LexSchem

A Large Subcategorization Lexicon for French Verbs

Cédric Messiant, Anna Korhonen and Thierry Poibeau

30 May 2008

Lexical Resources

- Lexical resources are key components for most NLP applications:
  - Machine translation;
  - Parsing;
  - Information extraction.
- These applications require information about the relations between a predicate and its arguments

Lexical Resources

- Lexical resources are key components for most NLP applications:
  - Machine translation;
  - Parsing;
  - Information extraction.
- These applications require information about the relations between a predicate and its arguments

$$\Rightarrow \text{Subcategorization frames (SCF)}$$

Subcategorization Frames

- SCFs are the combinations of arguments that a predicate can take.
- SCFs don't include adjuncts (e.g. locative or temporal phrases).
- SCFs are one of the most useful lexical information for NLP applications.

An Example

[Ces propriétaires]	[achètent]	[le carburant]	[à la compagnie] .
[These owners]	[buy]	[the fuel]	[from the company] .
[NP]	[VERB]	[NP]	[PP<à>] .

Subcategorization Acquisition

- We need automatic methods to acquire such lexicons from raw corpora.
- Such methods has been proposed for several languages
  - English: Briscoe & Carroll (1997), Korhonen & al (2006), Preiss & al (2007)...
  - German: Schulte Im Walde (2000)
- No experiment has been done so far for French, except on a very small scale: Chesley & Salmon-Alt (2006)

Subcategorization Acquisition

- We need automatic methods to acquire such lexicons from raw corpora.
- Such methods has been proposed for several languages
  - English: Briscoe & Carroll (1997), Korhonen & al (2006), Preiss & al (2007)...
  - German: Schulte Im Walde (2000)
- No experiment has been done so far for French, except on a very small scale: Chesley & Salmon-Alt (2006)
- In this talk, we present **LexSchem**, a subcategorization lexicon for French verbs which has been generated by **ASSCI**, our subcategorization acquisition system.

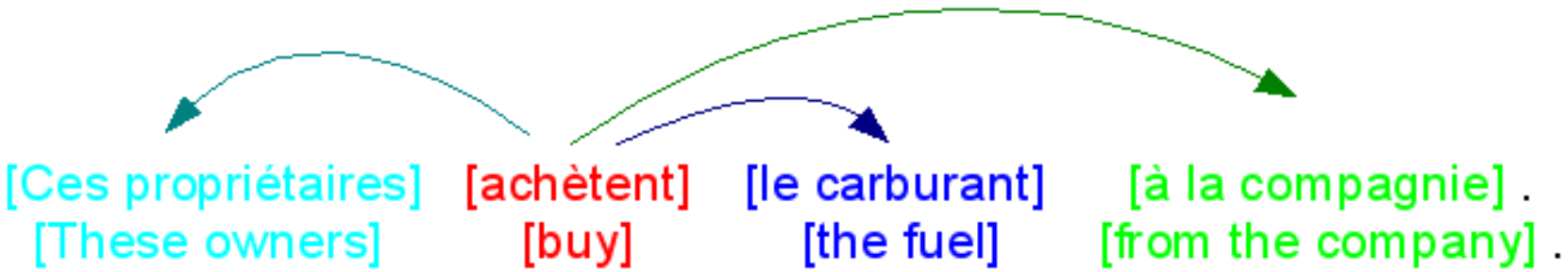
ASSCI: An Overview

- **ASSCI** is an automatic SCF acquisition system for French.
- **ASSCI** takes raw corpus data as input and extracts a subcategorization lexicon.
- **ASSCI** is made of 4 successives modules:
  1. Preprocessor,
  2. Pattern extractor,
  3. SCF builder,
  4. SCF filter.

ASSCI: Preprocessor

- **TreeTagger**: lemmatizer and morphosyntactic tagger (Schmid, 1994).
- **Syntex**: shallow dependency parser for French (Bourigault, 2005). It makes no distinction between arguments and adjuncts: every dependency is attached to the verb
  ⇒ statistical information can help the disambiguation process.

Example



ASSCI: Pattern Extractor

- Input: A corpus that has been tagged and parsed.
- Output: for each occurence of each verb, its list of dependencies.
- The subject is ignored (almost always present in French).

Example

[Ces propriétaires] [achètent] [le carburant] [à la compagnie] .
[These owners] [buy] [the fuel] [from the company] .
Extracted pattern:

```
Verb|acheter + Noun|carburant + Prep|à (+ Noun|compagnie).
Verb|to buy + Noun|fuel + Prep|from (+ Noun|company).
```

ASSCI: SCF builder

- Input: the ouput of the pattern extractor.
- Output: SCF candidates for each verb and the total number of occurrences of each tentative SCF (for each verb) in the corpus.

Some SCF candidates and their number of occurences in the corpus for "acheter (to buy)"

NP (2379)
PP[à+NP] (101)
NP_PP[à+NP] (379) *(Ces propriétaires achètent le carburant à la compagnie.)*
NP_PP[pour+NP] (123)

ASSCI: SCF Filter

- Input: output of the SCF builder, which is always noisy due to tagging or parsing errors.
- Output: the filtered lexicon i.e. a list of SCFs with their frequencies for each verb.
- Method: Maximum likehood estimates (comparison of the relative frequency of each SCF (for a verb) with a simple threshold).

