Adding frequencies to the LGLex lexicon with IRASUBCAT

Elsa Tolone and Romina Altamirano

FaMAF, Universidad Nacional de Córdoba,
elsa.tolone@univ-paris-est.fr, romina.altamirano@gmail.com

Abstract. We present a method for enlarge a lexicon (with frequencies information), that is useful for parsing and others NLP applications. We show an example enlarging the verbal LGLex lexicon of French [8], using several corpora extracted from the evaluation campaign for French parsers Passage [5]. To do that, we use the results of the FRMG parser [7] with IRASUBCAT [1], a tool that automatically acquires subcategorization frames from corpus in any language and that also allows to complete an existing lexicon. We obtain the frequencies of occurrence for each input and each subcategorization frame for 14,068 distinct lemmas.

Keywords: Lexicon-Grammar, syntactic lexicon, French lexicon, subcategorization, frequency of occurence.

1 Lexicon-Grammar, LGLex lexicon and FRMG parser

Lexicon-Grammar tables are currently one of the major sources of syntactic lexical information for the French language [4]. Moreover, several Lexicon-Grammar tables exist for other languages, such as Italian, Brazilian Portuguese, Modern Greek, Korean, Romanian, and others.

We improved the Lexicon-Grammar tables to make them usable in various NLP applications, in particular parsing [8]. So we generated a French syntactic lexicon for verbs, nouns playing the predicative role frozen expressions including verbal and adjectival idioms, and adverbs from the Lexicon-Grammar tables, called LGLex [3].

Then, we converted the verbs and predicative nouns into the Alexina framework, that is the one of the Lefff lexicon (Lexicon of French inflected form) [6], a large-coverage morphological and syntactic lexicon for French.

This enables its integration in the FRMG parser (French MetaGrammar) [7], a large-coverage deep parser for French, based on Tree Adjoining Grammars (TAG), that usually relies on the Lefff. The result is a variant of the FRMG parser, that we shall call FRMG\textsubscript{LGLex}, to distinguish it from the standard FRMG\textsubscript{Lefff}.

\begin{enumerate}
\item All tables are fully available under the LGPL-LR license at \url{http://infolingu.univ-mlv.fr/english > Language Resources > Lexicon-Grammar > Download}.
\item Lefff is available at \url{http://gforge.inria.fr/projects/alexina/}.
\item FRMG is available at \url{http://gforge.inria.fr/projects/mgkit/}, with a visualization of the grammar FRMG on \url{http://alpage.inria.fr/frmgdemo}.
\end{enumerate}
In this article we present a method for enlarge a lexicon (with frequencies information), that is useful for parsing and others NLP applications. We show an example enlarging the verbal LGLex lexicon of French, using several corpora extracted from the evaluation campaign for French parsers Passage [5]. To do that, we use the results of the FRMG parser with the IRASUBCAT tool [1].

2 IRASUBCAT

IRASUBCAT is a tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, syntactic function, etc.) in a corpus [1]. We are interested in using it to acquire information about verbs. IRASUBCAT takes as input a corpus in XML format. The output of IRASUBCAT is a lexicon, also in XML format, where each of the verbs under inspection is associated to a set of subcategorization patterns. The lexicon also provides information about frequencies of occurrence for verbs, patterns, and their co-occurrences in corpus.

Moreover, IRASUBCAT allows to integrate the output lexicon with a pre-existing one, merging information about verbs and patterns with information that had been previously extracted, possibly from a different corpus or even from a hand-built lexicon.

3 Experiment with IRASUBCAT and the LGLex lexicon of French

We want to use the results of FRMG parser on a big corpus with IRASUBCAT in order to improve the LGLex lexicon of French, adding the frequencies of occurrence for each entry and each subcategorization frame. To do this, we must:

– find a corpus with millions of words (using a small part for the experiment);
– parse the corpus with the FRMG parser, with and without the LGLex lexicon (i.e. only with the Lefff lexicon) – results with FRMG_{LGLex} and FRMG_{Lefff};
– convert both the processed corpus and the LGLex lexicon into XML format;
– use IRASUBCAT in order to add the frequencies of occurrence for each entry and each subcategorization frame in the LGLex lexicon from the corpus.

