Corpus et Web : deux alliés pour la construction de l'enrichissement automatique de classes conceptuelles

Nicolas Béchet, Mathieu Roche, Jacques Chauché

TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009

Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Després	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

SESSION 3

Corpus et Web : deux alliés pour la construction et l'enrichissement automatique de classes conceptuelles

Nicolas Béchet, Mathieu Roche, Jacques Chauché

Résumé : Cet article présente des méthodes permettant de construire et d'enrichir des classes conceptuelles. Ces classes sont construites en utilisant les informations syntaxico-sémantiques issues d'un corpus. La première méthode d'enrichissement se fonde sur l'utilisation du corpus et d'objets de verbes dits complémentaires. Nous présentons alors un protocole d'évaluation automatique permettant de valider la pertinence de ces objets réduisant ainsi le travail de l'expert. La seconde méthode permet d'enrichir les concepts avec des termes plus généraux en s'appuyant sur les ressources du Web.

Mots-clés : terminologie, classes conceptuelles, enrichissement, Web

1. Introduction

La terminologie est un domaine ayant de nombreuses applications en TAL (Traitement Automatique des Langues). Elle peut-être vue comme l'étude des mots techniques propres à un domaine et de leurs significations. Nous distinguons deux types d'études terminologiques : l'approche *sémasiologique* et l'approche *onomasiologique*. La première s'intéresse à l'étude des significations partant du *mot* pour en étudier le sens. La seconde, propose de partir du *concept*.

Un concept peut être défini comme la *représentation mentale d'une chose ou d'un objet* (Desrosiers-Sabbath 1984). Nous proposons de définir un concept comme un *ensemble de connaissances partageant des caractéristiques sémantiques communes*. Nous utilisons dans nos travaux l'approche *sémasiologique* en apportant un début de réponse aux problèmes générés par ce type d'approches. La terminologie ainsi extraite est en effet très dépendante du corpus. Cela implique alors qu'une terminologie répondant à des besoins spécifiques est vouée à une faible réutilisabilité (Roche 2005).

Nous proposons dans nos travaux de construire, dans un premier temps, des classes conceptuelles spécifiques en nous appuyant sur les données issues de corpus. Pour cette tâche, nous construisons des classes conceptuelles en étudiant la dépendance syntaxique des termes d'un corpus (**section 2**). Pour cela, nous nous fondons sur les relations syntaxiques *Verbe-Objets*. De telles relations sont assez représentatives d'un domaine. Par exemple dans un texte du domaine de l'informatique, le verbe *charger* prendra comme objet un nom qui appartient à la classe conceptuelle *logiciels* (L'Homme 1998).

Une extraction terminologique ainsi effectuée s'apparente à une analyse distributionnelle "à la Harris" (Harris 1968). Il existe de nombreux travaux effectuant une telle analyse pour l'acquisition de ressources terminologiques ou ontologiques à partir de textes. Citons par exemple (Bourigault et Lame 2002) dans le domaine du droit et (Nazarenko *et al.* 2001) dans le domaine médicale.

Une fois les classes conceptuelles constituées, nous proposons de les enrichir avec deux méthodes distinctes. Une première utilise les ressources syntaxico-sémantiques du corpus afin de proposer de nouveaux termes (**section 3**). Ces termes sont alors ordonnés automatiquement avant d'être soumis à un expert. L'autre méthode d'enrichissement se fonde sur l'utilisation du Web afin de générer de nouvelles ressources terminologiques qui sont moins dépendantes du corpus (**section 4**).

2. Méthode pour la construction de classes conceptuelles

Cette section présente une méthode de construction de classes conceptuelles fondée sur l'utilisation d'informations syntaxiques d'un corpus : les relations syntaxiques de type Verbe-Objet. Ainsi un concept est formé avec les objets des verbes d'un corpus qui ont été jugés "proches". Nous détaillons ci-dessous la construction de tels concepts.

La première étape consiste à extraire les relations syntaxiques d'un corpus. Nous utilisons pour cela l'analyseur morpho-syntaxique *Sygfran* (Chauché 1984). Ainsi, avec cette analyse syntaxique, nous avons par exemple extrait de la phrase : "*Thierry Dusautoir brandissant le drapeau tricolore sur la pelouse de Cardiff après la victoire*". la relation syntaxique : "*verbe : brandir, objet : drapeau*"

La seconde étape de notre méthode consiste à rassembler les objets des verbes jugés proches. Pour cela, nous émettons l'hypothèse que des verbes peuvent être considérés comme sémantiquement proches s'ils partagent un nombre important d'objets en communs. Ainsi, nous utilisons le score d'Asium, se fondant sur cette hypothèse (Faure 2000) afin de mesurer la proximité des verbes de notre corpus. Le principe d'Asium est similaire à (Bourigault *et al.* 2002), se fondant sur une analyse distributionnelle.

