Structuration de domaines sémantiques

Gaël De Chalendar, Brigitte Grau

Structuration de domaines sémantiques

Gaël de Chalendar et Brigitte Grau

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

Résumé

Cet article présente la structuration de domaines sémantico-pragmatiques obtenue sur le français et sur l'anglais par l'application de SVETLAN'. Ce système regroupe les mots de sens proches selon une méthode distributionnelle. Les problèmes dus à la polysémie sont restreints par le regroupement de termes à l'intérieur de contextes thématiques spécialisés, ces contextes étant déterminés automatiquement lors de l'analyse du corpus traité. Nous présentons la méthode utilisée, ainsi qu'une comparaison des résultats obtenus sur les deux langues.

Mots clés : Structuration du lexique, connaissances sémantico-pragmatiques, approche distributionnelle

1 INTRODUCTION

Alors qu'il est évident que les systèmes qui sont amenés à gérer des documents requièrent des connaissances sémantiques et pragmatiques, on peut se demander quels types de connaissances sont nécessaires pour quels types d'applications. Une réponse courante en Ingénierie des Connaissances consiste à supposer que chaque application existe au sein d'un domaine bien déterminé qui manipule un ensemble de concepts et de relations entre ces concepts, une ontologie (Assadi, 1998). La représentation des connaissances directement liées à l'activité de l'organisation utilisant le système informatisé est donc suffisante dans cette optique.

Malgré tout, l'identification et le codage de ces connaissances est un problème difficile. Chaque métier, chaque spécialité générant des documents écrits, l'hypothèse de Z. Harris, largement reprise par le groupe TIA[1] (Terminologie et Intelligence Artificielle) en France, serait donc qu'il existe des sous-langages spécifiques à ces groupes et au sein desquels un *terme* ferait référence à un concept. Le but des applications proposées au sein du groupe TIA est alors d'aider à identifier et structurer les termes du domaine au sein d'un corpus constitué de textes collectés à cette occasion. Le cogniticien ou le spécialiste du domaine peuvent alors utiliser ces informations pour construire ou réviser l'ontologie du domaine.

Aussi précieuse que soit l'aide apportée par ces systèmes, il n'en reste pas moins que la constitution des ressources sémantiques sur lesquelles sont fondés les systèmes à base de connaissances (SBC) reste très coûteuse. Or, les métiers et les techniques évoluent ; les SBC doivent donc pouvoir être adaptés à l'évolution de leur domaine d'activité. De plus, dans certains cas, on peut considérer les changements non plus comme des évolutions mais comme l'apparition de nouveaux domaines. Dans ce cas, il peut être utile de construire des connaissances à propos de ces nouveaux domaines avant même que la nécessité de les prendre en compte au sein de l'organisation ne se soit fait sentir. Enfin, les systèmes sont de plus en plus ouverts et interconnectés. Les données strictement métier et les autres circulent au sein des mêmes flux, sans nécessairement de distinction formelle. Il faut donc pouvoir distinguer ces diverses informations. Il ne s'agit pas pour autant de simplement filtrer et éliminer les informations hors métier qui peuvent, comme dans le cas précédent, être les prémisses d'une future activité intéressante.

Pour résoudre les trois problèmes soulignés ci-dessus, il faut pouvoir traiter un flux de documents sans faire l'hypothèse du sous-langage. Une première réponse consiste à observer qu'il faut disposer de mots de sens proches. Dans ce cas, il est important de pouvoir évaluer la proximité de sens en contexte. Ainsi, si l'on est confronté aux phrases suivantes : *"… commencé à remplacer les barres de combustibles…"*, *"… commencé à remplacer le combustible du réacteur…"* et *"… remplacer les films et batteries des caméras… "*, dans un contexte de centrale nucléaire on voudra associer les mots *combustible* et *barre* et écarter le mot *film*.

Ce type de sémantique est particulièrement nécessaire pour réaliser des applications robustes, c'est-à-dire pouvant être utilisées quel que soit le domaine, par opposition à des applications dédiées à des domaines de spécialité. Les meilleurs exemples de ce type d'applications viennent de la recherche d'information. Les connaissances qui y sont utilisées peuvent provenir de bases destinées à couvrir la langue en général qui sont généralement conçues manuellement, telle WordNet. Dans ce cas, la difficulté réside dans la sélection des connaissances. Ainsi, le nom *care* possède 6 sens dans WordNet 1.6. Si l'on est intéressé par le domaine médical, on voudra sans doute éviter de sélectionner des documents possédant le mot *care*, utilisé dans sa 4ème acception, (« une raison de se sentir concerné, *a cause to feeling concerned*»), et choisir ceux où ce mot apparaît sous son sens premier : « le fait de prendre soin de quelqu'un ».

On peut alors se demander si, même en refusant le préalable du sous-langage, une classification universelle des concepts n'est pas une utopie, et cela principalement à cause

1 http://www.biomath.jussieu.fr/TIA/

de la polysémie des mots. Une solution peut être d'appréhender le vocabulaire général d'une langue en le modélisant par des classifications spécialisées qui se recouvrent partiellement. Nous nous sommes donc intéressés au problème de l'acquisition automatique d'une telle base. A cette fin, nous partons de trois hypothèses. Premièrement, une partie du sens des mots figure dans les textes. Deuxièmement, on peut en extraire cette connaissance automatiquement. Et enfin, cette acquisition n'est envisageable que si elle s'effectue dans le cadre de contextes très précis.