3.1 Conversion of the verbal LGLex lexicon

The input is the verbal LGLex lexicon, or more precisely, the extensional lexicon of LGLex lexicon in Lefff format, which contains each inflected form of the lemma and every possible redistribution.

In the output lexicon converted into XML format as IRASUBCAT output lexicon, each lemma is associated to a set of subcategorization patterns. In fact,
Adding frequencies to the LGLex lexicon with IRASUBCAT we simplify by omitting the realizations. So, we have only the syntactic functions because it’s more easy to find them in the processed corpus.

For each lemma represented by his identifier (for example, verb="achever__V_1_1"), which correspond to the 1st entry in the verb class 1), a count of occurrences of this lemma is initialized at 0 (count_oc_verb="0"). We extracted the set of subcategorization patterns from all his inflected forms and all his redistributions and the number of different pattern is indicated (for example, different_patterns="6"). For each pattern (["obj", 'suj'], ['obl', 'suj'], ['obl2', 'suj'] and ['obl', 'obl2']), a count of occurrences of this pattern for this lemma and a count of occurrences of this pattern for all verbs are both initialized at 0 (count_w_verb="0" total_count="0").

We have in total 14,068 distinct lemmas. An example of the output lexicon:

```
<dictionary>
  <entry verb="achever__V_1_1" count_oc_verb="0">
   <tag name="fs" different_patterns="6">
     <pattern id="['obj', 'suj']" count_w_verb="0" total_count="0" rejected_patterns_freq_test="NO">
     ...</pattern>
   </tag>
  </entry>
</dictionary>
```

3.2 Conversion of the processed corpus with the FRMG parser

To use the result of the parsing in NLP applications of high-level, Forest utils5 represents the forest of dependencies in format XMLDep [7]. Basically, we represent in XMLDep format a graph of dependencies with nodes (lemmas), grouped in clusters (forms), with arcs describing the syntactic dependencies between nodes.

The processed corpus with FRMG\textsubscript{LGLex} used for the experiment is CPJ (Corpus Passage Jouet) with 100K sentences of AFP (Agence France-Presse), Europarl, Wikipedia and Wikisources, extracted from the corpus of the evaluation campaign (in 2009) for French parsers Passage [5].

The input is the processed corpus CPJ with the FRMG parser, more precisely, with FRMG\textsubscript{LGLex}. We want to extract only the useful information in a format directly readable by IRASUBCAT.

In the output in XML format, for each sentence of the corpus (for example, <sentence ID="12" corpus="frwikipedia_012" s="12" >), we extracted the verbs (cat="v") with their identifiers (for example, lemmaid="achever__V_1_1"). For each verb, we extracted the syntactic functions and we indicated the number of arguments (nb\_fs="2") and then, each syntactic function (fs) one by one (for example, fs="suj" for subject, and fs="obl2" for oblique).

5 Forest utils is a set of Perl scripts to convert between various formats for shared derivation forest produced by parsers for TAG (available at https://gforge.inria.fr/projects/lingwb/).
4 Conclusion

Using IRASUBCAT with the converted lexicon and the relevant information extracted of the processed corpus we can complete the lexicon with the frequencies of occurrence for each verb and each syntactic function. The processed corpus is the results of the FRMG parser with LGLex lexicon, so it could find wrong sense.

The next step is to consider the information on realizations, that we must extract from processed corpus, but it is not a straightforward task. Then we have to use the FRMG parser with Lefff lexicon only, without the LGLex lexicon influences the results. We could also use IRASUBCAT with another parser which is statistical, such as MaltParser, MSTParser, or Berkeley Parser [2]. And we could do a comparison using the original lexicon and the enlarged lexicon with that different parsers to verify that the accuracy is better using more information.

References