The total tag set thus comprises 149 symbols which makes the labeling task challenging.

The rich morphology of Spanish can be observed in the treebank through word form variation. Table 1 shows some figures extracted from the corpus (training, development and test). In particular the word form/lemma ratio is 1.54, which is similar to other Romance language treebanks (French FTB and Italian ITB).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of tokens</td>
<td>94 907</td>
</tr>
<tr>
<td># of unique word forms</td>
<td>17 979</td>
</tr>
<tr>
<td># of unique lemmas</td>
<td>11 642</td>
</tr>
<tr>
<td>ratio word form/lemma</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Table 1: C3LB properties

Thus, we are confronted with a small treebank with a rich tagset and a high word diversity. All these conditions make the corpus a case in point for building a parsing architecture for morphologically-rich languages.

3 Experiments

We conducted experiments on the Cast3LB development set in order to test various treebank modifications, that can be divided in two categories:

1. modification of the preterminal symbols of the treebank by using simplified POS tagsets;
2. modification of the terminal symbols of the treebank by replacing word tokens by lemmas.

3.1 Experimental Setup

In this section we describe the parsing formalism and POS tagging settings used in our experiments.

3.1.1 PCFG-LAs

To test our hypothesis, we use the grammatical formalism of Probabilistic Context-Free Grammars with Latent Annotations (PCFG-LAs) (Matsuzaki et al., 2005; Petrov et al., 2006). These grammars depart from the standard PCFGs by automatically refining grammatical symbols during the training phase, using unsupervised techniques. They have been applied successfully to a wide range of languages, among which French (Candito and Seddah, 2010), German (Petrov and Klein, 2008), Chinese and Italian (Lavelli and Corazza, 2009).

For our experiments, we used the LORG PCFG-LA parser implementing the CKY algorithm. This software also implements the techniques from Attia et al. (2010) for handling out-of-vocabulary words, where interesting suffixes for part-of-speech words are collected on the training set, ranked according to their information gain with regards to the part-of-speech tagging task. Hence, all the experiments are presented in two settings. In the first one, called generic, unknown words are replaced with a dummy token UNK, while in the second one, dubbed IG, we use the collected suffixes and typographical information to type unknown words.\(^1\) We retained the 30 best suffixes of length 1, 2 and 3.

The grammar was trained using the algorithm of Petrov and Klein (2007) using 3 rounds of split/merge/smoother\(^2\). For lexical rules, we applied the strategy dubbed simple lexicon in the Berkeley parser. Rare words – words occurring less than 3 times in the training set – are replaced by a special token, which depends on the OOV handling method (generic or IG), before collecting counts.

3.1.2 POS tagging

We performed parsing experiments with three different settings regarding POS information provided as an input to the parser: (i) with no POS information, which constitutes our baseline; (ii) with gold POS information, which can be considered as a topline for a given parser setting; (iii) with POS information predicted using the MElt POS-tagger (Denis and Sagot, 2009), using three different tagsets that we describe below.

MElt is a state-of-the-art sequence labeller that is trained on both an annotated corpus and an external lexicon. The standard version of MElt relies on Maximum-Entropy Markov models (MEMMs). However, in this work, we have used a multiclass perceptron instead, as it allows for much faster training with very small performance drops (see Table 2). For training purposes, we used the training section of the Cast3LB (76,931 tokens) and the Leff lex-

\(^1\)Names generic and IG originally come from Attia et al. (2010).

\(^2\)We tried to perform 4 and 5 rounds but 3 rounds proved to be optimal on this corpus.
con (Molinero et al., 2009), which contains almost 800,000 distinct (form, category) pairs.  

We performed experiments using three different tagsets:

- a baseline tagset which is identical to the tagset used by Cowan and Collins (2005) and Chrupała (2008); with this tagset, the training corpus contains 106 distinct tags;
- the reduced2 tagset, which is a simplification of the baseline tagset: we only retain the first two characters of each tag from the baseline tagset; with this tagset, the training corpus contains 42 distinct tags;
- the reduced3 tagset, which is a variant of the reduced2 tagset: contrarily to the reduced2 tagset, the reduced3 tagset has retained the mood information for verb forms, as it proved relevant for improving parsing performances as shown by (Cowan and Collins, 2005); with this tagset, the training corpus contains 57 distinct tags.

