The challenge of syntactic dependency parsing adaptation for the patent domain

Alicia Burga¹, Joan Codina¹, Gabriela Ferraro³, Horacio Saggion¹, Leo Wanner¹,²
¹ Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra
² Institució Catalana de Recerca i Estudis Avançats (ICREA)
³ NICTA & Australian National University
{first-name.last-name}@upf.edu
gabriela.ferraro@nicta.com.au

Abstract

Patents are legal documents with a proper complex discourse making it difficult to use off-the-shelf syntactic parsers to adequately process them. The annotation of a training corpus is a titanic task that cannot be afforded for a single project. In this paper, we present a methodology for adapting a dependency parser to the patent genre which only requires the addition of minimal genre specific annotated sentences and minor domain-adaptations to the treebank. After identifying the principal problems faced by the parser, we added to the training corpus sentences that condense and maximize the information brought to the model. The resulting models allow parsing of patents with performances similar to newspaper data.

1 Introduction

Patent language is difficult to be read and understood due to the complex structures it contains. Thus, there is a need to develop tools that help humans to understand patent contents and, at the same time, to manage large amounts of patent data. The first step for “understanding” the patent content beyond “bag of words” is to analyse its content by means of NLP tools. These tools, particularly syntactic parsers, are usually trained with newspapers corpora that drastically differ from the patent genre. In this paper, using Bohnet’s transition-based parser (Bohnet and Kuhn, 2012), we show that after increasing the training corpus by 9% with domain specific data, and modifying the treebank annotation, the parser achieves on unseen patent data similar results to those obtained when analysing newspaper text.

In Section 2 we explain why parsing patents is a challenging task. In Section 3, the performance of one off-the-shelf parser is tested in newspaper text and in patents, and the error sources are explained. In Section 4, our approach to patent domain adaptation is described. We present in Section 5 the experiments and evaluation of our approach. In the last two sections, the related work and the final conclusions are presented.

2 Patent language phenomena

Patents have a pre-defined document structure that consists on two main sections: Claims and Description. The latter does not have so peculiar and unique linguistic specifications as the former, but it still carries some linguistic characteristics. Thus, we need to identify and analyze those particular configurations in patents in order to appropriately deal with them. Those phenomena are the following:

- Long sentences: Patents, especially their claim section, contain very long sentences: sentences with more than 100 tokens are common, in contrast to newspaper text, whereas sentences contain around 30 tokens. Thus, when working in patent domain, we deal with long-distance dependencies.

- Complex syntactic structure: Long sentences usually have a complex syntactic structure. Thus, sentences with several subordination clauses -and levels- are often encountered in patents. Also, as the claim below shows, we find coordinative constructions that are difficult to interpret because their coordinated heads (underlined) are separated by a long distance in the text (whereas at the same time other complex internal syntactic configurations hold).
A device for coupling a sail batten to a mast in a board sail, the device comprising: a body having a first end for rotateably bearing against a mast and a second end for receiving a batten tip, a sail-widget for attachment to the sail, the sail-widget movably locatable with the batten and body, and a connector for movably coupling the body, the batten tip and the sail-widget together . . .

**Peculiar multi-word expressions or terminology:** In patents, we find a big amount of multi-word expressions or terminology, due to their nature of describing new inventions and/or methods. Many of those multi-word expressions are longer and more complex than noun phrases found in newspaper text. Thus, noun phrases like *percussive like portable power-operated drill*, *long bar member matching apparatus* and *magneto optical recording and reproducing unit*, are very common in patent language. The complexity of the syntactic structure is also reflected in the internal configuration of these multi-word expressions.

**Different PoS distribution of characteristic tokens:** There are tokens that in newspaper text carry a PoS different from that one carried in patents. The most salient cases are *said* and *means*. *Said*, which in newspaper data is always the past participle of the verb *say* (PoS=VBD), in patents it is always a determiner (PoS=DT), which signals that the introduced noun has been already presented. *Means*, which in newspaper data is ambiguous between a noun (PoS=NN) and a verb (PoS=VBZ), in patents it never works like a verb.

### 3 Assessment of the performance of Bohnet’s parser on patent material

As already mentioned, patents sentences can be very long, raising very long-distance dependencies. Therefore, transition based parsers, which typically have a linear or quadratic complexity (Nivre and Nilsson, 2004; Attardi, 2006) are better suited to parse patent sentences than graph based parsers, which usually have a cubic complexity. For our experiments on domain adaptation, we use the transitional version of Bohnet’s (Bohnet and Kuhn, 2012) parser, one of the best performing parsers in the CoNLL Shared Task 2009 (Hajič et al., 2009).