Some SCFs for "acheter (to buy)" (threshold= 0.035)

```
NP (0.497)
NP_PP[à+NP] (0.079)
PP[à+NP] (0.021)
NP_PP[pour+NP] (0.025)
```

ASSCI: SCF Filter

- Input: output of the SCF builder, which is always noisy due to tagging or parsing errors.
- Output: the filtered lexicon i.e. a list of SCFs with their frequencies for each verb.
- Method: Maximum likehood estimates (comparison of the relative frequency of each SCF (for a verb) with a simple threshold).

Some SCFs for "acheter (to buy)" (threshold= 0.035)

```
NP (0.497)
NP_PP[à+NP] (0.079)
PP[à+NP] (0.021)
NP_PP[pour+NP] (0.025)
```

Experiment settings

- A large lexicon has been produced for French.
- An experiment has been done on 10 years from the French newspaper *Le Monde*:
  - 200 millions words,
  - Various domains, but mainly politics and foreign affairs.
- We used *TreeTagger* and *Syntex* for the preprocessing stage.

Lexical Entries

- **ID**: the identifier of the entry in the lexicon;
- **SUBCAT**: a summary of the target verb and SCF;
- **VERB**: the verb;
- **SCF**: the subcategorization frame;
- **COUNT**: the number of corpus occurences found for the combination of the verb with this SCF;
- **RELFREQ**: the relative frequency of the SCF for this verb;
- **EXAMPLES**: 5 corpus occurrences exemplifying this entry (the examples are provided in a separate file).

Lexical Entries: An Example

```
:ID:        00109
:SUBCAT:    acheter :  NP_PP[à+NP]
:VERB:      ACHETER+acheter
:SCF:       NP_PP[à+NP]
:COUNT:     379
:RELFREQ:   0.107
:EXAMPLE:   525;526;527;528;529
```

LexSchem

- 11,149 lexical entries in total (i.e. 11,149 different combinations of verb + SCF);
- 3268 verb types (a verb and its reflexive form are counted as 2 different types);
- 336 distinct SCFs.

Web distribution of LexSchem

- **LexSchem** is freely available under the Lesser General Public License For Linguistic Resources (LGPL-LR).
- A web interface is provided at the same address. It enables viewing lexical entries for each verb along with practical examples.
- `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`

Gold Standard

- Several lexicons exist for French (Lexicon Grammar, Lefff, DicoValence, TreeLex, TLFI...)
- None of them can directly be used as a gold standard $\Rightarrow$ manual adaptation of the resource is required!
- For more details about the issues of the comparison to a gold standard, see (Poibeau & Messiant, LREC2008).

Evaluation

We calculated type precision, type recall and F-measure for 20 verbs against a gold standard (*Trésor de la Langue Française Informatisé (TLFI)*).

	Our work	Chesley & Salmon-Alt (2006)	Korhonen *& al.* (2006)
Precision	0.79	0.87	0.81
Recall	0.55	0.54	0.46
F-Measure	0.65	0.67	0.58

Table: Comparison with recent works in French and English

Conclusion

LexSchem

- an automatically acquired subcategorization lexicon for French,
- large scale (more than 3000 verbs),
- freely available on the web.

Future Work

- We plan to perform a better evaluation (more verbs),
- We plan to evaluate the method on other domains (e.g. medical domain),
- We plan to extract semantic classes from **LexSchem** (Levin, 1993).

Thank you for your attention!

INTERFACE LEXSCHEM - REMERCIER

Choisir un schéma de sous-catégorisation : SN_SP[de+VINF] (0.152) Voir les informations Afficher les analyses de syntex

VERBE	CADRE DE SOUS-CATÉGORISATION	NOMBRE D'OCCURENCES	FRÉQUENCE RELATIVE
remercier	SN_SP[de+VINF]	113	0.152

Pour lui dire du plus près qu' ils l' aiment et le remercient de l' aimer .

Remercions la Radiodiffusion française d' avoir rendu cet hommage au très grand musicien que négligent vraiment trop nos associations symphoniques .

Ce matin , on a vu un jeune homme d' une de ces sociétés très en vue de la Silicon Valley - qu' elle ne nommera pas - venir commander une Porsche en anticipant la hausse de ses stock options . « Acheter une Porsche avec de l' argent d' emprunt ! » , s' indigne Dorothy Miller . « S' il y a une correction , c' est à ces gens -là que ça fera mal . » En attendant , tous les matins , chez elle , quand elle « parle à son mari » brutalement emporté l' an dernier par une attaque cardiaque , elle le remercie de lui avoir confié , il y a cinq ans , son plan

Il fait aussi le plein des nominations " techniques " (photo , scénario , musique , son , décor , montage , costumes) , comme si la profession remerciait Claude Berri d' exister , de prendre des risques , de faire de gros films qui donnent du boulot à beaucoup de monde ...

Remercions en revanche La Cinquième d' avoir fêté sa naissance avec la projection du superbe film de Jean Renoir , le Fleuve .