Les premiers travaux portant sur la modélisation de connaissances pragmatiques ont proposé des formalismes de représentation des situations de la vie courante, comme les scripts de Schank (Schank, 1982). Les connaissances y sont décrites manuellement et il est très difficile d'étendre cette approche à de nombreux domaines. CYC (Lenat, 1986) constitue un autre exemple de base de connaissances à visée universelle. CYC visait à construire une base de connaissances encyclopédique couvrant toute la langue. En réalité, comme pour WordNet, CYC doit être adapté manuellement pour chacune des applications l'utilisant.

Par ailleurs, des méthodes variées ont été appliquées avec un certain succès pour acquérir des connaissances sémantiques sur des domaines spécialisés : cooccurrences, approches statistiques (Zernik, 1991), approches distributionnelles appliquant les idées de Harris (Harris, 1968), techniques de classification (Agarwal, 1995), utilisation d'indices linguistiques (Roark & Charniak, 1998), etc. En France, ces diverses techniques ont été explorées par les membres du groupe TIA dont nous avons déjà parlé. On pensera par exemple à (Nazarenko & al., 1997) ou à (Assadi & Bourigault, 1998).

On peut alors se demander s'il ne serait pas possible d'appliquer ce type d'approche sur la « langue générale ». Aussi, notre proposition consiste à appliquer une approche distributionnelle sur des textes, et même des segments de texte, à condition qu'ils appartiennent à un même domaine. Cette approche conduit notre système, SVETLAN' (Chalendar & Grau, 2000), à former des classes de mots sémantiquement proches en contexte.

La théorie sous-tendant l'approche distributionnelle consiste à définir les verbes par un cadre de sous-catégorisation. Ce cadre spécifie que, par exemple, le sujet d'un verbe donné doit être une instance d'un concept particulier. L'ensemble des objets référencés par les noms, sujets du verbe, représentent ce concept en extension. Une description de cette extension est donc donnée par cet ensemble de noms, ensemble qui constitue une classe formée par notre système.

Les classes ainsi formées peuvent être utilisées en l'état comme représentant des concepts non étiquetés dans le domaine ou bien être soumises à un analyste (cogniticien ou spécialiste du domaine) pour formalisation et intégration dans un SBC (Séguéla & Aussenac, 1999).

2 ARCHITECTURE DE SVETLAN'

SVETLAN' (cf. figure 1) possède en entrée des domaines sémantiques avec les unités thématiques (UT) qui leur ont donné naissance. Les domaines sont des ensembles de mots pondérés, mots qui se sont révélés pertinents pour décrire un même thème. Ils sont automatiquement appris par ROSA (Ferret, 1998) qui agrège des unités thématiques similaires, constituées d'ensembles de mots. Ces unités thématiques sont construites par un processus automatique de segmentation de texte fondé sur la cohésion lexicale. Les textes traités sont des articles de presse. Ainsi, même si nous ne manipulons pas la « langue générale » dont l'existence même est discutable, nous traitons des corpus dont les domaines ne sont pas délimités *a priori* et sur lesquels l'hypothèse du sous-langage n'est pas applicable en l'état.

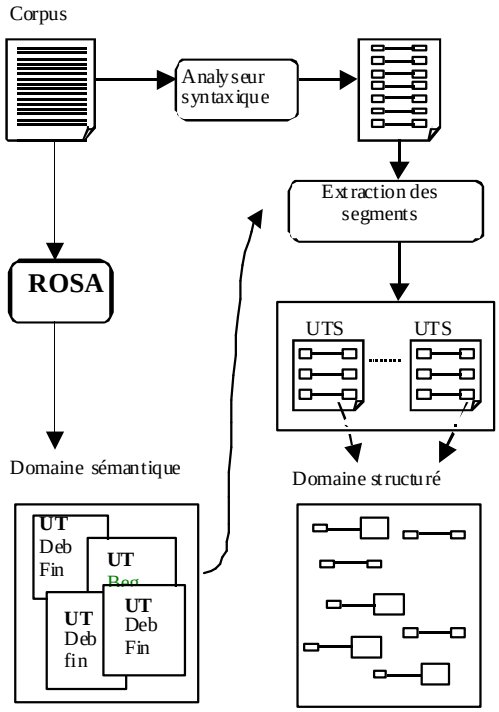

Figure 1. Apprentissage des domaines structurés

La première étape de SVETLAN' consiste à faire une analyse syntaxique du corpus (à l'aide d'un analyseur syntaxique externe) de manière à produire les unités thématiques structurées (UTS) correspondant à chaque unité thématique. Les UTSs sont constituées d'un ensemble de triplets – un verbe, un nom constituant la tête d'un argument et son rôle syntaxique – extraits des résultats de l'analyse syntaxique. Les UTSs relatives à un même domaine sont agrégées de manière à former un domaine structuré. L'agrégation consiste à regrouper les noms jouant un même rôle par rapport à un même verbe, ce regroupement définissant une classe sémantique. Comme ces agrégations sont effectuées entre UTSs relevant d'un même domaine, les classes formées sont dépendantes du contexte, ce qui leur assure une meilleure homogénéité. Une étape de filtrage fondée sur l'importance des noms dans le domaine permet d'éliminer beaucoup de noms non pertinents et de limiter le bruit.