In order to test the performance of the chosen parser with patents, we evaluated it using in-domain and out-of-domain data as test sets. As in-domain data test set we used the CoNLL 2009 testing set. For the out-of-domain data test set, we compiled a corpus composed by 8 sentences (684 tokens) extracted from patent documents. The sentences were automatically analysed with Bohnet’s parser and, then, the part-of-speech tags and dependency structures were manually corrected in order to get a gold standard set (from now on, we refer to this test set as *patent test set*).

As the quality of the dependency parser heavily depends on correct part-of-speech tags, the morphological tagger provided with the Bohnet’s parser was also evaluated. For its evaluation, the morphological tagger was trained with the CoNLL 2009 training data and the results were evaluated with the CoNLL test set and with the patent test set.

As shown on Table 1, the tagging accuracy drops almost three points when dealing with patent sentences (*#toks* stands for number of tokens and *% of corr* stands for the percentage of correct tokens).

### Table 1: Evaluation of PoS tagging with in and out domain data.

<table>
<thead>
<tr>
<th>PoS Tagging Experiment</th>
<th>#toks</th>
<th>% of corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL train CoNLL test</td>
<td>57676</td>
<td>97.69%</td>
</tr>
<tr>
<td>CoNLL train, patent test</td>
<td>684</td>
<td>94.59%</td>
</tr>
</tbody>
</table>

For the evaluation of the syntactic parser, we followed the same methodology as above. First, the parser was trained with the CoNLL 2009 training data and the results were evaluated with the CoNLL test set and with the patent test set. The evaluation results of the syntactic parser are shown in Table 2. The LAS measure (Labelled Attachment Score) stands for the proportion of edges with correct governor and dependent and the right label on the edge, and the UAS (Unlabelled Attachment score) stands for the proportion of edges with correct governor and dependent (Eisner, 1997; Collins et al., 1999).\(^1\)

There is a significant drop in LAS and UAS when the parser is tested on the patent test set, dropping more than eight points for LAS and near six points for the UAS. The results suggest that it is necessary to adapt the parser to the patent domain. Our goal is

\(^1\)LAS and UAS scores are calculated taking into account punctuation.
Table 2: Evaluation of the Bohnet’s syntactic dependency parser with in and out domain data.

to achieve, with a minimum effort, similar results as with in-domain-data.

3.1 Detection of the most problematic phenomena

In patents, we can identify two very different writing styles: one similar to newspaper data, found in the description; and one unique and typical of the patent domain, found in the claims section.² The claims’ writing style is very defined and regular along any English patent corpus. It has its own peculiarities and characteristics, such as the lack of a principal verb (being a noun the syntactic root in a whole claim), the different PoS that some words carry (with respect to newspaper data), etc. Thus, even though our patent test set is small, the analysis of the parser results over it allows detecting the most problematic phenomena:

Problems derived from PoS tagger: Given that the parser partially bases its decisions about the dependency relation involved, as well as about its respective governor and dependent, on the PoS of the syntactic tree nodes, an error in the PoS tagging brings a high probability of erroneous relations and/or members connected. Given the wrong PoS assignation of said and means explained in Section 2, dependency problems rise when these words are involved. Thus, as said is not always tagged as a determiner but as a verb, it is wrongly considered by the parser as the governor of some verbal relation (and, many times, it is considered as the root of the sentence). As expected, parallel problems arise with a wrong tagging of means.

Another related problem concerns nouns modified by a posterior participle. As the PoS tagger erroneously assigns the past tense tag to a past participle, instead of establishing a relation between the noun (as head) and the participle as a modifier, the parser assigns most of the times -overall when no verb is found as root- a relation between the participle (as head, given its verbal PoS) and the previous noun (as subject).

Confusion between adverbial phrases and relative clauses: A frequent problem that arises from a PoS tagging confusion between verbs and nouns is that the parser analyzes adverbial phrases—which imply a relation ADV/LOC/TMP headed by a verb and linked to a subordinating element that at the same time heads the relation SUB linked to the subordinated verb– as relative clauses—which imply a relation NMOD headed by a noun and the subordinated verb– and vice versa (see Figure 1).

Despite these two very different styles within patents, a classifier to distinguish between the two sections is not needed given that claims and descriptions are usually separated in patent data bases.