Melt POS tagging accuracy on the Cast3LB development set for these three tagsets is given in Table 2, with overall figures together with figures computed solely on unknown words (words not attested in the training corpus, i.e., as high as 13.5 % of all tokens).

### 3.2 Baseline

The first set of experiments was conducted with the baseline POS tagset. Results are summarized in Table 3. This table presents parsing statistics on the Cast3LB development set in the 3 POS settings introduced above (i) no POS provided, (ii) gold POS provided and (iii) predicted POS provided. For each POS tagging setting it shows labeled precision, labeled recall, labeled F1-score, the percentage of exact match and the POS tagging accuracy. The latter needs not be the same as presented in Section 3.1.2 because (i) punctuation is ignored and (ii) if the parser cannot use the information provided by the tagger, it is discarded and the parser performs POS-tagging on its own.

As already mentioned above, this tagset contains 106 distinct tags. On the one hand it means that POS tags contain useful information. On the other hand it also means that the data is already sparse and adding more sparseness with the IG suffixes and typographical information is detrimental. This is a major difference between this POS tagset and the two following ones.

### 3.3 Using simplified tagsets

We now turn to the modified tagsets and measure their impact on the quality of the syntactic analyses. Results are summarized in Table 4 for the reduced2 tagset and in Table 5 for reduced3. In these two settings, we can make the following remarks.

- Parsing results are better with reduced3, which indicates that verbal mood is an important feature for correctly categorizing verbs at the syntactic level.

---

Note that MEl does not use information from the external lexicon as constraints, but as features. Therefore, the set of categories in the external lexicon need not be identical to the tagset. In this work, the Leffe categories we used include some morphological information (84 distinct categories).
- When POS tags are not provided, using suffixes and typographical information improves OOV word categorization and leads to a better tagging accuracy and F1 parsing score (78.94 vs. 81.81 for reduced2 and 79.69 vs. 82.44 for reduced3).

- When providing the parser with POS tags, whether gold or predicted, both settings show an interesting difference w.r.t. to unknown words handling. When using reduced2, the IG setting is better than the generic one, whereas the situation is reversed in reduced3. This indicates that reduced2 is too coarse to help finely categorizing unknown words and that the refinement brought by IG is beneficial, however the added sparseness. For reduced3 it is difficult to say whether it is the added richness of the POS tagset or the induced OOV sparseness that explains why IG is detrimental.

To make the parser less sensitive to lemmatization and tagging errors, we train both tools on a 20 jackknifed setup. Resulting lemmas and POS tags are then reinjected into the train set. The test corpora is itself processed with tools trained on the unmodified treebank. Results are presented Table 7. They show an overall small gain, compared to the previous experiments but provide a clear improvement on the richest tagset, which is the most difficult to parse given its size (106 tags).

First we remark that Tagging accuracy with the baseline tagset when no POS is provided is lower than previously observed. This can be easily explained: it is more difficult to predict POS with morphological information when morphological information is withdrawn from input.

Second, and as witnessed before, reduction of the POS tag sparseness using a simplified tagset and increase of the lexical sparseness by handling OOV words using typographical information have adverse effects. This can be observed in the generic Predicted POS section of Table 7 where the baseline

<table>
<thead>
<tr>
<th>Tagset</th>
<th>ALL</th>
<th>SEEN</th>
<th>UNK (13.84%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>98.39</td>
<td>99.01</td>
<td>94.55</td>
</tr>
<tr>
<td>reduced2</td>
<td>98.37</td>
<td>98.88</td>
<td>95.18</td>
</tr>
<tr>
<td>reduced3</td>
<td>98.24</td>
<td>98.88</td>
<td>94.23</td>
</tr>
</tbody>
</table>

Table 6: Lemmatization performance on the Cast3LB.
Table 7: Lemmatization Experiments

tagset is the best option. On the other hand, in IG Predicted POS, using the reduced3 is better than baseline and reduced2. Again this tagset is a trade-off between rich information and data sparseness.