Alors, nous regroupons tous les objets dont les verbes ont été jugés proches. Nous illustrons dans la figure 1 un exemple de verbes sémantiquement proches : *agiter* et *brandir*. Les **objets communs** à ces deux verbes constituent le concept : *Objets symboliques*. Notons que le concept formé ici est d'une thématique peu fréquente montrant l'intérêt d'une telle construction de classes conceptuelles spécifiques.

La troisième étape de notre méthode propose de distinguer deux types d'objets pour construire une classe conceptuelle. Les objets **communs** aux deux verbes de la figure 1 *drapeau* et *fleur* sont des instances de qualité du concept *Objets symboliques* mais qu'en est-il des deux autres objets *pancarte* et *rasoir* ? Ces objets, qui ne sont objets que d'un seul verbe sont appelés objets **complémentaires.** Ils permettent d'induire de l'information. En effet, la relations syntaxique *brandir pancarte* n'est pas originalement présente dans le corpus et est induite de la relation *agiter pancarte* et du verbe *brandir*.

La méthode définie dans cet article considère « nativement » un objet **commun** comme instance d'un concept. Néanmoins, les objets **complémentaires** nécessitent d'être validés. En effet, l'objet *pancarte* est une instance pertinente du concept *Objets symboliques*, contrairement à *rasoir*.

Nous présentons dans la section suivante des solutions de validation des objets complémentaires ainsi que des protocoles d'évaluation afin d'estimer la qualité de ces validations.

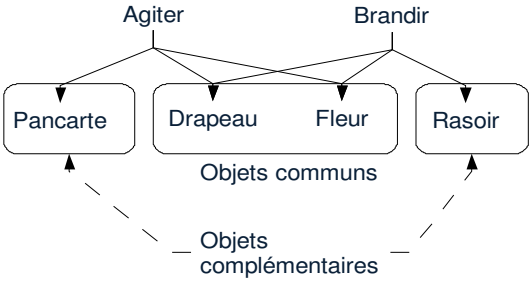

Figure 1. Objets communs et complémentaires des verbes "agiter" et "brandir"

3. Traitement des objets complémentaires

3.1. Les mesures de validation

Cette section présente brièvement les approches permettant d'ordonner en termes de qualité les relations syntaxiques induites. Elles sont présentées en détail dans (Béchet *et al.* 2009a). Nous émettons l'hypothèse qu'une relation syntaxique induite cohérente signifie que l'objet complémentaire la composant est une instance pertinente du concept formé par les objets communs des deux verbes.

Vecteurs sémantiques (VS)

Une première validation consiste à mesurer la pertinence de l'association d'un verbe avec son objet complémentaire. Ainsi, nous allons valider la proximité sémantique entre le verbe et l'objet d'une relation induite avec l'autre verbe et le même objet. Sur l'exemple de la figure 1, il s'agit de mesurer la proximité sémantique des relations *agiter pancarte* (relation originale) et *brandir pancarte* (relation induite). Les relations syntaxiques sont représentées par des **vecteurs sémantiques**.

Un tel vecteur est construit en représentant un ou plusieurs termes en le(s) projetant sur un espace de dimension finie de 873 concepts. Ces concepts sont organisés comme une ontologie de concepts définis dans le thésaurus (Larousse 1992). Chaque mot est indexé par un ou plusieurs éléments. Par exemple, "brandir" est associé à "agitation" et "drapeau" à "paix, armée, funérailles, signe, cirque". La représentation d'une relation syntaxique résulte d'une combinaison linéaire entre la représentation du verbe et de l'objet, dont les coefficients prennent en compte la structure syntaxique

(Chauché 1990). La proximité des vecteurs est finalement mesurée par un calcul de cosinus entres les deux vecteurs sémantiques (vecteurs propres aux relations originales et aux relations induites).