2Despite these two very different styles within patents, a classifier to distinguish between the two sections is not needed given that claims and descriptions are usually separated in patent data bases.
4 The Approach

On the basis of the error detection phase, we considered more appropriate for our needs to perform the domain adaptation as a pre-processing task. Our domain adaptation approach consists of modifying the training data and adding sentences that contain the manual correction of the problems described above (no post-process is applied).

Correction of parsing errors in the CoNLL corpus: we decided to reannotate, keeping in mind our purposes, 200 sentences from the CoNLL corpus that contain at least one of the principal parsing problems (nouns followed by a number, adverbial phrases interpreted as relative clauses or vice versa, chains of modifying nouns and nouns modified by a posterior participle).

Corpus enrichment with in-domain data:

(a) Correction of part-of-speech tags: as explained in previous sections, some important errors of parsing come from a wrong PoS tagging. Thus, in order to correct those problems, it is crucial to assign the appropriate PoS to each element. So, we have corrected the PoS of problematic lexemes in 100 sentences from claims.

(b) Correction of 26 syntactic structures extracted from the claim and description sections of patents. One of the principal goals of this step concerned the appropriate analysis of multiword expressions (i.e. detection of the right head, identification of all its dependents as well as their corresponding dependencies).

(c) Insertion of structures manually created of the type A comprises B; C; D: Given that many claims in patents carry such configuration, it is important for a correct parsing in patents that the parser learns that the semicolon can be interpreted as a coordinative conjunction or, in other words, that the elements B, C and D are members of a coordinated object. As real claims usually carry an internal structure much more complex than the simplified configuration we want the parser to learn, instead of taking the risk that such complex internal structure becomes noise for a correct parser learning, we decided to manually simplify claim sentences and thus get to five simple sentences containing such basic configuration. These sentences were added to the training set.

We also took advantage of the Brown’s clusters (Brown et al., 1992) based on the English Gigaword corpus, using them as features for the tagger training.

5 Experiments and evaluation

With the data and resources for domain adaptation described in Section 4 it is possible to re-train Bohnet’s tagger and syntactic parser. In what follows, we present the training experiments and their evaluation.

5.1 Re-training of Bohnet’s parser

For the training of the syntactic parser, the CoNLL corpus and the corpus data described in Section 4 were merged into a single training set. After experimenting with different versions of the corpus, the best results were achieved when the data from items (b) and (c) (Section 4) was multiplied by three (see Table 3).

<table>
<thead>
<tr>
<th>Adaptation source and process</th>
<th># of structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL corpus, correct dependencies</td>
<td>200</td>
</tr>
<tr>
<td>Claims, correct PoS</td>
<td>100</td>
</tr>
<tr>
<td>Patents, correct dependencies</td>
<td>78 (26*3)</td>
</tr>
<tr>
<td>New, simulation of claim style</td>
<td>15 (5*3)</td>
</tr>
</tbody>
</table>

Table 3: Description of the expansion of the training corpus.

After re-train, the parser was evaluated with the CoNLL test set and with the patent test set. The results are shown in Table 4 (for the sake of comparison, the table also presents the results of Table 2, in grey color.)

<table>
<thead>
<tr>
<th>Syntactic parser training</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL train, CoNLL test</td>
<td>88.21</td>
<td>90.75</td>
</tr>
<tr>
<td>CoNLL+patent train, CoNLL test</td>
<td>88.47</td>
<td>91.07</td>
</tr>
<tr>
<td>CoNLL train, patent test</td>
<td>79.61</td>
<td>84.87</td>
</tr>
<tr>
<td>CoNLL+patent train, patent test</td>
<td>82.60</td>
<td>87.42</td>
</tr>
</tbody>
</table>

Table 4: Training of Bohnet’s syntactic parser with in and out domain data

---

3The adaptation tasks were done using the graph transducer MATE (Bohnet et al., 2000).
The evaluation with the CoNLL test is almost the same as before re-training, which means that the accuracy of the model is not degraded once the domain data is included. Regarding the evaluation with the patent test set, the re-trained parser gets more than two points than before re-training. The improvements in LAS and UAS are statistical significant with \( p < 0.01 \) (paired t-test).

