Conférence TALN 2000, Lausanne, 16-18 octobre 2000

SVETLAN'
ou
Comment Classer des Noms en fonction de leur Contexte

Gaël de Chalendar et Brigitte Grau, LIMSI/CNRS, BP 133, 91 403 Orsay Cedex, France,
{Gael.de.Chalendar,Brigitte.Grau}@limsi.fr

Résumé

L'utilisation de connaissances sémantiques dans les applications de TAL améliore leurs performances. Cependant, bien que des lexiques étendus aient été développés, il y a peu de ressources non dédiées à des domaines spécialisés et contenant des informations sémantiques pour les mots. Dans le but de construire une telle base, nous avons conçu le système SVETLAN', capable d'apprendre des catégories de noms à partir de textes, quel que soit leur domaine. Dans le but d'éviter de créer des classes générales regroupant tous les sens des mots, les classes sont apprises en fonction de l'usage des mots en contexte.

1. INTRODUCTION

L'amélioration des performances des applications de TAL nécessite d'effectuer des analyses de plus en plus profondes des documents dans le but d'extraire leur signification plus précisément. Dans ce domaine, les techniques existantes correspondent à deux approches, radicallement différentes.

D'une part, les analyses de surface sont fondées sur les distributions des mots et leur importance dans un corpus. Elles utilisent des connaissances lexicales sur les mots et, éventuellement, des connaissances sémantiques générales, comme la classe sémantique d'un mot et les relations entre les mots et les concepts. De telles informations sont encodées par exemple dans WordNet ou dans des thésaurus. Ce type d'approche est applicable à de grosses bases de textes, quel que soit leur sujet. En retour, ces systèmes ne peuvent mettre en œuvre une analyse détaillée et précise du fait qu'ils ne disposent pas d'une base de connaissance détaillée et structurée. Même en utilisant WordNet, ce problème persiste. Les mots dans WordNet font partie d'un ou plusieurs Synsets, qui sont des ensembles de synonymes. Ces catégories sont larges et il arrive que deux mots présents simultanément dans un ou plusieurs Synsets, mais considérés dans un contexte particulier, ne partagent dans ce cas plus la moindre part de sens.

D'autre part, des systèmes effectuent une analyse en profondeur. Ils nécessitent l'utilisation de bases de connaissances sémantiques extrêmement structuées aussi bien que des connaissance pragmatiques à propos des situations auxquelles les textes font référence (événements, liens causaux et entités). Ces systèmes ont pour but de produire une analyse sémantique des phrases

et de construire une représentation de la signification des textes. Leurs limites proviennent en grande partie de ces indispensables connaissances : il est très difficile et couteux de les produire, et ce même dans des domaines restreints. De plus, leur réutilisation dans des application ou domaines connexes est toujours problématique.

Le problème dont traite cet article est de tenter d'améliorer les premiers systèmes sans perdre leur capacité de large couverture dans le but de tendre vers une plus grande applicabilité des seconds. Nous effectuons cela par l'acquisition automatique d'une base de connaissances plus structurée que celles utilisées jusqu'à présent. Notre but est d'extraire ces connaissances de textes. En effet ceux-ci contiennent de nombreux exemples exploitables de l'usage des mots. Par contre, nous ne cherchons pas à modéliser un domaine choisi *a priori*. Au contraire, nous voulons traiter "la langue en général", à opposer à un langage de spécialité. Bien sur, il est reconnu qu'il nexiste pas de corpus de la langue générale. Chaque corpus est une partie du langage qui a sa propre spécificité. Nous utilisons deux corpus journalistiques : des dépêches de l'AFP (Agence France Presse) et des articles du journal "Le Monde". Ces deux corpus, bien qu'ayant un style journalistique assez fixe, couvrent des sujets très variés et, de ce fait, sont assez proche de ce qui pourrait être considéré comme un corpus de la langue générale.

