

LexSchem

Acquisition automatique d'informations sur la valence des verbes
à partir de gros corpus

Cédric Messiant et Thierry Poibeau

Laboratoire d'Informatique de Paris-Nord

9 juin 2008

Introduction

Le système d'acquisition ASSCI

Aperçu

Les modules du système ASSCI

Expériences

Protocole expérimental

LexSchem

Evaluation

Conclusion

Ressources lexicales

- ▶ Les ressources lexicales sont essentielles pour les applications de TAL :
 - ▶ Traduction automatique ;
 - ▶ Analyse syntaxique ;
 - ▶ Extraction d'information.
- ▶ Toutes ces applications ont notamment besoin d'informations sur la relation entre les prédicats et leurs arguments

Ressources lexicales

- ▶ Les ressources lexicales sont essentielles pour les applications de TAL :
 - ▶ Traduction automatique ;
 - ▶ Analyse syntaxique ;
 - ▶ Extraction d'information.
- ▶ Toutes ces applications ont notamment besoin d'informations sur la relation entre les prédicats et leurs arguments
 - ⇒ Schémas de valence (sous-catégorisation)

Motivations pour une approche automatique (1/2)

- ▶ On ne veut pas concurrencer les travaux “manuels”.
 - ▶ On obtient (évidemment) des performances moins bonnes (données en partie bruitées).
 - ▶ On obtient (évidemment) des données moins riches.
- ▶ On veut en revanche produire automatiquement (donc à moindre coût) des ressources quand il n'y a rien de disponible dans un domaine donné
 - ▶ Médecine, santé, etc.
 - ▶ Droit, jurisprudence, etc.

Motivations pour une approche automatique (2/2)

- ▶ On veut aussi étudier (à moyen terme) l'importance de la redondance de la langue et ce que l'on peut en tirer (y compris du point de vue de l'acquisition du langage)
 - ▶ Peut-on désambiguïser facilement les formes de surface (sans connaissances extérieures) ?
 - ▶ Peut-on faire la différence entre argument et modifieur sur une base statistique ?
 - ▶ Quel peut être l'apport des méthodes d'apprentissage non supervisées pour le TAL ?
 - ▶ Quelles sont les particularités des langues de spécialité ?

Schémas de valence

Qu'est-ce qu'un schéma de valence ?

- ▶ Un schéma inclut le nombre et le type des arguments autour du prédicat.
- ▶ Il n'inclut pas les adjoints (e.g. compléments de temps, sauf exceptions).

Ces informations sont très utiles pour le TAL.

Un exemple

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

Etat de l'art

- ▶ Nous cherchons à définir une méthode pour l'acquisition de ces schémas à partir de corpus bruts.
- ▶ Des méthodes proches sont disponibles pour plusieurs langues
 - ▶ Anglais : Briscoe & Carroll (1997), Korhonen & al (2006), Preiss & al. (2007)...
 - ▶ Allemand : Schulte Im Walde (2000)
- ▶ Aucune expérience d'envergure sur le français (cf. l'expérience de Chesley & Salmon-Alt (2006) porte sur une centaine de verbes seulement)

Acquisition des schémas de valence

- ▶ Attention : l'acquisition à partir d'un Treebank est une tâche relativement différente (qui obtient évidemment des taux de succès beaucoup supérieurs) (Kupsc, 2007)

Acquisition des schémas de valence

- ▶ Attention : l'acquisition à partir d'un Treebank est une tâche relativement différente (qui obtient évidemment des taux de succès beaucoup supérieurs) (Kupsc, 2007)
- ▶ Nous présentons ici la ressource **LexSchem**, un dictionnaire de schémas de valence qui a été généré par le système d'acquisition **ASSCI**.

ASSCI: un aperçu

- ▶ **ASSCI** est un système d'acquisition automatique pour le français.
- ▶ **ASSCI** prend en entrée un corpus brut et extrait dynamiquement les schémas de valence associés aux verbes.
- ▶ **ASSCI** est composé de 4 modules :
 1. Prétraitements,
 2. Extraction des compléments,
 3. Construction du cadre,
 4. Filtrage des cadres.

ASSCI : Prétraitement

- ▶ **TreeTagger**: lemmatiseur et analyseur morpho-syntaxique (Schmid, 1994).
- ▶ **Syntex**: analyseur de surface pour la détection des dépendances (Bourigault, 2005). Syntex ne fait pas la distinction entre arguments et modifieurs. Tous les compléments sont directement rattachés au verbe
⇒ des informations statistiques peuvent aider à la désambiguïsation.

Exemple

[Ces propriétaires] [achètent] [le carburant] [à la compagnie] .

[These owners] [buy] [the fuel] [from the company] .

ASSCI: Extraction de patrons

- ▶ Entrée : Le corpus analysé et annoté.
- ▶ Sortie : pour chaque occurrence de chaque verbe, sa liste de dépendances.
- ▶ Le sujet est laissé de côté (approximation de la nature du sujet).

Exemple

[Ces propriétaires] [achètent] [le carburant] [à la compagnie] .

[These owners] [buy] [the fuel] [from the company] .

Patrons extraits:

Verb|acheter + Noun|carburant + Prep|à (+ Noun|compagnie).

Verb|to buy + Noun|fuel + Prep|from (+ Noun|company).

ASSCI: Constructeur de cadres

- ▶ Entrée : la sortie de l'extracteur de patrons.
- ▶ Sortie : Les cadres candidats pour chaque verbe et le nombre d'occurrences pour chaque candidat dans le corpus.

Exemple pour le verbe “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 : le filtre de cadres

- ▶ Entrée : la sortie du constructeur de cadres (qui est toujours bruitée du fait des erreurs d'analyse).
- ▶ Sortie : le dictionnaire filtré, i.e. la liste des cadres pour chaque verbe, avec leur fréquence.
- ▶ Méthode : *Maximum likelihood estimates* (comparaison de la fréquence relative des cadres avec des seuils adapté à chaque cadre).

