Extraction of LTAG-based supertags from the French Treebank
Bladier, Tatiana; van Cranenburgh, Andreas; Kallmeyer, Laura

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Introduction. In the present paper we present an approach to automatically extract Lexicalized Tree Adjoining Grammars [LTAG; 6] from the French Treebank [FTB; 1], which can be used for the LTAG supertagging and parsing. We discuss the challenges encountered while extracting the grammars, propose our solutions, and evaluate our French LTAG grammars on the supertagging task.

LTAG is a linguistically and psychologically motivated grammar formalism. Productions in LTAGs support an extended domain of locality, which allows them to express linguistic generalizations that are not captured by typical statistical parsers based on context-free grammars or dependency parsing. Parsing with LTAGs can be facilitated through the intermediate step of supertagging—a task of assigning a sequence of LTAG tree templates (supertags) for a given sentence. Supertagging is referred to as being “almost parsing”, since it takes much of syntactic disambiguation before applying a costly parsing algorithm [3]. Several supertagging approaches have been proposed for LTAGs [7, 2], with the most recent advances including approaches based on Recurrent Neural Networks (RNN) [9]. While supertagging experiments were reported for English [3, 9] and German [12], to our best knowledge, no research was reported on supertagging with French LTAGs.

Extraction of a feasible LTAG supertag lexicon. Neural supertagging with LTAGs is a sequence labeling problem. Thus, the performance of the supertagger on the classification task strongly depends on the size of the supertag lexicon—in case of a too large number of supertags, the supertagging system has to make a choice between a bigger number of options and thus suffers a drop of performance. Recent research on supertagging with automatically induced LTAG grammars for English and German [9, 12] shows that the manageable size of a supertag lexicon contains around 4000 distinct supertags with roughly a half of the supertags appearing only once (see Table 1).

LTAG Induction. We used the French Treebank for the grammar extraction due to its being the currently largest available and the most widely used resource for the French language. In order to extract an LTAG from the FTB, we applied the heuristic top-down procedure described by Xia [13]. For facilitation of the LTAG induction we carried out pre-processing steps described in Candito et al. [4] and Crabbé and Candito [5] including extension of the original POS tag set in FTB from 13 to 26 POS tags, undoing most multiword expressions with regular syntactic patterns and raising some complements (e.g. raising the PPs of the VPinf constituents). We experimented with the following LTAGs for French: including 13 or 26 POS tags, with or without compounds, including and excluding punctuation marks (see Table 2).

Challenges while extracting supertags from FTB. A big number of flat multi-word expressions (MWEs) in FTB leads to a large number of rather infrequent distinct extracted supertags. About 14 % of the word tokens in FTB belong to flat MWEs. After rewriting compounds with regular syntactic patterns, the number of MWEs is reduced to approximately 5 %. Extending the set of part-of-speech tags provided by FTB to more fine-grained 26 POS-tags theoretically helps the supertagger to better learn the dependencies between the supertags, however, a bigger number of POS tags leads to a higher number of supertags, which causes a drop of performance.

Left- and right-sister-adjunction. The tree structures in FTB are rather flat and allow any ordering of arguments and modifiers. In order to preserve these original flat structures as far as possible we decided against the traditional notion of adjunction in TAG which relies on nested structures and apply sister-adjunction; i.e., the root of a sister-adjoining tree can be attached as a daughter of any node of another tree with the same node label. Since a modifier can appear on the right or on the left side relative to the position of the constituent head, we distinguish between right- and left-sister-adjoining trees—marked with * on the left or the right side of the root label as shown in Figure 1.

Evaluation and future work. We evaluated our extracted grammars on experiments with an implemented supertagger similar to the one described in Kasai et al. [9] and Samih [10]. Our results show that the
extracted grammars get comparable results with the data for English and German LTAG. In our future work we plan to improve the supertagger and to use it for graph-based parsing. In particular, we aim at adapting the A*-based PARTAGE parser for LTAGs developed by Waszczuk [11] for parsing with extracted supertags. We also intend to add deep syntactic features and information on semantic roles to the supertags in order to test whether extracted LTAGs can be used for semantic role labeling.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Supertags</td>
<td>5145</td>
<td>2516</td>
<td>3426</td>
<td>4727</td>
</tr>
<tr>
<td>Supertags occur. once</td>
<td>2693</td>
<td>1123</td>
<td>1562</td>
<td>2165</td>
</tr>
<tr>
<td>POS tags</td>
<td>13</td>
<td>53</td>
<td>53</td>
<td>36</td>
</tr>
<tr>
<td>Sentences</td>
<td>21550</td>
<td>28879</td>
<td>50000</td>
<td>44168</td>
</tr>
<tr>
<td>Avg. sentence length</td>
<td>31.34</td>
<td>17.51</td>
<td>17.71</td>
<td>appr. 20</td>
</tr>
<tr>
<td>Supertagging accuracy</td>
<td><strong>78.54</strong></td>
<td>85.91</td>
<td><strong>88.51</strong></td>
<td><strong>89.32</strong></td>
</tr>
</tbody>
</table>

Table 1: Comparison of LTAGs extracted from different treebanks

<table>
<thead>
<tr>
<th>Extracted French LTAG</th>
<th># supertags</th>
<th># supertags once</th>
<th>Supertagging accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 POS, undone comp.</td>
<td>5145</td>
<td>2693</td>
<td><strong>78.54</strong></td>
</tr>
<tr>
<td>13 POS, with compounds</td>
<td>6847</td>
<td>3738</td>
<td>76.78</td>
</tr>
<tr>
<td>26 POS, with compounds</td>
<td>5831</td>
<td>3008</td>
<td>74.84</td>
</tr>
<tr>
<td>13 POS, undone comp., no punct. marks</td>
<td>5015</td>
<td>2557</td>
<td>74.44</td>
</tr>
</tbody>
</table>

Table 2: Supertagging experiments with different LTAGs extracted from the FTB.

Figure 1: Left- and right-sister-adjunction for L’activité ne suffit pas ("The activity does not suffice")

References