Les processus automatiques qui extraient des connaissances depuis les textes appartiennent à des approches statistiques ou syntaxiques, et ce avec beaucoup de variations entre ces deux extrêmes. Les approches purement statistiques comme (Zernik, 1991) obtiennent souvent des groupes de mots qui caractérisent le sens d'un mot cible. D'autres systèmes intègrent des traitements syntaxiques divers et des interventions humaines, comme ARIOSTO (Basili *et al.*, 1993), ASIUM (Faure & Nedellec, 1998), STARTEX (Rousselot *et al.*, 1996) ou ZELLIG (Habert & Fabre, 1999). Ils obtiennent de façon générale des classes de mots ayant des sens similaires[1]. Ces systèmes sont conçus pour traiter des textes de langages spécialisés, dont le vocabulaire est bien défini et souvent peu ambigü. Appliqués à des textes non-spécialisés, ils conduisent à la création de classes excessivement larges, comme cela a été montré par (Fabre *et al.*, 1997) dans le cas de ZELLIG. Le système que nous avons développé, SVETLAN', contourne cette difficulté par l'utilisation d'une procédure de reconnaissance précise du contexte. Cela lui permet de construire des classes de mots ayant des sens similaires malgré la généralité du corpus. SVETLAN' détermine le sujet d'un texte en utilisant SEGAPSITH (Ferret & Grau, 1998), puis applique une méthode distributionnelle qui groupe les noms ayant un même rôle syntaxique par rapport à un verbe.

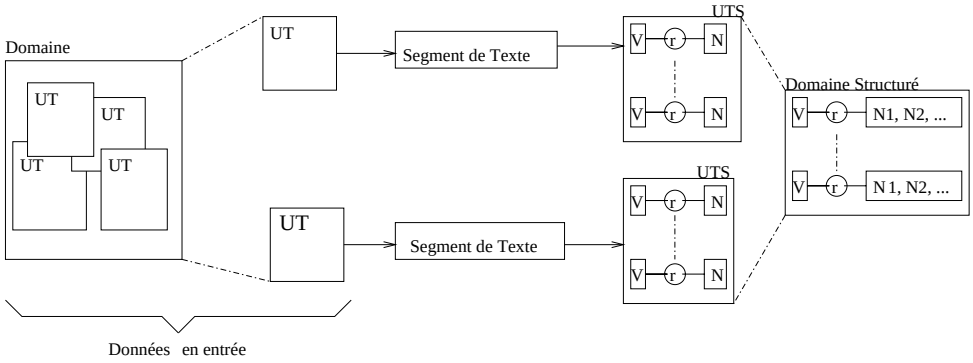
2. APERÇU DU SYSTÈME

Les données en entrée de SVETLAN' (voir fig. 1) sont des "domaines sémantiques" et les "unités thématiques" qui leur ont donné naissance. Les unités thématiques sont construites par un processus de segmentation thématique fondé sur un principe de cohésion lexicale et qui s'applique à des textes tels que des articles de journaux. Chaque unité thématique contient les mots pleins lemmatisés d'un segment de texte. Les domaines sont automatiquement appris par l'agrégation des unités thématiques similaires. Les domaines sont donc des ensembles de mots associés à un poids et qui représentent un sujet spécifique.

La première étape de SVETLAN' consiste à retrouver les segments des textes d'origine as-

[1]Dans la suite de cet article, nous utiliserons le terme "mots ayant des sens similaires", pour parler de mots qui se réfèrent à des concepts ayant un ancêtre immédiat ou quasi-immédiat dans une hypothétique ontologie. Une classe de tels mots sera dite "homogène".

FIG. 1 – Schéma de l'apprentissage d'un Domaine Structuré



sociés aux différentes unités thématiques, cela pour pouvoir les analyser syntaxiquement. Le système extrait alors des résultats de l'analyseur tous les triplets constitués d'un verbe, du nom tête d'un syntagme et de la fonction syntaxique de ce nom dans le but de produire des "unités thématiques structurées". Les unités thématiques structurées relatives à un même domaine sont agrégées de façon à apprendre un "domaine structuré". L'agrégation conduit au regroupement des noms jouant un même rôle syntaxique par rapport à un verbe afin de former des classes. Comme ces agrégations sont effectuées à l'intérieur d'unités thématiques appartenant à un même domaine, les classes sont fonction du contexte d'utilisation des mots, ce qui leur assure une grande homogénéité. Une étape de filtrage, fondée sur les poids des mots dans leurs domaines, permet au système d'éliminer certains noms des classes quand ils ne sont pas vraiment importants dans le domaine.

