IRASubcat

a highly customizable, language independent tool for the acquisition of verbal subcategorization from corpus

Ivana Romina Altamirano

NLP group
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Argentina

Laura Alonso i Alemany

IRASubcat obtains a lexicon of verbs with their subcategorization patterns from a corpus. Patterns are found in the corpus and associated with verbs with which they co-occur if they pass a hyphotesis test and a frequency test. Possible adjuncts are also identified in patterns. The system is flexible enough to take in input corpora annotated at various levels, in XML format. It is highly customizable, and provides default values for most parameters. IRASubcat is a free and open source tool available at http://irasubcat.com.ar

Input of the system

Corpus

XML format
UTF-8 encoding
Tags for verbs are mandatory
Other analyses are optional

```
<englishCorpus>
  <phrase id='5'>
    <w lem='the' pos='DT'>The</w>
    <w lem='dog' pos='N'>dog</w>
    <w lem='be' pos='V'>is</w>
    <w lem='big' pos='A'>big</w>
  </phrase>
</englishCorpus>
```

```
<corpus>
  <sentence id='1'>
    <w>The</w>
    <w>children</w>
    <w mp='v'>sleep</w>
  </sentence>
</corpus>
```

Command line

```
IRASubcat.py corpus.xml morph='v' phrase lem  [config.cfg]
```

Existing Dictionary

List of verbs to study

Configuration File

Configuration option	Value Range	Default Value
VERB LIST =	NO / <VerbalList.txt>	NO
EXISTING DICTIONARY =	NO / <Dictionary.xml>	NO
LENGTH OF VERBAL CONTEXT =	ALL / <number>	3
COMPLETE WITH EMPTY WORD =	NO / <EMPTY>	NO
KEEP ORDER =	NO / YES	NO
TARGET TAGS =	<target tags>	sint
USE LEXICAL FORM OF WORDS =	NO / YES	NO
INTRODUCE VERBAL MARK =	NO / YES	NO
COLLAPSE PATTERNS =	NO / YES	NO
MAX ITERATION TO COLLAPSE PATTERNS =	<number>	NO
MIN FREQUENCY OF VERBS =	0 / <number>	0
MIN REL FREQUENCY OF PATTERNS =	0 / <number>	0
USE LIKELIHOOD RATIO TEST =	NO / YES	NO

Consistency of arguments

Filter sentences with verbs to study

```
<Corpus>
  <s id='5'> … <w verb='talk'></w> … </s>
  <s id='6'> … <w verb='run'></w> … </s>
  <s id='7'> … <w verb='speak'></w> …  </s>
</Corpus>
```

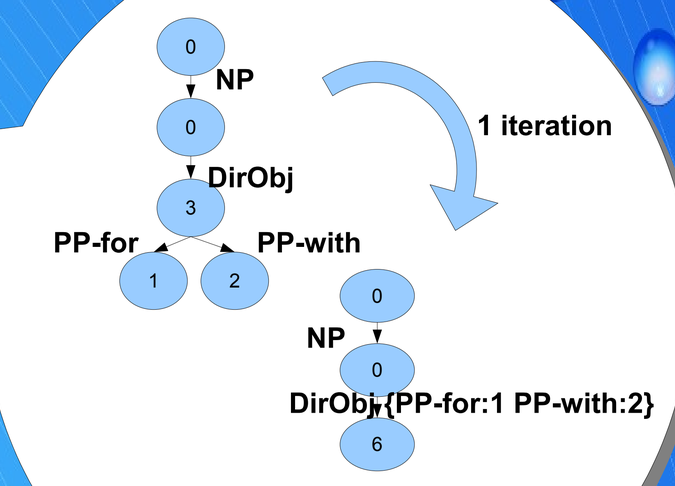
Identify patterns

EAT → [NP DirObj PP-with] [NP DirObj] [NP DirObj] [NP DirObj PP-with] [NP DirObj] [NP DirObj PP-for]

Verbs → patterns

EAT → [NP DirObj PP-with] → 2, [NP DirObj] →3, [NP DirObj PP-for] → 1

Collapse Patterns



[NP DirObj] (PP-with:2, PP-for:1)

Verbs → patterns

Associate Verbs with Patterns

Frequency Filter

Hypothesis Test

Resulting Dictionaries

EAT
[NP DirObj] {PP-with:2, PP-for:1}
in sentences 2, 5,8
[NP]
in sentences 1, 3, 7

DICTIONARY OF VERBS AND PATTERNS

```
<dictionary  execute=...>
  <entry verb="caminar" count_oc_verb="1">
    <tag name="sint" different_patterns="1">
      <pattern id="['ADJ', 'ART', 'NC', 'PREP', 'V']"
        count_w_verb="1" total_count="2"
        rejected_patterns_freq_test="NO"
        rejected_likelihood_test="NO_DECIDE">
      </pattern>
    </tag> ...
  </entry> …
  ...
</dictionary>
```

STATISTICS OF THE EXECUTION

Count of sentences
Count of verb studied
Count of patterns founded
Count of rejected patterns (frequency test)
Count of rejected patterns (Likelihood Ratio test)

Evaluation

Spanish

SenSem Corpus
- manually annotated
- experiments for syntactic functions
- evaluated against verbal lexicon

Filter	Prec	Rec	F
Freq	.79	.70	.74
Lik90	.42	.46	.44
Lik 99	.31	.36	.33

Russian

- journalistic raw corpus
- 1000000 words
- tagged with TreeTagger
- anecdotic evaluation

eat **transitive patterns**
sleep **intransitive**
give **ditransitive**

Я	ела	яблоко
Subj	Verb	DirObj
Ты	спал	
Subj	Verb	

Useful for ...

- Parsing
- Sense Disambiguation
- Information Extraction

Further information

http://irasubcat.com.ar
http://www.cs.famaf.unc.edu.ar/~pln
http://grial.uab.es/projectes/sensem

Acknowledgements

This research has been partially funded by projects KNOW, TIN2006-15049-C03-01 and Representation of Semantic Knowledge TIN2009-14715-C04-03 of the Spanish Ministry of Education and Culture, and by project PAE-PICT-2007-02290, funded by the National Agency for the Promotion of Science and Technology in Argentina.