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 enlarging 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.

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 Lefk 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 BerkeleyParser.
* And we could do a comparison using the original lexicon and the enlarged lexicon with different parsers to verify that the accuracy is better using more information.

RESULTS

The frequencies of occurrence for each input and each subcategorization frame for 14,068 distinct lemmas.

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 Lefk lexicon) - results with FRMG-LGLex and with FRMG-Le
* 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.

SYNTHETIC-SEMANTIC CLASSES OF SIMPLE VERBS

Lexicon-Grammar tables

Electronic inventory of lexical & syntactic information

Conversion with LGetract

Constant & Tolone, 2010

LGLex

NLP oriented syntactic lexicon

Conversion and integration in FRMG parse

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

IRASubcat

A tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, semantic function, etc.) or combination of them, from corpus.

* 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.

LGLex-Leff used with FRMG-LGLex

Syntactic-semantic information in IRASubcat format

Syntactic-semantic classes of simple verbs

Corpus

CPJ (Corpus Passage Jouet) with 108K sentences of AFP (Agency France-Presse), Europap, Wikisources, extracted from the corpus of the evaluation campaign (in 2009) for French parsers Passage.

LGLex

Lexicon-grammar tables

Electronic inventory of lexical & syntactic information

Conversion with LGetract

Constant & Tolone, 2010

LGLex

NLP oriented syntactic lexicon

Conversion and integration in FRMG parse

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

IRASubcat

A tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, semantic function, etc.) or combination of them, from corpus.

* 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.

LGLex-Leff used with FRMG-LGLex

Syntactic-semantic information in IRASubcat format

Syntactic-semantic classes of simple verbs

Corpus

CPJ (Corpus Passage Jouet) with 108K sentences of AFP (Agency France-Presse), Europap, Wikisources, extracted from the corpus of the evaluation campaign (in 2009) for French parsers Passage.