3. APPRENTISSAGE DES DOMAINES SÉMANTIQUES

Nous ne donnons ici qu'un bref aperçu du module d'apprentissage des domaines sémantiques, SEGAPSITH. Celui-ci est décrit plus précisément dans (Ferret & Grau, 1998). SEGAPSITH construit incrémentalement des représentations des sujets abordés dans les textes. Ces représentations sont constituées de mots pondérés, extraits des segments de textes délimités par SEGCOHLEX (Ferret, 1998). Il fonctionne sans classification *a priori* ou connaissances entrées manuellement. Les textes traités sont typiquement des articles de journaux du Monde ou de l'AFP. Ils sont prétraités de façon à ne conserver que les mots pleins lemmatisés (adjectifs, noms simples ou composés et verbes).

La segmentation d'un texte en sujets implémentée dans SEGCOHLEX repose sur un important réseau de collocations construit à partir de 24 mois du journal Le Monde. Dans ce réseau, un lien entre deux mots provient de la mesure de l'Information Mutuelle entre eux. À chaque position dans un texte, une valeur de cohésion est calculée grâce à ces liens. Les mots d'un segment de texte portant sur le même sujet sont fortement liés dans le réseau de coocurrences et induisent une valeur de cohésion élevée. Au contraire, une valeur de cohésion faible indique un changement de sujet. L'analyse de la courbe de cohésion permet de délimiter les segments. Les segments ayant des valeurs de cohésion suffisament élevées, appelés unités thématiques, sont conservés. La taille des unités thématiques est proche de celle d'un paragraphe. Les segments de textes, même quand ils relèvent du même sujet, développent bien souvent des points de vue différents et donc risquent de n'avoir que peu de mots en commun, ce qui limiterait leur similarité et risquerait d'empêcher leur agrégation. C'est pourquoi, pour enrichir la description donnée par un texte, le système ajoute aux unités thématiques les mots du réseau de cooccurrences qui

sont particulièrement liés aux mots du segment.

L'apprentissage d'une description complète d'un sujet consiste à agréger tous les points de vue successifs, c'est-à-dire les unités thématiques similaires, dans une seule unité thématique mémorisée, appelée domaine sémantique ou simplement domaine. Chaque agrégation d'une nouvelle unité thématique accroît les connaissances du système à propos d'un sujet en renforçant les mots récurrents et en ajoutant les nouveaux. Les poids sur les mots représentent l'importance de chacun par rapport au sujet et sont calculés d'après le nombre d'occurences de ces mots dans les unités thématiques. Cette méthode amène SEGAPSITH à apprendre des représentations de sujets spécifiques au contraire de (Lin, 1997), par exemple, dont la méthode construit des représentations de sujets généraux comme l'économie, le sport, etc.

FIG. 2 – Les mots les plus représentatifs d'un domaine relatif à la justice. Chaque mot est suivi de son nombre d'occurrences et de son poids dans le domaine.

```
juge d'instruction / 58 / 0,501, charge / 49 / 0,421,
emprisonner / 45 / 0,417, cour d'appel / 47 / 0,412, re-
cel / 42 / 0,397, supposer / 45 / 0,382, police judi-
ciaire / 42 / 0,381, fraude / 42 / 0,381
```

Nous avons appliqué le module d'apprentissage de SEGAPSITH sur un mois (mai 1994) de dépêches de l'AFP. La figure 2 montre un exemple de domaine portant sur la justice et regroupant 69 unités thématiques.

4. APPRENTISSAGE DE DOMAINES STRUCTURÉS

Comme dans (Faure & Nedellec, 1998), les verbes nous permettent de catégoriser les noms. Une classe est définie par les noms qui jouent le même rôle relativement à un même verbe. Dans le but d'apprendre des classes très homogènes, nous n'appliquons ce principe qu'aux mots appartenant à un même contexte, c'est-à-dire un domaine.

4.1. Analyse Syntaxique

Pour trouver les verbes et leurs arguments dans les textes, nous utilisons l'analyseur syntaxique Sylex (Constant, 1991)(Constant, 1995). La figure 3 montre une petite partie des résultats de Sylex pour une phrase. La première partie exhibe des informations lexico-syntaxiques pour les mots et ce pour quatre interprétations différentes ("taux 4") dues au fait que Sylex ne peut choisir, entre autres, entre les diverses interprétations du mot "laisse" : le verbe "laisser" et le nom "laisse". La seconde partie montre les liens syntaxiques trouvés par Sylex. Entre parenthèses on trouve des références aux mots des précédentes analyses. Ici, Sylex a trouvé quatre fois la même interprétation. Dans ce cas, on ne compte qu'une seule occurrence. Par contre, s'il trouve plusieurs fois la même relation entre le verbe et différents mots, par exemple plusieurs sujets possible, alors nous conservons les diverses possibilités. En effet, nous n'avons aucun moyen de choisir automatiquement la bonne. Nous faisons l'hypothèse raisonnable que les mauvaises interprétations auront beaucoup moins d'occurrences dans le corpus et de ce fait, seront filtrées durant la suite du processus.

