Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français

Présentation des auteurs

- Cédric Messiant, chercheur au laboratoire d'Informatique de Paris-Nord ;
- Kata Gábor, chercheuse au département Language Technology dans un institut de recherche à Budapest ;
- Thierry Poibeau, chercheur au laboratoire LaTTiCe.

Plan

- Description de la méthode d'acquisition à partir de corpus (ASSCi) ;
- Présentation du lexique obtenu (LexSchem) ;
- Évaluation du lexique obtenu ;
- Production automatique de classes de verbes à partir du lexique de sous-catégorisation.

Le système ASSCi

- Système d'acquisition automatique de schémas de sous-catégorisation ;
- Développé par les auteurs de l'article en 2008 ;
- 3 modules principaux : extracteur de pré-schémas de sous-catégorisation, constructeur de schémas candidats, filtre de schémas non pertinents.

ASSCi – Pré-traitement

- Lemmatisation, annotations morpho-syntaxiques et analyse de surface
- Sortie du pré-traitement (TreeTagger + SYNTEX) :

```
PRO:PER     il          | Pro|il|Il|1|SUJ;4|
PRO:PER     la|le       | Pro|le|les|2|OBJ;4|
PRO:PER     lui         | Pro|lui|lui|3|PREP;4|
VER:subp    reprocher   | VCONJS|reprocher|reproche|4||SUJ;1,OBJ;2,PREP;3,PREP;5
PRP:det     au          | Prep|au nom de|au nom du|5|PREP;4|NOMPREP;6
NOM         nom         |
PRP:det     du          |
NOM         Sartre      | NomPrXXInc|Sartre|Sartre|6|NOMPREP;5|
PRO:REL     que         | CSub|que|qu'|7||COMP;9
PRO:PER     il          | Pro|il|il|8|SUJ;9|
VER:pres    aimer       | VCONJS|aimer|aime|9|COMP;7|SUJ;8
SENT        .           | Typo|.|.|10||
```

ASSCi – Extracteur

- Module n°1 : extracteur de schéma de sous-catégorisation locaux (pré-SSC) ;
- Pré-SSC à partir de la phrase « *Il les lui reproche au nom de Sartre qu'il aime* » :

```
0100.anasynt!d686339p6_2!21
REPROCHER+reprocher
[P-OBJ:SP<au_nom_de+SN>:Sartre, SUJ:SN:il, OBJ:SN:le, A-OBJ:SP<à+SN>:lui]
```

ASSCi – Constructeur

- Module n°2 : constructeur de schémas de sous-catégorisation candidats (= non filtrés) à partir des pré-SSC ;
- Pas de liste de schémas prédéfinie ;
- Les schémas correspondent à un ensemble d’occurrences du corpus ;
- Les éléments des SSC sont ordonnés selon leur fonction, selon un ordre défini.

ASSCi – Constructeur : autres rôles

- Suppression des compléments doublons des pré-SSC ;
- Comptabilisation du nombre d'occurrences de chaque SSC pour chaque verbe ;
- Calcul de la fréquence relative des SSC ;
  - ces données seront utilisées lors de l'étape de filtrage

ASSCi – Filtrage

- Module n°3 : filtrage des SSC candidats ;
- Compare la fréquence relative des SSC candidats à un seuil ;

  → Si la fréquence est inférieure au seuil, le SSC candidat est rejeté par le module.
- Seuils différenciés pour certains SSC.

LexSchem – Présentation

- Application d'ASSCi sur le corpus LM10 → lexique de sous-catégorisation pour le français (LexSchem) ;
- Corpus LM10 : corpus journalistique, articles du quotidien *Le Monde* sur 10 ans (200 millions de mots) ;
- Version 3 à l'époque de l'article (2010).

LexSchem – Illustration

- 10 928 entrées correspondant à des combinaisons verbes-SSC ;
- Exemple d'une entrée de LexSchem :

```
<ID>          2610
<VERB>        REPROCHER+reprocher
<VERB_NB_OCC> 9757
<SCF>         [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]
<NB_OCC>      2128
<VERB_NB_SCF> 118
<REL_FREQ>    0.218099825766117
<SEQ_ID>      0100.anasynt!d6863p6_2!4, 0100.anasynt!d6835p2_7!9, ...
<NB_ARGS>     3
<ARG0>        il,on,...
<ARG1>        le,manque,...
<ARG2>        lui,secrétaire,...
<PASS>        oui
```

Évaluation – comparaison

- LexSchem comparé à TreeLex et DicoValence :
  - Une partie des ressources de référence n’a pas été retrouvé dans LexSchem (formes réduites de SSC complexes : un post-traitement réduit de 50 % le nombre de SSC manquants) ;
  - Évaluation manuelle sur 150 verbes : 108 nouveaux SSC valides pourraient être ajoutés à TreeLex et 75 à DicoValence.

Évaluation – LexSchem-Europarl

- Application de la méthode à un nouveau corpus : Europarl (fr) ;
- Europarl : corpus parallèle, actes du Parlement européen entre 1993 et 2003 ;
  - Comparaison avec les résultats obtenus sur LexShem-LM10 : apparition de nouveaux SSC dus au contexte particulier du corpus d'Europarl.

Production de classes de verbes

- Hypothèse : des verbes partageant des comportements syntaxiques similaires peuvent former des classes homogènes sur le plan sémantique ;
- Algorithme à partir de LexSchem ;
- Résultats : 20 % des classes sont parfaitement homogènes, 43 % contiennent 1 verbe incorrect au maximum.

Conclusion

- Le système ASSCi :
  - Est capable de repérer des données nouvelles afin d'enrichir des lexiques existants ;
  - Permet d'acquérir des données profilées en fonction d'un corpus donné.