3 LE SYSTEME ROSA

Nous ne donnerons ici qu'un rapide aperçu de ROSA, qui est composé de deux modules, SEGCOHLEX et SEGAPSITH. Pour plus de détails, consulter (Ferret & Grau, 1998). ROSA construit des représentations de thèmes de manière incrémentale. Ces représentations, les domaines sémantiques (appelés plus simplement *domaines*), sont constituées de mots lemmatisés possédant chacun un poids représentant leur importance pour le thème décrit. Les domaines sont formés de l'agrégation de segments de discours délimités par SEGCOHLEX (Ferret, 1998). Le processus complet ne nécessite aucune classification *a priori* des textes traités, ni aucune connaissance sémantique constituée manuellement. Les textes sont typiquement des articles de journaux venant du journal « Los Angeles Times » ou de dépêches de l'Agence France Presse. Ils sont pré-traités de manière à ne garder que leurs mots pleins sous forme lemmatisée (adjectifs, noms simples et composés et verbes).

La segmentation thématique réalisée par SEGCOHLEX est fondée sur l'utilisation d'un réseau de collocations construit à partir de 24 mois du journal « Le Monde ». Les liens dans ce réseau visent à coder des relations sémantico-pragmatiques entre les mots. Ils sont pondérés par la valeur de l'information mutuelle entre les deux mots liés[2]. Le processus de segmentation calcule une valeur de cohésion pour chaque mot du texte, à l'intérieur d'une fenêtre glissante. On suppose que des mots relatifs à un même thème sont fortement liés dans le réseau et entraînent une valeur de cohésion élevée lorsque la fenêtre est centrée sur chacun de ces mots. Quand la fenêtre est positionnée sur un ensemble de mots peu liés, on peut supposer qu'elle est à cheval sur plusieurs thèmes. Une faible valeur de cohésion permet alors de détecter un changement de thème.

Une fois les segments thématiques délimités dans les textes, seuls les segments de plus forte valeur de cohésion, appelés unités thématiques (UT), sont conservés pour l'apprentissage des domaines. Ce type de segmentation conduit à décomposer un texte en petites unités, dont la taille est environ celle d'un paragraphe. Notre but est ensuite d'agréger les unités thématiques qui sont descriptives d'un même sujet. Or, ces unités thématiques développent souvent des points de vue différents du même sujet, et de ce fait, ont peu de vocabulaire en commun. C'est pourquoi nous enrichissons la représentation d'une unité thématique par les mots du réseau de collocations qui sont les plus liés à l'ensemble des mots figurant dans les textes.

L'apprentissage d'une description complète d'un thème consiste à fusionner incrémentalement tous les points de vue, c'est-à-dire les UTs similaires, en une seule entité, décrivant ainsi un domaine sémantique (appelé, on le rappelle, *domaine*). La similarité d'une unité thématique avec un domaine mémorisé est donnée par la formule suivante :

$$sim(dom, UT) = \sqrt{ratio_d \cdot ratio_u}$$

avec les deux ratios calculés ainsi :

$$ratio = \sqrt{\frac{\sum_c poids(w_c)}{\sum_t poids(w_t)} \cdot \frac{\sum_c nbocc(w_c)}{\sum_t nbocc(w_t)}}$$

où l'indice c fait référence aux mots communs à l'UT (u) et au domaine (d) tandis que l'indice t désigne l'ensemble des mots constituant respectivement l'UT et le domaine. Les mots possédant un poids trop faible dans les domaines (poids inférieur à 0,1) ne sont pas retenus pour le calcul de similarité car ils sont globalement assimilés à du bruit. Le seuil en dessous duquel un nouveau domaine est créé est fixé à 0,25. Dans l'esprit du *tf*idf*, cette formule permet de privilégier les mots apparaissant souvent avec un poids élevé sans donner trop d'importance aux contenants (UTs et domaines) de grande taille.

mots	occ.	poids
juge d'instruction	58	0.501
garde_à_vue	50	0.442
bien social	46	0.428
inculpation	49	0.421
écrouer	45	0.417
chambre d'accusation	47	0.412
recel	42	0.397
présumer	45	0.382
police judiciaire	42	0.381
escroquerie	42	0.381

Figure 2. Les mots les plus représentatifs dans un domaine sur la justice

Chaque agrégation permet d'augmenter la connaissance du système sur un sujet particulier en renforçant le poids des mots récurrents et en ajoutant les mots nouveaux. Les poids sur les mots représentent l'importance de chaque mot pour le domaine et sont calculés à partir de leur significativité, de leur nombre d'occurrences et du nombre d'agrégations du domaine (cf. à nouveau (Ferret & Grau, 1998) pour plus de détails sur le calcul de ce poids). La figure 2 montre un exemple de domaine obtenu après l'agrégation de 69 UTs. Cette méthode, implémentée dans SEGAPSITH, conduit à la formation de descriptions spécifiques, contrairement à ce que l'on peut trouver dans des travaux comme ceux de Lin (Lin 1997), ou Bigi (Bigi 1998).