Les résultats de Sylex sont très détaillés et difficiles à analyser avec, par exemple, un langage tel que Perl. De plus, nous n'avons pas besoin de toutes les informations extraites. En réalité, nous n'avons besoin que des verbes avec ses liens et les noms têtes des groupes nominaux

FIG. 3 – Un extrait du résultat de l'analyse d'une phrase par Sylex

```
                                                    ...<coupé>....
********** Partie 1   193-466   taux 4 **********
"L'état de santé critique du pilote autrichien Karl Wendlinger (Sauber-Mercedes), victime d'un grave accident
jeudi matin lors des premiers essais du Grand Prix de Monaco de Formule Un, laisse planer une menace sur le
déroulement de la course, dimanche en Principauté."
                                                    <InformationsLexico-Syntaxiques>
 193-195   (164) "L'" "le" [gs.1,avn,pdet.1] pdet : singulier elision dmaj
 195-208   (165) "état de santé" "état de santé" [gs.1,nom.1] nom : masculin singulier mot_compose locsw
                                                    ...<coupé>....
 382-388   (203) "laisse" "laisse" [gs.12,nom.1] nom : feminin singulier
 389-395   (204) "planer" "planer" [gs.13,verbe] verbe : infinitif
                                                    ...<coupé>....
 382-388   (211) "laisse" "laisser" [gs.13,verbe] verbe : singulier autoontif antiontif anontif present
indicatif subjonctif imperatif
 389-395   (212) "planer" "planer" [gs.14,verbe] verbe : infinitif
                                                    ...<coupé>....
                                                <Liens Syntaxiques>
'L'état de santé critique' (164) ->- cn head ->- 'du pilote autrichien' (170)
'planer' (204) ->- a2 head ->- 'une menace' (205)
                                                    ...<coupé>....
'planer' (153) ->- a2 head ->- 'une menace' (154)
                                                    ...<coupé>....
```

pointés par ces liens. C'est pourquoi, nous avons développé une grammaire formelle qui extrait les relations entre un verbe et ses arguments depuis ces analyses brutes. Les liens sont extraits dans le format suivant :

i # j verb # **token1** # *lemme1* # k rel #**token2** # *lemme2* # l

où i et j sont les frontières de la phrase qui contient le lien dans le corpus ; **token1** et *lemme1* sont respectivement la forme fléchie et le lemme du verbe ; rel est la relation syntaxique qui peut être "sujet", "COD" ou une préposition ("vers", "de", etc.) ; **token2** et *lemme2* sont la forme fléchie et le lemme du nom tête du groupe nominal pointé par la relation ; enfin, k et l sont les indices dans le corpus de **token1** et **token2** respectivement. La figure 4 montre quelques liens extraits des résutats de Sylex.

FIG. 4 – Exemples de liens extraits

token1	*lemme1*		rel	**token2**	*lemme2*
plane	*planer*		sujet	**menace**	*menace*
joue	*jouer*		COD	**coupe**	*coupe*
apprennons	*apprendre*		de	**sources**	*source*

Sylex, comme les autres analyseurs syntaxiques, éprouve des difficultés avec certaines constructions. Cela a pour conséquence d'introduire des erreurs qui peuvent causer des problèmes au reste du système. Un type d'erreur commun est la mauvaise interprétation de la voie passive qui cause l'interprétation du sujet comme un COD et inversement, le COD est vu comme un sujet. Une autre erreur commune est le fait de ne trouver aucun lien dans une phrase. C'est ce que nous nommons "silence". Nous verrons au paragraphe 5 que nous pouvons obtenir des résultats satisfaisants malgré ces problèmes grâce à la redondance nécessaire pour valider les liens dans les étapes suivantes du processus. Mais une autre conséquence de cette nécessité de redondance et du silence est que le système doit utiliser de grandes quantités de textes pour créer des classes d'une taille satisfaisante.