Validation Web (VW)

Une autre validation utilise le **Web** afin de mesurer la dépendance entre verbe et objet d'une relation syntaxique induite. Elle s'inspire des travaux de (Turney 2001). Ainsi, une requête est soumise à un moteur de recherche, sous forme de chaîne de caractère. Le nombre de pages de résultats retournés constitue la mesure de dépendance. De plus, nous appliquons diverses mesures statistiques telles que l'*Information Mutuelle* (Church et Hanks 1990) ou le *coefficient de Dice* (Smadja *et al.* 1996). Ceci permet de pondérer l'importance de la relation syntaxique en fonction du verbe et de l'objet dans les résultats obtenus. Nous utiliserons dans cet article uniquement l'Information Mutuelle suite aux expérimentations effectuées. L'Information Mutuelle adaptée à notre approche est définie comme suit :

$$IM(v,o) = \frac{nb(v,o)}{nb(v)nb(o)}$$

Avec *nb(v), nb(o) et nb(v,o)* étant respectivement le nombre de pages retourné par le moteur de recherche lors de la soumission du verbe *v,* de l'objet *o* et de la relation syntaxique *vo*. La particularité de la validation Web est qu'elle utilise des ressources extérieures afin de mesurer la cohérence d'un candidat à un concept. Ainsi, à partir d'informations spécifiques propres à un corpus, nous obtenons une évaluation plus globale de la pertinence des concepts.

Les combinaisons

Nous proposons de combiner les deux approches présentées précédemment avec deux types de combinaisons.

- La première combinaison introduit un paramètre $k \in [0,1]$ pour donner un poids supplémentaire à l'une ou l'autre des approches. Pour une relation syntaxique *r*, nous combinons les approches avec le calcul suivant :

$$combine\_score_r = k \times VS + (1-k) \times VW$$

Avec VS et VW étant respectivement les scores obtenus pour la relation syntaxique *r* avec l'approche des vecteurs sémantique et la validation Web.

- La seconde combinaison consiste à classer la totalité des relations syntaxiques par l'approche VS. Puis, les *n* premières relations syntaxiques sont de nouveau ordonnées avec l'autre approche VW.

L'ensemble des approches présentées dans cette section vont fournir une liste de candidats aux différents concepts ordonnés par valeurs décroissantes des différentes mesures (VS, VW, combinaison 1, combinaison 2). Nous ne présenterons dans cet article que les résultats de la validation Web et de la seconde combinaison, mesures obtenant les meilleures performances (Béchet *et al.* 2009b). Les autres résultats sont proposés en annexe.

La section suivante présente différents protocoles d'évaluation afin de mesurer la qualité de nos approches.

3.2. Protocoles expérimentaux

Nous avons montré dans de précédents travaux la qualité de ces approches (Béchet *et al.* 2009b) en les évaluant par un protocole automatique. Nous proposons dans cet article d'effectuer une évaluation manuelle sur un échantillon de relations syntaxiques. Ainsi, nous pourrons confirmer la qualité de nos approches et également la qualité du protocole d'évaluation automatique décrit ci-dessous.

Nos expérimentations utilisent un corpus écrit en français. Il est extrait du site Web d'informations de Yahoo (http://fr.news.yahoo.com) appartenant au domaine *actualités avec un style journalistique*. Il contient 8 948 articles. Les expérimentations ont été effectuées à partir de 60 000 relations produites dans (Béchet *et al.* 2009b). Dans nos expérimentations, nous sélectionnons *manuellement* cinq concepts, dont les instances sont les objets communs des verbes ayant généré le concept[1], présentés dans le tableau 1.

Concepts	*Organisme /Administration*	*Fonction*	*Objets symboliques*	*Sentiment*	*Manifestation de protestation*
Instances	parquet	négociateur	drapeau	mécontentement	protestation
	mairie	cinéaste	fleur	souhait	grincement
	gendarme	écrivain	spectre	déception	indignation
	préfecture	orateur		désaccord	émotion
	pompier			désir	remous
	onu				tollé
					émoi
					panique

Tableau 1. Les cinq concepts sélectionnés et leurs instances

1 Issus des concepts de verbes ayant un score de plus de 0,7 avec le score d'Asium (Faure 2000)

L'objectif de nos protocoles présentés ci-dessous est d'évaluer la qualité des objets complémentaires pouvant enrichir les cinq concepts définis. L'objectif est donc de proposer à l'expert les objets complémentaires les plus pertinents. Dans ce but, nous allons utiliser les approches présentées en section 3.1 qui classeront ces objets. Dans la section suivante, nous proposons d'évaluer la qualité du classement obtenu.