4 STRUCTURATION DES DOMAINES

Les domaines sont analogues aux classes formées par (Zernik, 1991). SVETLAN' délimite de petites classes à l'intérieur des domaines et les associe aux verbes qu'elles permettent ainsi de définir, comme on le trouve dans

[2] Les valeurs de l'Information Mutuelle entre un mot et tous les mots auxquels celui-ci est lié permettent de calculer pour le mot une valeur nommée *significativité* qui mesure, comme son nom l'indique, l'information qu'apporte ce mot de façon absolue. Pour plus de détails sur ce calcul, cf. (Ferret, 1998)

d'autres approches distributionnelles (Faure & Nedellec, 1998) (Pereira & al., 1993). Une classe est définie par l'ensemble des noms jouant un même rôle par rapport à un verbe, ces noms devant être reliés par des liens sémantiques forts. Ainsi, même si les noms ne dénotent pas toujours un même objet, les objets dénotés jouent un rôle similaire dans le cadre précis défini par un domaine.

4.1 Formation des unités thématiques structurées

Les textes sont analysés par un analyseur syntaxique de manière à relier les verbes et leurs compléments. Pour le français, nous avons utilisé Sylex (Constant 1991, 1995) et pour l'anglais le "link grammar" (Grinberg & al., 1995). Le système extrait tous les triplets trouvés par l'analyseur, constitués d'un verbe, d'une relation syntaxique et du mot tête du groupe nominal complément. Les relations retenues sont : sujet, compléments d'objet direct et indirect, et la relation nommée par la préposition pour un complément formé d'un groupe prépositionnel. Quand l'analyseur détecte une ambiguïté de rattachement, nous conservons toutes les analyses (pour le français seulement, l'analyseur anglais ne proposant qu'une analyse). Cela provoquera du bruit dans certaines classes, mais cela permettra aussi de former plus de classes par rapport à l'option de ne conserver aucun triplet dès qu'il y ambiguïté, ce qui est fréquent.

Après l'analyse de tous les textes du corpus, SVETLAN' regroupe les triplets selon les unités thématiques conservées. Nous définissons ainsi une unité thématique structurée par l'ensemble des triplets <*Verbe*à *relation syntaxique* à *Nom*> instanciés par des verbes et des noms particuliers. Nous appellerons ces triplets *relation syntaxique instanciée*.

4.2 Agrégation

Les unités thématiques structurées reliées à un même domaine sont agrégées pour former des domaines structurés. L'agrégation d'une UTS avec un domaine structuré suit le processus suivant :

- Agrégation des relations syntaxiques identiques liées au même verbe ; on construit ainsi un ensemble de mots argument d'un verbe par une même relation, formant ainsi une classe sémantique.
- Ajout de nouvelles relations syntaxiques consistant à ajouter une nouvelle relation et son mot argument à un verbe existant ou à ajouter un nouveau verbe avec son argument.

Les noms ne sont pas pondérés à l'intérieur des classes, ils conservent le poids qu'ils ont par rapport au domaine. Ainsi, l'ensemble des mots constituant une classe sont ceux qui jouent le même rôle par rapport à un verbe dans un contexte similaire, la similarité de contexte étant définie par une similarité lexicale calculée sur la totalité du domaine (cf. paragraphe3).

Il est important de noter que, à la différence de la plupart des approches distributionnelles, nous ne classifions pas à partir d'ensembles de contextes syntaxiques partagés (Habert & Fabre, 1999) ou de segments répétés (Rousselot & al., 1996). Au contraire, un seul verbe et une relation suffisent pour regrouper deux noms. Cela est dû au fait que seuls les mots ayant un poids suffisant dans les domaines sont intégrés aux classes. Or, le poids des mots est directement fonction de leur nombre d'occurrences dans le domaine. Ainsi, le rôle de la répétition dans les autres approches est pris en compte à une étape précédente dans la nôtre. Comme nous l'a fait remarquer un des relecteurs de cet article, ajouter un critère de répétition pourrait peut-être améliorer la précision de nos regroupements. Par contre, cela réduirait encore la couverture lexicale qui, comme nous le verrons ci-dessous, ne croit que lentement par rapport à la taille du corpus.

5 RESULTATS

Les classes sont formées compte tenu de deux niveaux de contextualisation dans l'emploi des mots : le contexte thématique et la pertinence locale dans un domaine, critère que nous avons ajouté pour contraindre davantage la formation des classes et éliminer les mots non pertinents du domaine.

Mettre	*sujet*	police, condition, chantage, table, ministre, président, parti, accord, ressource, régime, roi, village, déclaration, gens, gouvernement, preuve, décision, cas, film, réalisateur, banque, déficit, crime, opération, rencontre, sniper, ambition, blague, tractation
répondre	*à*	inquiétude, conflit, notification, double, document, question, liste, résolution, force, besoin, invitation, presse, réglementation, injonction, menace, proposition, appel, pénurie

Figure 3. Exemples de classes sans utilisation du contexte

Afin d'illustrer l'effet induit par la similarité thématique, la figure 3 montre une classe regroupant tous les noms complément d'objet direct trouvés pour les verbes *mettre* et *répondre* dans le corpus. On peut voir qu'il n'y a pas de proximité sémantique entre tous les noms de la classe.