Les liens syntaxiques des phrases des textes sont groupés en fonction de leur appartenance à un segment de texte et donc à une unité thématique. Ainsi, nous définissons une "unité théma-

tique structurée" comme un ensemble de structures de la forme : <Verbe → relation syntaxique → Nom>, c'est à dire une relation syntaxique instantiée avec un verbe et un nom. Nous nous référerons par la suite à ces structures sous le nom de "relations syntaxiques instantiées". La mise en relation des liens extraits des résultats de Sylex et des mots contenus dans les domaines est possible car chaque domaine de la mémoire thématique mémorise quelles unités thématiques ont été utilisées pour le créer. Et de même, chaque unité thématique se souvient de la partie de texte de laquelle il provient.

4.2. Agrégation

Dans le but de construire des classes de mots ayant des sens très voisins, nous groupons les noms jouant le même rôle syntaxique par rapport à un verbe dans un domaine. Nous définissons alors un "domaine structuré" comme un ensemble de structures de la forme <Verbe → relation syntaxique → Nom_1, Nom_2, ..., Nom_n>, c'est-à-dire des relations syntaxiques instanciées agrégées.

Les unités thématiques structurées relatives au même domaine sont agrégées pour former les domaines structurés. Cette agrégation consiste à agréger leurs relations syntaxiques instanciées contenant le même verbe puis à ajouter les nouvelles relations syntaxiques instanciées, c'est-à-dire ajouter les nouveaux verbes avec leurs arguments composés d'une relation syntaxique et de la forme lemmatisée d'un nom.

La figure 5 montre l'agrégation d'un domaine structuré et de trois relations syntaxiques instanciées. Les éléments en gras représentent les données ajoutées ou mises à jour. Cet exemple montre tous les effets possibles de l'agrégation. Agréger une relation syntaxique instanciée dans un domaine structuré qui contient déjà le verbe de la relation syntaxique instanciée cause l'incrémentation du nombre d'occurrences du verbe. De même, les nombres d'occurrences des noms reliés au verbe par la même relation sont mis à jour et les nouvelles relations sont ajoutées au verbe avec leur nom associé. Enfin, une relation syntaxique instanciée avec un nouveau verbe est simplement ajoutée avec un nombre d'occurrences de 1.

FIG. 5 – Un exemple de l'agrégation de trois relations syntaxiques instanciées dans un domaine structuré

Domaine Structuré source			*Domaine Structuré résultat*		
jouer [4]	COD	coupe [3], match [1]	jouer **[5]**	COD	coupe [3], match **[2]**
	avec	balle [1]		avec	balle [1]
gagner [2]	sujet	joueur [1]		**sujet**	**champion [1]**
	COD	match [1]	gagner [2]	sujet	joueur [1]
3 *Relations Syntaxiques Agrégées sources*				COD	match [1]
jouer	sujet	champion	**perdre [1]**	**COD**	**championnat [1]**
	COD	match			
perdre	COD	championnat			

Les classes de noms dans les domaines structurés produits contiennent beaucoup de mots qui gênent leur homogénéité. Ces mots appartiennent souvent à des parties des différentes unités thématiques ayant produit le domaine structuré qui ne portent pas vraiment sur le sujet décrit. Cela correspond à une signification du verbe peu utilisée dans le contexte courant. Une autre possibilité est que la relation syntaxique instanciée résulte d'une erreur de Sylex. Comme ces mots ont un poids peu élevé dans les domaines correspondants, les données peuvent être filtrées :

tous les noms possédant un poids inférieur à un certain seuil sont supprimés de la classe. Par cette sélection, nous renforçons la composante contextuelle dans l'apprentissage des classes. La figure 6 montre, dans sa partie gauche, deux liens agrégés obtenus sans filtrage et les liens filtrés correspondants dans sa partie droite, les parties supprimées étant en gras. Le lien pour le verbe "établir" a été totalement supprimé tandis que le lien du verbe "répondre" avec la préposition "à" a été réduit par la suppression de "liste". Nous pouvons voir dans cet exemple que le filtrage est efficace : le verbe "établir" n'est pas particulièrement lié au domaine des "armes nucléaires" duquel cet exemple est tiré et l'usage de "répondre à une liste" a une très faible probabilité. Nous donnerons plus de détails concernant les effets du filtrage dans la section 5 sur les résultats.

