Frequencies of occurrence of entries and subcategorization frames in 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 other 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.

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 Leff lexicon) - results with FRMG-LGLex and with FRMG-Leff.
* 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 the processed corpus, but it is not a straightforward task.
* Then we have to use the FRMG parser with Leff 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.