Quand ces classes sont formées en contexte, cf. figure 4, on obtient alors des classes homogènes. Ainsi, même des verbes aussi généraux que *mettre* et *répondre* (il est possible d'appliquer ces verbes à beaucoup de choses), sont des critères pertinents de regroupement quand ils sont employés dans des textes portant sur un même domaine.

Domaine	**verbe**	***relation***	**classe**
Festival cinéma	Mettre	*sujet*	film, réalisateur
Irlande du Nord	répondre	*à*	document, question, liste

Figure 4. Effet du contexte thématique sur la constitution des classes

Restreindre l'application de la méthode par la proximité de contextes est nécessaire mais ce n'est pas suffisant pour

obtenir des classes homogènes. En effet, beaucoup de mots n'ayant pas de liens sémantiques entre eux appartiennent aux classes. Ces noms se trouvent souvent dans des phrases du segment de texte qui ne sont pas très liées au contexte, comme par exemple les dates d'un événement. Dans ce cas, nous disposons d'un critère pour les filtrer qui consiste à s'appuyer sur le poids des mots dans le domaine. Aussi, dès qu'un mot possède un poids inférieur à un seuil, il est supprimé de la classe. Par cette sélection, nous renforçons l'effet du contexte dans l'apprentissage des classes. La figure 5 montre deux classes obtenues sans filtrage à l'intérieur de leur domaine dans la partie haute, et leur contrepartie filtrée en bas. La classe constituée à partir des COD du verbe *établir* a disparu, alors que le mot *liste* a été supprimé de la classe constituée par les compléments introduits par *à*.

établir	*objet*	base, zone
répondre	*à*	document, question, liste
~~**établir**~~	~~*objet*~~	~~base, zone~~
répondre	*à*	document, question, ~~liste~~

Figure 5. Filtrage des classes selon l'importance des mots dans le domaine

La figure 6 montre quelques exemples de classes obtenues à partir d'un corpus de dépêches de presse en français (cf . section 6 pour plus de détails sur le corpus et des résultats quantitatifs).

Domaine	Verbe	Relation	Class
Guerre	Qualifier	Objet	président, dirigeant
Aide alimentaire	Réfugier	dans	pays, région
Tour de France	Franchir	Objet	étape, tour
Sport	Rencontrer	en	match, finale
Economie	Dégager	Objet	million, milliard
Festival cinéma	Raconter	Sujet	cinéaste, film
Conflit Croate	Reprendre	Objet	négociation, discussion
Economie	Réduire	Objet	excédent, déficit

Figure 6. Exemples de classes de noms

Même quand un verbe est polysémique, ce qui est le cas pour plusieurs verbes des exemples, la contrainte de domaine conduit à construire des classes pertinentes. On peut aussi constater que toutes les relations syntaxiques sont des critères appropriés pour regrouper des mots sémantiquement liés.

Une originalité de SVETLAN' consiste à produire des classes avec leur contexte de référence. Comme ce contexte est explicitement défini par un ensemble de mots, cela fournit des clés pour choisir entre une classe ou une autre pour un mot figurant dans un texte ou une phrase, et obtenir ainsi des mots voisins. Cependant, les résultats de SVETLAN' ne sont pas seulement la formation de classes, c'est aussi la structuration des domaines. Au lieu d'un ensemble de mots, ceux-ci sont maintenant décrits par des verbes associés à des classes précisant la nature de leurs arguments. Ce type de connaissance est une première étape vers une représentation des connaissances pragmatiques sous forme de schémas (Schank, 1982), (Chalendar & al., 2000). Un tel exemple, provenant du corpus de journaux américains, est donné figure 7.

Verbe	Relation	Classe
To accuse	Subject	Indictment, prosecutor
	By	Prosecutor, jury
To make	Subject	Prosecutor, indictment
	Direct Object	Jury, prosecutor
To show	Subject	Juror, defendant
	Direct Object	Jury, scheme
To tell	subject	Magistrate, informant
	Direct Object	Juror, jury
To give	Direct Object	Sentence, prosecutor, trial
	From	Sentence, prosecution
	To	Jury, defendant

Figure 7. Exemple d'un domaine sur la justice structuré par ses verbes

6 EXPERIMENTATIONS SUR DEUX LANGUES

Nous avons réalisé deux expérimentations sur des corpus de même nature, l'un en français et l'autre en anglais. Les deux expérimentations ont été réalisées en suivant les mêmes principes : segmentation thématique du corpus et création de la mémoire thématique (*i.e.* l'ensemble des domaines) ; analyse syntaxique et extraction des unités thématiques structurées ; création de la mémoire structurée (*i.e.* l'ensemble des domaines structurés) ; et enfin, évaluation du résultat. Pour celle-ci nous avons comptabilisé le nombre de bonnes classes, c'est-à-dire le nombre de classes contenant des mots partageant un lien sémantique direct. Ce jugement, intuitif, ne forme pas en soi une validation du système. Nous travaillons actuellement à l'intégration de l'usage des connaissances acquises dans plusieurs applications, entre autres du type de celles que nous avons esquissées en introduction. En ce qui concerne les classes jugées mauvaises, nous avons comptabilisé le nombre de ces classes dues à des erreurs d'analyse syntaxique.