FIG. 6 – Deux liens agrégés dans un domaine portant sur les "armes nucléaires"

établir	COD	base, zone	**établir**	**COD**	**base, zone**
répondre	à	document, question, liste	répondre	à	document, question**, liste**

Dans le principe, les opérations décrites ne sont pas très complexes. Les difficultés proviennent de la nécessité de travailler sur des données produites par des outils variés. De plus, pour des raisons de faisabilité et de performance, nous n'appliquons pas la chaîne de traitement texte par texte. La façon naturelle de voir le processus serait de :

- lire un texte ;
- en extraire les unités thématiques ;
- extraire les unités thématiques structurées correspondantes ;
- ajouter chaque unité thématique à son domaine ;
- et ajouter chaque unité thématique structurée au domaine structuré correspondant.

Dans les faits, chaque étape de calcul est effectuée sur le corpus dans son ensemble et les résultats sont alignés par la suite. Cela nous permet de gagner du temps de calcul car nous n'avons pas à exécuter chaque outil un grand nombre de fois. En revanche, nous avons à gérer des dictionaires et des indexs pour des fichiers et outils variés. De façon à simplifier cette gestion, nous avons développé un format de représentation des données commun à l'ensemble de nos outils.

5. RÉSULTATS

Les expériences préliminaires que nous avons menées avaient pour but de montrer que SVETLAN' peut apprendre des classes dont les mots ont des sens similaires dans le domaine. Pour obtenir de tels résultats, nous avons choisi d'exécuter le système sur un mois de dépêches de l'AFP (Agence France Presse), ce qui forme un corpus stylistiquement cohérent mais qui couvre des sujets variés en utilisant des verbes très polysémiques et peu spécifiques.

Ces dépêches sont constituées de 4 500 000 mots dans 48 000 phrases et 6 000 textes. L'analyse thématique renvoie 8 000 unités thématiques agrégées dans 2 000 domaines. De ces 48 000 phrases, Sylex extrait 117 000 différentes relations syntaxiques instanciées, dont 24 000 correspondent à des liens sujet, COD ou compléments circonstanciels introduits par une préposition. Ces liens sont intégrés dans 1 531 domaines structurés.

Après l'agrégation, mais avant le filtrage, le système obtient 431 liens agrégés avec deux arguments ou plus, équivalents à 431 classes de mots. Certaines, comme <*fabriquer*→ *COD*→ *bombe, arme*> sont satisfaisantes. Mais d'autres classes sont hétérogènes, comme <*rendre*→ COD→ *territoire, bande, contexte, synagogue*> (ici bande est un extrait de "Bande de Gaza"),

ou mélangent clairement des sens différents du verbe, comme <*quitter→ COD→ base, gouvernement*> qui mélange les sens de “quitter un lieu” et de “se retirer d’une institution”. Pour ces deux derniers cas, on peut voir l’intérêt qu’il y a à tenir compte du fait que les domaines contiennent des mots avec des poids variés qui représentent leur importance dans ce domaine. Plus le poids est élevé, plus le mot est important dans le domaine. Nous appliquons donc le filtre mentionné ci-dessus à nos classes et ne retenons que les mots dont le poids est supérieur à un seuil. La classe <*territoire, bande, contexte, synagogue*> est corrigée en <*territoire, bande*> et la classe <*base, gouvernement*> est supprimée.

Parmi les mauvaises classes, certaines sont dues à des erreurs de Sylex, comme <*décerner→ COD→ prix, acteur*> où “acteur” devrait être lié à “décerner” par la préposition “à”. Les autres sont dues à l’usage intensif de différents sens du verbe dans un même domaine, comme pour : <*diriger→ COD→ délégation, négociation*>. Ce type d’erreur est inhérent à la méthode que nous utilisons et devrait être traité par d’autre moyens. Il faut noter que la qualité des liens a été jugée manuellement par nous-mêmes.

Nous avons essayé deux seuils de filtrage : 0,05 et 0,1. La figure 7 montre les résultats pour les deux.

FIG. 7 – Résultats du filtrage pour deux seuils

Seuil	**Total**	**Bonnes**	**Erreurs Sylex**	**Autres Erreurs**
0,05	73	63% (46)	18% (13)	19% (14)
0,1	38	71% (27)	18% (7)	11% (4)

Après filtrage, un grand nombre de classes est supprimé mais celles qui restent sont la plupart du temps bien fondées. Un exemple d’une classe retenue est : <*blesser→Sujet →colon, soldat*>. Avec un seuil fixé à 0,1 plutôt que 0,05, nous ne retenons que 38 classes, mais nous avons un gain de précision de 8%. Si nous ignorons les erreurs dues à Sylex, la précision réelle de SVETLAN’ est dans la premier cas de 78% et dans le second de 87%. Ces valeurs élevées montrent l’intérêt qu’il y a à choisir un seuil adapté, même s’il nous faut à nouveau relever que l’évaluation actuelle faite d’après notre intuition devra être confirmée par une procédure plus formelle.