5.2 Impact of the PoS tagging during syntactic parsing

For the re-training of the tagger, we used the CoNLL corpus and the data from item (a) (Section 4). In order to measure the impact of the PoS tagging process, we evaluate the performance of parser model from Section 5.1 with test sets that contained PoS from different sources. As shown in Table 5, the parser performed best when processing the test set that contained perfect PoS (TestSet-PerfectPoS). When the PoS tagging analysis is done by Bohnet’s tagger (TestSet-BohnetsPoS), which is trained with CoNLL data, the LAS score dropped almost two points and the UAS almost three points. Finally, we evaluate the parser using a test set that contained PoS tagging information derived from our re-trained tagger (TestSet-PatentsPoS). As expected, we are able to achieve reasonable good tagging results when using a tagging model trained with patents data.

When the PoS tagging errors are known, as in this context, adapting the part-of-speech tags is not costly and its effect on parsing is considerable.

<table>
<thead>
<tr>
<th>PoS tagging test sets</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestSet-PerfectPoS</td>
<td>82.25</td>
<td>87.43</td>
</tr>
<tr>
<td>TestSet-BohnetsPoS</td>
<td>79.53</td>
<td>84.65</td>
</tr>
<tr>
<td>TestSet-PatentsPoS</td>
<td>81.29</td>
<td><strong>86.26</strong></td>
</tr>
</tbody>
</table>

Table 5: The impact of PoS tagging in patents dependency parsing.

6 Related work

Parser domain adaptation is a costly task. That is why many works investigate which are the most appropriate improvement actions that enable a reasonable parsing quality with out-of-domain data. Improvement actions can be applied as pre-processing, as a post-processing, or as both. For example, (Seeker et al., 2010) presents a hybrid approach, which consists on PoS tagging improvement (obtaining an error reduction of 45%), tree-bank modification (specifically, restructuring of PPs); deletion of morphological information not directly relevant to syntax (e.g.: mood and tense); combination of different parsing strategies; and a post-processing re-labelling based on labels distribution (e.g.: a verb can have just one subject). After these modifications, the labelled accuracy of the parser increases from 87.64% to 89.40%. (Attardi, 2006) uses Active Learning (AL) as a strategy for domain adaptation. For doing so, the following steps are applied: i) target domain sentences are parsed; ii) a small number of sentences with the highest perplexity are manually corrected and added to the training corpus, iii) the process is repeated three times. This strategy improves the LAS results from 74.82% to 79.48% on the target domain. The parser domain adaptation proposed by (Dredze et al., 2007) includes modification of the feature set, training with many parsers, and target focus learning (learning technique that favours training examples that are similar to the target). Their goal was to adapt a parser trained on a single domain to a new target domain using only unlabelled data. They report negative results as they were not able to be substantially above the state of the art baseline. Domain adaptation can also be focused on the lexicon as proposed by (Candito et al., 2011). For improving parsing on biomedical French texts, the authors used word clusters to reduce lexical data sparseness, achieving an error reduction of 21%.

Kwon (2009) proposed a method to improve patent parsing in the context of an English-Korean machine translation system. The improvement actions include: adapting the PoS tagger to the patent domain (complex symbols words, chemical and mathematical formulas, programming code, etc.) and simplification of long sentences through the identification of coordinated structures in order to reduce parsing complexity. Similarly, in order to generate patent claim paraphrases and summaries,
(Bouayad-Agha et al., 2009) present a rule-based simplification system that reduces the complexity of patent claim sentences before syntactic parsing.

We observe that one of the most popular and effective adaptation task is the improvement of the PoS tagging annotation and algorithms. Hence, we also follow this trend. Differently from many state-of-the-art approaches, we have also enriched the training corpus with domain data. We show that, despite manual annotation is costly, a considerable improvement is obtained with a small number of domain structures.

7 Conclusions

We presented an approach for adapting a state-of-the-art syntactic dependency parser to the patent domain. Even though the results we obtained are competitive, since we are able to augment the labelled and unlabelled accuracy of the parser, it still remains the principal parsing challenge: facing the style duality of patents and the peculiar claims’ writing style, avoiding that the special treatment assigned by the parser to the claims is spread to other sections of the patent. A possible solution would be to train two parsers, one for the claim section and another for the rest.

We found several trends for future work. First we plan to augment the patent test set in order to get a more realistic evaluation. Another adaptation task regarding tree-bank modification is to remove from the CoNLL corpus, structures which syntactic analysis is correct but interferes with the patent writing style. Also, we want to try the new parser from (Bohnet and Nivre, 2012), which is a parsing system with a joint tagger.

We would like to thank B. Bohnet and M. Ballesteros for their very helpful comments. At the time of the reported research, the third author was member of the NLP group, Universitat Pompeu Fabra.

References