Pour les deux expérimentations, nous n'avons gardé que les unités thématiques qui avaient permis de construire des domaines stables, soit les domaines formés à partir de 10 agrégations minimum.

Les deux corpus dont nous sommes partis étaient des textes bruts balisés SGML. Leur niveau de langue est élevé, les textes étant rédigés en style journalistique. Les deux portent sur des thèmes variés et nombreux. La taille du corpus français est 1,5 millions de mots, alors que le corpus anglais contient 7,3 millions de mots.

Les deux sous-sections suivantes détaillent les résultats obtenus sur les deux corpus et la troisième en fait une synthèse.

6.1 Français

Le corpus français est composé de 3 mois de dépêches de l'Agence France Presse (AFP). La mémoire thématique créée comporte 87 domaines stables. La figure 8 montre quelques exemples de classes contenues dans des domaines structurés.

Domaine	Verbe	Relation	Classe
Loi italienne	Dénoncer	Sujet	Loi, disposition
Loi italienne	Abroger	Objet	Loi, article, disposition
Israël Palestine	Entrer	Dans	Bande, territoire

Figure 8. Exemples de classes dans des domaines structurés pour le français

On peut constater que les classes d'un domaine ne forment pas une partition. Par exemple les mots *loi* et *disposition* du domaine portant sur la loi italienne appartiennent à 2 classes.

Dans cette expérimentation, nous avons testé deux seuils de filtrage : 0,05 et 0,1. La table 1 montre que, avec un seuil à 0,05, 63% de bonnes classes sont formées, alors que ce pourcentage passe à 71 % avec un seuil à 0,1. Bien entendu, l'utilisation d'un seuil plus élevé réduit le nombre total de classes construites, mais c'est la contrepartie inévitable d'une plus grande précision. Dans l'expérimentation suivante, nous avons choisi de privilégier la précision plutôt que le rappel, aussi nous avons conservé le seuil de filtrage à 0,1.

Seuil	Nombre de classes	Bonne	Erreurs d'analyse	Autres erreurs
0.05	73	63 %	18 %	19 %
0.1	38	71 %	18 %	11 %

Table 1. Résultats en français avec 2 seuils de filtrage

La table 1 montre que 18% des classes jugées mauvaises sont dues à des erreurs d'analyse. Cependant nous ne savons pas combien de ces classes seraient correctes si les erreurs étaient corrigées. De toute façon, certaines classes sont intrinsèquement incorrectes, et illustrent les limites de la méthode ; les domaines ne peuvent résoudre tous les problèmes de polysémie des verbes, aussi des mots apparaissant avec le même rôle pour des sens différents d'un même verbe seront groupés. Cependant, dans l'état actuel, seules 11% des classes sont dans ce cas. Ce pourcentage est assez bas et reflète l'efficacité de l'étape de filtrage.

6.2 Anglais

Le corpus anglais est constitué de 3 mois d'articles du « Los Angeles Times ». La mémoire thématique créée possède 138 domaines stables. La figure 9 montre quelques exemples des classes formées dans un domaine portant sur la médecine.

Verbe	Rel[on]	Classe
To take	Under	Home, residence
To meet	Object	Care, physician
To carry	Object	Virus, antibody
To get	Subject	Treatment, care

Figure 9. Exemples de classes pour l'anglais

Ces exemples font apparaître deux classes contenant le mot *care*. Elles instancient deux relations sémantiques différentes : dans la classe <care, treatment> il existe une relation *instrument* entre les termes (un traitement est un moyen de soigner un malade), et dans la classe <care, physician>, le lien est de type *agent* (le médecin prend soin de ses patients). D'autres classes de ce domaine contiennent le mot *care*, avec les 2 sens relevés. Les classes ne partitionnent donc pas les mots comme nous l'avons vu pour le français, mais elles ne créent pas non plus une partition des concepts. Dans une étape ultérieure, nous projetons d'étudier la possibilité de fusionner des classes proches.

Nombre de classes	Bonne	Erreurs de l'analyseur	Autres erreurs
149	58 %	7 %	35 %

Table 2. Résultats sur l'anglais avec un seuil fixé à 0,1

6.3 Comparaison des résultats des deux expérimentations

La table 3 fait apparaître des points de comparaison entre les résultats obtenus sur le français et sur l'anglais. Nous apprenons principalement deux choses. Premièrement, il est évident qu'avec un corpus de plus grande taille nous obtenons plus de domaines et donc plus de classes. Cependant, l'accroissement du nombre de domaines et de classes n'est pas linéaire par rapport à la taille du corpus, le ralentissement étant même plus important pour les domaines. Cela n'est pas surprenant dans la mesure où les mêmes sujets sont traités régulièrement dans les articles et viennent donc conforter des domaines existants. Les nouveaux domaines sont donc de plus en plus rares au fur et à mesure de la progression dans le traitement du corpus. A l'inverse, lorsque de nouvelles unités thématiques sont ajoutées à des domaines, de nouveaux mots apparaissent régulièrement, des verbes comme des noms, et de nouvelles classes se créent ou se complètent dans des domaines existants. On peut s'attendre à ce que le taux d'accroissement du nombre de classes diminue aussi, mais avec de plus grands corpus encore.