Exemple pour “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 : le filtre de cadres

- ▶ Entrée : la sortie du constructeur de cadres (qui est toujours bruitée du fait des erreurs d'analyse).
- ▶ Sortie : le dictionnaire filtré, i.e. la liste des cadres pour chaque verbe, avec leur fréquence.
- ▶ Méthode : *Maximum likelihood estimates* (comparaison de la fréquence relative des cadres avec des seuils adapté à chaque cadre).

Exemple pour “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)

Expérience

- ▶ Un dictionnaire à large couverture a été produit pour le français.
- ▶ Le dictionnaire a été produit à partir de l'analyse automatique de 10 ans du journal *Le Monde*:
 - ▶ 200 millions de mots,
 - ▶ Plusieurs domaines avec une prépondérance pour la politique et les affaires étrangères.
- ▶ Le prétraitement est effectué avec *TreeTagger* et *Syntax*.

Structure d'une entrée lexicale

une entrée lexicale comporte les informations suivantes :

- ▶ **ID**: identifiant de l'entrée dans le dictionnaire;
- ▶ **SUBCAT**: forme schématique du couple verbe-schéma de valence;
- ▶ **VERB**: le verbe en question;
- ▶ **SCF**: le cadre de valence;
- ▶ **COUNT**: le nombre d'occurrences en corpus du couple verbe-cadre;
- ▶ **RELFREQ**: fréquence relative du cadre pour le verbe en question;
- ▶ **EXAMPLES**: 5 exemples illustrant l'entrée en question (exemples issus du journal *Le Monde*).

Un exemple d'entrée lexicale

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

LexSchem

- ▶ 11 149 entrées lexicales (i.e. 11 149 combinaisons différentes verbe-cadre);
- ▶ 3 268 verbes différents (les formes réfléchies sont comptées à part mais il faudrait une analyse plus fine du phénomène des réfléchis);
- ▶ 336 cadres de valence différents.

Distribution de LexSchem sur le web

- ▶ **LexSchem** est disponible librement sous la licence LGPL-LR (*Lesser General Public License For Linguistic Resources*).
- ▶ Une interface graphique en ligne permet d'effectuer des requêtes directement dans la base et d'avoir accès aux exemples .
- ▶ <http://www-lipn.univ-paris13.fr/~messiant/lexschem.html>

Références (*Gold Standard*)

- ▶ Il existe différents dictionnaires ou ressources électroniques développées à la main pour le français (Lexique Grammaire, Lefff, DicoValence, TreeLex, TLFi...)
- ▶ Aucune de ces ressource ne peut être utilisée directement comme référence \Rightarrow une adaptation manuelle de ces ressources est donc nécessaire !
- ▶ Pour plus de détails sur les difficultés liées à l'évaluation, voir (Poibeau & Messiant, LREC2008).

Evaluation

On calcule la précision, le rappel et la F-mesure pour 20 verbes, en comparant ce que l'on obtient à une référence, ici le TFLi (*Trésor de la Langue Française Informatisé*).

	Notre travail	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: Comparaison avec des travaux récents sur l'anglais et le français

Que signifient ces chiffres ?

- ▶ Pas grand chose !?
 - ▶ Il faut une analyse du rappel (les éléments manquants dans LexSchem sont-ils absents du Monde ? Correspondent-ils à des emplois anciens enregistrés dans le TLFi ?)
 - ▶ Il faut une analyse de la précision
 - ▶ Il faut une analyse par cadre de valence
- ▶ Cette analyse est en cours au LIPN.

Conclusion

LexSchem

- ▶ Une ressource avec des informations de valence acquises automatiquement à partir de corpus non annotés manuellement,
- ▶ large couverture (plus de 3 000 verbes – ça ne suffit évidemment pas à en faire une ressource “complète”),
- ▶ disponible librement sur le web.

Perspectives

Perspectives

- ▶ Une évaluation plus fine sur plus de verbes est en cours (avec une analyse plus fine des résultats au niveau linguistique : quelles sont les distinctions qui ne sont pas faites en l'état ? Qu'est-ce qui est faisable automatiquement ? etc.),
- ▶ Une évaluation sur un autre domaine est planifiée (*a priori* sur le domaine médical) : il s'agit de valider le fait que la méthode peut produire des ressources utiles pour des domaines où il y a peu de données disponibles.
- ▶ On projette enfin d'utiliser les informations obtenues pour générer des classes lexicales, en reprenant les hypothèses de (Levin, 1993).

Collaborations

Adaptation aux noms prédicatifs

- ▶ Besoin exprimé par une équipe de l'Université de Lille 3 : avoir accès à une base de connaissance sur la valence des noms (notamment les noms d'actions).
- ▶ ASSC peut facilement être adapté pour répondre à ce besoin
- ▶ L'outil permet d'offrir très rapidement une base avec des informations sur la productivité aux linguistes intéressés
- ▶ Un travail de validation linguistique reste évidemment nécessaire mais l'approche est
 - ▶ plus rapide qu'un travail purement manuel
 - ▶ plus systématique
 - ▶ moins coûteuse

Références

- ▶ Cédric Messiant. 2008. ASSCI: A Subcategorization Frames Acquisition System for French Verbs. In Association for Computational Linguistics (ACL, Student Research Workshop), Columbus, Ohio.
- ▶ Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In Language Resource and Evaluation Conference (LREC), Marrakech.
- ▶ Thierry Poibeau and Cédric Messiant. 2008. Do we still need gold standard for evaluation ? In Proceedings of the Language Resources and Evaluation Conference (LREC), Marrakech.