Dans le but de constater de façon évidente l’intérêt qu’il y a à construire des classes de mots en étant guidés par leur appartenance à un domaine, il est intéressant de regarder quelles classes auraient été obtenues par le mélange de tous les domaines, autrement dit : créer des classes hors-contexte. C’est pourquoi, nous avons appliqué les mêmes principes d’agrégation au même corpus mais sans prendre en compte les domaines. Ci-dessous, nous montrons une classe créée hors-contexte pour le verbe “remplacer”. Les mots en gras sont ceux qui appartienent à la même classe mais dans un domaine portant sur le nucléaire. Ce verbe est très général, quasiment tout peut être remplacé !

remplacer	COD	texte, constitution, pantalon, **combustible**, loi, dinar, **barre**, film, circulation, juge, saison, appareil, parlement, bataillon, police, président, traité

Ce groupe de mots mêle des sens très différents tandis que la classe formée par les mots en gras, beaucoup plus réduite, est meilleure car les mots qu’elle contient réfèrent à des concepts très similaires : une barre d’uranium est du combustible nucléaire. Un avantage de notre mé-

thode est donc d'obtenir des classes cohérentes pour des verbes très généraux et polysémiques. Les mots, utilisés en contexte, ne dénotent qu'une seule signification, généralement la même dans les diverses occurrences d'un même contexte. En construisant des classes en fonction du contexte d'apparition des mots, nous évitons de grouper les divers sens d'un mot dans une même classe. De tels résultats sont ainsi visibles dans les classes (loi, constitution) et (loi, article, disposition) où les mots très polysémiques constitution et article n'ont pas causé l'agrégation de mots gênant l'homogénéité de la classe.

SVETLAN', en collaboration avec SEGAPSITH, permet l'apprentissage automatique de domaines sémantiques structurés. Plutôt que de n'avoir que des ensembles de mots avec des poids pour décrire des domaines sémantiques, nous avons maintenant des ensembles de verbes reliés à des classes de mots par des liens syntaxiques. À l'inverse, nous pouvons aussi considérer ces classes comme des classes sémantiques, chacune d'entre elles étant en relation avec son contexte d'interprétation.

6. RELATIONS AVEC D'AUTRES TRAVAUX

Comme nous l'avons écrit dans l'introduction, beaucoup de travaux sont dédiés à la formation de classes de mots. Ces classes ont des statuts très variés. Elles peuvent, entre autres, contenir des mots appartenant au même champ sémantique ou des quasi-synonymes. Ci-dessous, nous donnons quelques détails sur trois systèmes qui permettent de mettre en lumière la spécificité de SVETLAN'.

IMToolset, de Ury Zernik (Zernik, 1991), extrait, pour un mot, un certain nombre de groupes de mots depuis des textes. Chacun de ces groupes reflète un sens différent du mot étudié. Cette extraction est faite par l'observation du contexte local du mot, les 10 mots qui l'entourent dans le texte. Ces signatures sont analysées statistiquement et classées. Le résultat est constitué d'un ensemble de groupes de mots qui sont similaires à nos domaines mais dédiés à la description du sens d'un mot unique. Ronan Pichon et Pascale Sébillot utilisent un principe semblable dans une expérience qu'ils décrivent dans (Pichon & Sébillot, 1999), mais ils utilisent en plus le thème des textes où ils rencontrent les occurrences des mots. Ce thème est détecté par une analyse statistique de la distribution des noms les plus courants du corpus. Il est utilisé pour focaliser les classes sur un sens spécifique du mot étudié. Comme dans IMToolset, et au contraire de SVETLAN', les classes obtenues ne contiennent pas des mots référant à des concepts proches mais elles définissent un concept.