In all cases reduced2 is below the other tagsets wrt. to Parseval F1 although tagging accuracy is better. We can conclude that it is too poor from an informational point of view.

4 Discussion

There is relatively few works actively pursued on statistical constituency parsing for Spanish. The initial work of Cowan and Collins (2005) consisted in a thorough study of the impact of various morphological features on a lexicalized parsing model (the Collins Model 1) and on the performance gain brought by the reranker of Collins and Koo (2005) used in conjunction with the feature set developed for English. Direct comparison is difficult as they used a different test set (approximately, the concatenation of our development and test sets). They report an F-score of 85.1 on sentences of length less than 40.5

However, we are directly comparable with Chrupała (2008)6 who adapted the Collins Model 2 to Spanish. As he was focusing on wide coverage LFG grammar induction, he enriched the non terminal annotation scheme with functional paths rather than trying to obtain the optimal tagset with respect to pure parsing performance. Nevertheless, using the same split and providing gold POS, our system provides better performance (around 2.3 points better, see Table 8).

It is of course not surprising for a PCFG-LA model to outperform a Collins’ model based lexicalized parser. However, it is a fact that on such small treebank configurations PCFG-LA are crucially lacking annotated data. It is only by greatly reducing the POS tagset and using either a state-of-the-art tagger or a lemmatizer (or both), that we can boost our system performance.

The sensitivity of PCFG-LA models to lexical data sparseness was also shown on French by Seddah et al. (2009). In fact they showed that performance of state-of-the-art lexicalized parsers (Charniak, Collins models, etc.) were crossing that of Berkeley parsers when the training set contains around 2500–3000 sentences. Here, with around 2,800 sentences of training data, we are probably in a setting where both parser types exhibit similar performances, as we suspect French and Spanish to behave in the same way. It is therefore encouraging to notice that our approach, which relies on accurate POS tagging and lemmatization, provides state-of-the-art performance.

5 Conclusion

In this paper we presented several experiments reporting the impact of lexical sparseness reduction on non lexicalized statistical parsing. We showed that, by using state-of-the-art lemmatization and POS tagging on a reduced tagset, parsing performance can be on a par with lexicalized models that manage to extract more information from a small corpus exhibiting a rich lexical diversity. It remains to be seen whether applying the same kind of simplifications to the rest of the tagset, i.e. on the internal nodes, can further improve parse structure quality.

Finally, the methods we presented in this paper are not language specific and can be applied to other languages if similar resources exist.
Table 8: PARS EVAL F-score results on the Cast3LB test set

<table>
<thead>
<tr>
<th>TAGSET</th>
<th>MODE</th>
<th>TOKENS</th>
<th>ALL</th>
<th>≤ 70</th>
<th>≤ 40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reduced3</td>
<td>Gen.</td>
<td>pred. POS</td>
<td>83.92</td>
<td>84.27</td>
<td>85.08</td>
</tr>
<tr>
<td></td>
<td>eval. w/o CP/SBAR</td>
<td></td>
<td>84.02</td>
<td>84.37</td>
<td>84.24</td>
</tr>
<tr>
<td>baseline</td>
<td>IG</td>
<td>pred. lemma &amp; POS</td>
<td>84.15</td>
<td>84.40</td>
<td>85.26</td>
</tr>
<tr>
<td></td>
<td>eval. w/o CP/SBAR</td>
<td></td>
<td>84.34</td>
<td>84.60</td>
<td>85.45</td>
</tr>
<tr>
<td>reduced3</td>
<td>Gen.</td>
<td>gold POS</td>
<td>86.21</td>
<td>86.63</td>
<td>87.84</td>
</tr>
<tr>
<td></td>
<td>eval. w/o CP/SBAR</td>
<td></td>
<td>86.35</td>
<td>86.77</td>
<td>88.01</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>gold POS</td>
<td>83.96</td>
<td>84.58</td>
<td>–</td>
</tr>
</tbody>
</table>

(Chrupała, 2008)

Acknowledgments

Thanks to Grzegorz Chrupała and Brooke Cowan for answering our questions and making data available to us. This work is partly funded by the French Research Agency (EDyLex, ANR-09-COORD-008).

References