Deuxièmement, le taux de classes jugées correctes est plus faible en anglais qu'en français. Nous n'avons pas pour le moment d'explication complète pour ce phénomène. Cela

ne semble pas dû à l'analyseur syntaxique. En effet, d'après une première analyse, la précision du Link Parser dans l'identification des liens en anglais semble meilleure que celle de Sylex en français. Trois hypothèses sont en lisse pour justifier cette différence :

- la différence de langue : la sélection des arguments serait moins importante en anglais qu'en français ;
- notre maîtrise plus faible de l'anglais qui nous ferait refuser des classes qu'une personne dont l'américain est la langue maternelle aurait peut-être acceptées ;
- et enfin, une précision inférieure dans la constitution des domaines dont certains seraient devenus trop généraux.

Le premier point est appuyé en partie par la constatation de la forte amélioration des résultats en ignorant certains verbes (to be et to have). Nous avons aussi eu l'impression que les verbes très spécifiques au domaine étaient accompagnés de classes bien meilleures.

	Français	Anglais
Taille du corpus	1.5 M	7.3 M
Nombre de domaines	87	138
Nombre de classes	38	149
Classes correctes	71 %	58 %

Table 3. Synthèse des 2 expérimentations

Le second reçoit une certaine confirmation avec l'exemple suivant : nous avons jugé que la classe (mortgage, bondholder, dividend), dans un domaine sur la banque, n'était pas bonne. Or, ces trois mots font référence à des produits financiers. Peut-être une personne dont l'anglais est la langue maternelle aurait-elle accepté cette classe.

Ces résultats et ces doutes montrent la difficulté qu'il y a à faire reposer l'évaluation sur notre propre jugement. Cela montre qu'il est vraiment important de trouver un moyen extérieur pour juger les classes, comme une application dont on peut mesurer l'amélioration des performances avec ou sans les classes.

Par la suite, nous prévoyons d'expérimenter des seuils de stabilité plus bas pour la sélection des domaines, de manière à augmenter le nombre de classes apprises. Dans une telle configuration, il pourrait être utile d'utiliser comme confirmation des classes un critère de contextes partagés à la manière des autres approches distributionnelles.

7 TRAVAUX CONNEXES

Beaucoup de travaux portent sur la formation de classes de mots. Ces classes ont des statuts différents. Elles peuvent contenir des mots appartenant au même champ sémantique ou des presque synonymes.

WordNet (Fellbaum, 1998) est une base de connaissances élaborée par des lexicographes qui a pour but de représenter tout le lexique d'une langue. Les noms sont liés à des Synsets, qui représentent les concepts et regroupent les ensembles de synonymes. Ces Synsets sont hiérarchisés. La couverture de WordNet est très large, mais cette qualité est dans un sens son principal défaut. En effet, cette volonté de tout décrire rend l'utilisation de WordNet assez problématique sans une adaptation manuelle.

Les systèmes automatiques eux appliquent différents critères pour regrouper les noms, mais ils font tous usage de la notion de contexte et de mesures de proximité sémantique. IMToolset, par Uri Zernik (Zernik, 1991), regroupe les contextes locaux du mot étudié, contexte défini par les 10 mots qui l'entourent dans les textes. La proximité entre mots est évaluée par la mesure de l'information mutuelle, comme ce que nous faisons lors de la segmentation des textes. Le résultat d'IMToolset est constitué de groupes de mots similaires à nos domaines, mais plus focalisés sur le sens d'un seul mot.

Faure et Nedellec (Faure & Nedellec, 1998), avec Asium, ou Lin (Lin, 1998) appliquent une approche distributionnelle. Asium est dédié à la construction d'une ontologie d'un domaine spécialisé, aussi il n'y a aucune restriction contextuelle lors du regroupement des noms, la restriction étant effectuée lors de la constitution du corpus à traiter. Les classes de base y sont regroupées pour créer l'ontologie par le biais d'un algorithme d'apprentissage coopératif. Cette partie manuelle est une étape analogue à notre filtrage automatique. Lin, quant à lui, n'applique pas de sélection contextuelle sur les mots avant de les regrouper ; il définit une mesure de similarité entre mots de la même classe pour les ordonner en fonction de leur degré de similarité. Ce type de méthode conduit aussi à construire de larges classes, plus analogues à nos domaines sémantiques.

8 CONCLUSION

Le système que nous avons développé, en conjonction avec ROSA et un analyseur syntaxique, extrait des classes de noms à partir de textes bruts. Ces classes sont créées par le regroupement des noms qui jouent le même rôle par rapport au même verbe dans un contexte similaire. Ce contexte est défini par l'agrégation de segments de textes portant sur un même sujet. Nous avons réalisé deux expérimentations afin de montrer les résultats obtenus par notre système, l'une sur des articles de journaux en français, l'autre sur des articles en anglais. Nous obtenons de bons résultats, mais ces expérimentations confirment la nécessité de traiter un très gros volume de textes pour extraire suffisamment de classes utiles pour une application en grandeur réelle. Notre but est d'être à terme en mesure de fournir une base de connaissances lexicales utilisable dans de nombreuses applications de recherche d'information ou de désambiguïsation sémantique. Dans cette perspective, nous développons actuellement un module d'expansion de requêtes pour un système de question-réponse (Ferret et al. 2001) pour l'anglais.

