ASAI, September 16-20, 2013,
Córdoba, Argentina

Adding frequencies to the *LGLex* lexicon with IRASubcat

Elsa Tolone & Romina Altamirano

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

SUMMARY

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. To do that, we use the results of the FRMG parser with IRASubcat, 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.

Lexicon-Grammar tables

electronic inventory of lexical & syntactic information

LADL since 1970s ; LIGM since late 1990s

→ Conversion with *LGExtract*

Constant & Tolone, 2010

LGLex

NLP oriented syntactic lexicon

Tolone, 2012

→ Conversion and integration in FRMG parser

Tolone & Sagot, 2011 ; Tolone, Sagot & de La Clergerie, 2012

LGLex*-Le*fff* used with FRMG-*LGLex

syntactic lexicon integrated in a parser

Tolone, 2012

Syntactic-semantic classes of simple verbs

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ENT>	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	-	-	<E>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+	-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	+	Max s'est rendu à mon opinion
+	-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+	-	-	<E>	-	-	-	-	renoncer	-	-	+	+	-	-	+	-	-	Max renonce à son héritage

http://infolingu.univ-mlv.fr/english > Language Resources > Lexicon-Grammar > Download

Gross, 1975 ; 1994

+

IRASubcat

- a tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, syntactic function, etc.) or combination of them, from corpora.
- takes as input a corpus in XML format.
- the output 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.
- allows to integrate the output lexicon with a preexisting 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.

http://www.cs.famaf.unc.edu.ar/~romina/irasubcat/

Altamirano & Alonso Alemany, 2010

IRASubcat →

***LGLex* with frequencies**

syntactic lexicon in IRASubcat format

EXPERIMENT

To do the experiment with IRASubcat and the *LGLex* lexicon of French, we must:
- choose a corpus with millions of words, also we just only need a small part of this corpus for the experiment.
- parse the corpus with the FRMG parser, with and without the *LGLex* lexicon (i.e. only with the Le*fff* lexicon) – results with FRMG-*LGLex* and with FRMG-Le*fff*.
- convert both the processed corpus and the *LGLex* lexicon into XML format, required by IRASubcat.
- use IRASubcat in order to add the frequencies of occurrence extracted from the big corpus into the *LGLex* lexicon.

FUTURE WORK

- 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 Le*fff* 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.
- 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.

RESULTS

Corpus

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.

```
<dictionary>
  <entry verb="achever___V_1_1" count_oc_verb="1">
    <tag name="fs" different_patterns="4">
      <pattern id="['obj', 'suj']" count_w_verb="0" total_count="1001"
      rejected_patterns_freq_test="NO">
      </pattern>
      <pattern id="['obl', 'suj']" count_w_verb="0" total_count="214"
      rejected_patterns_freq_test="NO">
      </pattern>
      <pattern id="['obl2', 'suj']" count_w_verb="1" total_count="325"
      rejected_patterns_freq_test="NO">
      </pattern>
      <pattern id="['obl', 'obl2']" count_w_verb="0" total_count="0"
      rejected_patterns_freq_test="NO">
      </pattern>
    </tag>
  </entry>
</dictionary>
```

```
<ids_from>
  <entry verb="achever___V_1_1" total_count="1">
    <tag name="fs">
      <pattern id="['obj', 'suj']">
        <s_list>
        []
        </s_list>
      </pattern>
      <pattern id="['obl', 'suj']">
        <s_list>
        []
        </s_list>
      </pattern>
      <pattern id="['obl2', 'suj']">
        <s_list>
        ['12']
        </s_list>
      </pattern>
      <pattern id="['obl', 'obl2']">
        <s_list>
        []
        </s_list>
      </pattern>
    </tag>
  </entry>
</ids_from>
```

Number of occurrences of patterns

pattern	*total_count*
['obj', 'suj']	1001
['obl2', 'suj']	325
['obl', 'suj']	214
['att', 'suj']	142
['loc', 'suj']	92
['obj', 'suj']	91
['suj']	62
['objde', 'suj']	55
['obj']	26
['dloc', 'suj']	11
others	0

Number of occurrences of verbs

verb or number of verbs	*total_count*
être_____2	63
pouvoir___V_1_88	60
devoir___V_1_38	37
faire_____2	22
dire___V_9_130	19
vouloir___V_15_82	17
2	16
avoir___V_37E_10	13
2	12
3	10
4	9
3	8
8	7
12	6
14	5
30	4
63	3
192	2
740	1
13 043	0