Nous avons déjà décrit certaines caractéristiques de ASIUM par David Faure et Claire Nedellec, mais nous allons donner ici quelques détails supplémentaires. ASIUM apprend des cadres de sous-catégorisation de verbes et une ontologie en utilisant une analyse syntaxique (SYLEX, comme nous) et un algorithme de regroupement conceptuel. Le système utilise les résultats de l'analyseur syntaxique pour créer des classes de base constituées des mots apparaissant avec un même verbe et un même rôle syntaxique ou préposition, comme le fait SVETLAN'. Ces classes de bases sont ensuites regroupées hiérarchiquement pour créer une ontologie à l'aide d'un algorithme d'apprentissage coopératif. La principale différence avec SVETLAN' provient de cette approche coopérative : ASIUM dépend de l'expert qui doit valider, et éventuellement séparer en plusieurs morceaux, les classes proposées par l'algorithme. Cette approche est justifiée sur des textes techniques spécifiques, mais ASIUM, appliqué sur des textes tels que des dépêches de l'AFP risquerait de ne pas pouvoir extraire d'aussi bonnes classes de base que celles de SVETLAN'. De plus, comme dans ces textes chaque mot n'occure que peu de fois, la

distance utilisée n'aurait pas permis de regrouper un grand nombre de ces classes. Par contre, sur des textes techniques et avec la coopération d'un expert, ASIUM pourra obtenir de meilleurs résultats que les nôtres du point de vue de la couverture du domaine.

7. CONCLUSION

Le système que nous proposons, nommé SVETLAN', en conjonction avec SEGAPSITH et l'analyseur syntaxique Sylex, extrait des classes de mots depuis des textes bruts. Ces classes sont obtenues par le rassemblement de noms apparaissant avec la même relation syntaxique après un même verbe à condition d'être utilisés dans un même contexte. Ce contexte est donné par l'agrégation de segments de textes portant sur des sujets similaires. Les premières expériences que nous avons menées donnent de bons résultats. Cependant, elles confirment aussi que de très grandes quantités de données sont nécessaires pour extraire des connaissances lexicales substancielles par l'analyse des distributions syntaxiques. De plus, le très faible rappel de l'analyseur syntaxique et ses erreurs systématiques sur certaines constructions comme la voie passive, très courante dans le style journalistique de notre corpus, réduisent le nombre et la taille des classes formées. Pour résoudre ce problème, nous envisageons d'essayer d'utiliser conjointement un autre analyseur ou d'ajouter une étape de post-traitement à Sylex qui permet de détecter cette voie passive en utilisant les informations déjà présentes dans ses sorties. Ces adaptations et l'étude d'un corpus plus étendu nous permettront d'obtenir une meilleure couverture en nombre de domaines. Ainsi, nous serons capables de fournir des données sémantiques utiles à de nombreuses applications comme les systèmes de Recherche d'Information ou ceux de désambiguïsation des sens des mots.

Références

BASILI R., PAZIENZA M. T. & VELARDI P. (1993). What can be learned from raw texts ? *Machine Translation*, **8**, 147–173.

CONSTANT P. (1991). *Analyse Syntaxique par Couches.* PhD thesis, Ecole Nationale Supérieure des Télécommunications.

CONSTANT P. (1995). L'analyseur linguistique sylex. In *5ème école d'été du CNET*.

FABRE C., HABERT B. & LABBÉ D. (1997). La polysémie dans la langue générale et les langages spécialisés. *Sémiotiques*, (13), 15–30.

FAURE D. & NEDELLEC C. (1998). Asium, learning subcategorization frames and restrictions of selections. In Y. KODRATOFF, Ed., *10th European Conference on Machine Learning, ECML'98 - Workshop on text mining*.

FERRET O. (1998). How to thematically segment texts by using lexical cohesion ? In *ACL-COLING'98(student session)*, p. 1481–1483, Montreal, Canada.

FERRET O. & GRAU B. (1998). A thematic segmentation procedure for extracting semantic domains from texts. In *proceedings of the European Conference on Artificial Intelligence ECAI'98*, Brighton.

HABERT B. & FABRE C. (1999). Elementary dependency trees for identifying corpus-specific semantic classes. *Computer and the Humanities*, **33**(3), 207–219.

LIN C.-Y. (1997). *Robust Automatic Topic Identification*. PhD thesis, University of Southern California.

PICHON R. & SÉBILLOT P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In *TALN'99*, Cargese.

ROUSSELOT F., FRATH P. & OUESLATI R. (1996). Extracting concepts and relations from corpora. In *Corpus-Oriented Semantic Analysis ECAI'96 Workshop*, p. 74–78, Budapest, Hungary.

ZERNIK U. (1991). Train1 vs. train2 : Tagging word senses in corpus. In *RIAO'91*.