Cela ne nous fait pas oublier notre but à long terme, c'est-à-dire la réalisation d'un système incrémental d'acquisition de connaissances pragmatiques pour représenter des situations prototypiques du monde réel et ainsi raisonner efficacement à leur propos (Chalendar et al., 2000). C'est la difficulté de coder et de maintenir manuellement les connaissances sémantiques sur lesquelles repose un tel système qui nous a amené à tenter de les acquérir automatiquement. Le

fait d'identifier les classes acquises par Svetlan' à des concepts nous permettra à terme de les utiliser au sein de représentations des textes sous forme de graphes conceptuels (Sowa, 1984).

REFERENCES

R. Agarwal, Semantic feature extraction from technical texts with limited human intervention, PhD thesis, Mississippi State University, 1995.

H. Assadi. Construction d'ontologies à partir de textes techniques. Thèse de l'université Paris 6, 19/10/1998.

H. Assadi and D. Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. 10ème congrès reconnaissance des formes et intelligence artificielle, rfia-96, Rennes, France, 1/1996.

B. Bigi, R. de Mori, M. El-Bèze and T. Spriet. Detecting topic shifts using a cache memory, In *Proceedings of 5th International Conference on Spoken Language Processing*, Sydney, Australia, (1998).

P. Constant, Analyse Syntaxique Par Couches. Ph.D thesis, École Nationale Supérieure des Télécommunications, April, 1991.

P. Constant, L'analyseur linguistique SYLEX. 5ème ecole d'été du CENT, 1995.

G. de Chalendar and B. Grau. Svetlan' or how to classify words using their context. In R. Dieng and O. Corby, editors, Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000, volume 1937 of Lectures notes in artificial intelligence. Springer, 2000.

G. de Chalendar, B. Grau and O. Ferret. A cost-bounded algorithm to control events generalization'2000. In B. Ganter and G. W. Mineau, editors, Conceptual Structures : Logic, Linguistic, and Computational Issues. 8th International Conference on Conceptual Structures ICCS, volume 1867 of Lectures Notes in Computer Science. Springer, Darmstadt, Allemagne, 8/2000

D. Faure and C. Nedellec, ASIUM, Learning subcategorization frames and restrictions of selection. In Y. Kodratoff ed., proceedings of 10th ECML – Workshop on text mining, 1998

C. Fellbaum, WordNet: an electronic lexical database, The MIT Press, 1998.

O. Ferret, How to thematically segment texts by using lexical cohesion? Proceedings of ACL-COLING'98 (student session), pp. 1481-1483, Montreal, Canada, 1998.

O. Ferret and B. Grau, A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts, Proceedings of ECAI'98, Brighton, 1998.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, 2001, Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse, TALN 2001, Tours

G. Greffenstette, Explorations in automatic thesaurus discovery, Kluwer Academic Pub., Boston, 1994.

D. Grinberg, J. Lafferty and D. Sleator, A robust parsing algorithm for link grammars, Carnegie Mellon University Computer Science technical report CMU-CS-95-125, and Proceedings of the Fourth International Workshop on Parsing Technologies, Prague, September, 1995.

B. Habert and C. Fabre. Elementary dependency trees for identifying corpus-specific semantic classes. Computer and the Humanities, 33(3):207-219, 1999.

Z. Harris, Mathematical Structures of Language, Wiley, New York, 1968.

C.-Y. Lin, Robust Automated Topic Identification, Doctoral Dissertation, University of Southern California, 1997.

D. Lin. Automatic retrieval and clustering of similar words. In Proceedings of COLINGACL '98, pages 768--774, Montreal, Canada, August 1998.

A. Nazarenko, P. Zweigenbaum, J. Bouaud and B. Habert. Corpus-based identification and refinement of semantic classes. Journal of the American Medical Informatics Association, 4(suppl):585--589, 1997.

F. Pereira, N. Tishby and L. Lee, Distributional clustering of english words, Proceedings of ACL'93, 1993.

B. Roark and E. Charniak, Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. Proceedings of COLING-ACL'98, pp. 1110-1116, 1998.

F. Rousselot, P. Frath and R. Oueslati. Extracting concepts and relations from corpora. Corpus-oriented semantic analysis ecai'96 workshop, pages 74-78. Budapest, Hungary, 1996.

R. C. Schank, Dynamic memory. A theory of reminding and learning in computers and people, 1982. Cambridge University Press, 1982.

P. Séguéla and N. Aussenac. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In R. Teulier, editor, Actes de IC'99, 1999.

J. F. Sowa, Conceptual structures: information processing in mind and machine, London, 1984. Addison-Wesley, London, 1984.

U. Zernik, TRAIN1 vs. TRAIN2: Tagging Word Senses in Corpus, RIAO'91, 1991