Protocole d'évaluation automatique

Le principe de l'évaluation automatique est d'utiliser un second corpus écrit également en français, de taille plus conséquente que celui d'où proviennent les relations induites. Les deux corpus sont du même domaine. Nous jugeons alors comme bien formés des relations induites qui vont être présentes *nativement* dans le second corpus. Une telle relation sera alors qualifiée de **positive**. Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le second corpus n'est pas pour autant non pertinente. De plus, un objet complémentaire dans une relation syntaxique jugée pertinente peut également ne pas être un « bon » candidat pour un concept. C'est pourquoi nous présentons dans la section suivante un protocole d'évaluation manuel.

Protocole d'évaluation manuel

Pour mesurer la qualité de notre protocole automatique et vérifier le bon comportement des approches de validation des relations syntaxiques induites, nous proposons d'effectuer une validation manuelle des relations. Nous disposons de huit évaluateurs. Nous leurs avons alors soumis un formulaire. Celui-ci a pour objectif de faire valider manuellement des termes pouvant appartenir à un concept. Pour chacun des cinq concepts extraits, nous soumettons aux évaluateurs les objets candidats, qui ne sont autres que les objets **complémentaires** de l'un ou l'autre des verbes, tel que défini dans la section 2.2. L'évaluateur doit alors mesurer la pertinence d'un terme pour un concept donné en respectant le barème suivant :

- 2 : Parfaitement pertinent
- 1 : Susceptible d'être pertinent
- 0 : Non pertinent
- N : Ne se prononce pas

La figure 2 présente une capture d'écran du formulaire soumis aux experts.

Nous présentons alors deux variantes permettant d'utiliser les scores attribués par les juges : une moyenne des scores obtenus et un système de votes.

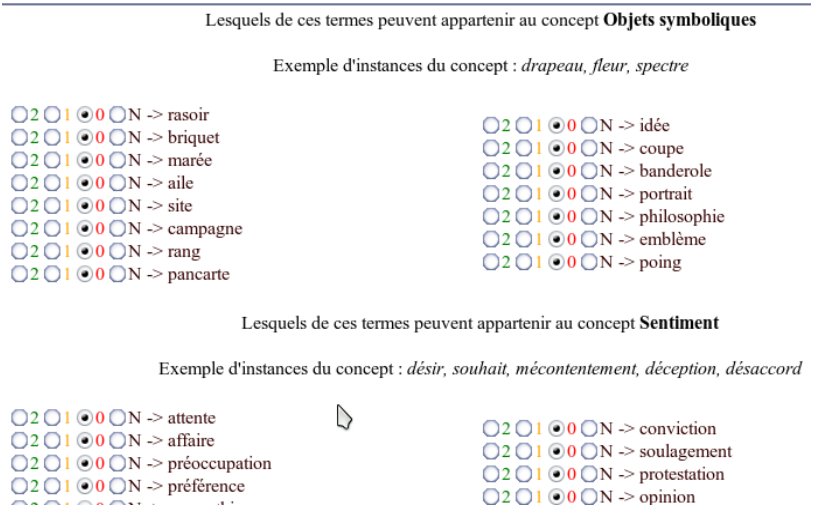

Figure 2. Capture d'écran du formulaire d'évaluation manuelle

La moyenne. Après l'évaluation des objets candidats (553 termes) par les experts, nous effectuons une moyenne des résultats obtenus en faisant varier la tolérance aux résultats obtenus. Nous distinguons alors différents intervalles afin de considérer un résultats comme positif ou non. Par exemple, un terme peut-être positif si son score est supérieur à 1.

Le vote. Une autre manière de qualifier un candidat de positif est de soumettre les scores donnés par les juges à un système de vote. Nous qualifions alors de pertinent un candidat si un pourcentage *p* de juges l'ont jugé pertinent. Un score pertinent d'un juge peut alors être *1* ou *2*.

Une fois la notion de *candidats pertinents* définie avec les protocoles présentés, nous proposons d'évaluer le classement issu de nos différentes approches en utilisant les courbes ROC. Cette méthode est décrite ci-dessous.

Les courbes ROC

Terme	***Validation Manuelle***
Conviction	+
Opinion	+
Pr f rence	-
Attente	-
Col re	+

Tableau 2. Exemple de classement de termes du concept *Sentiment*

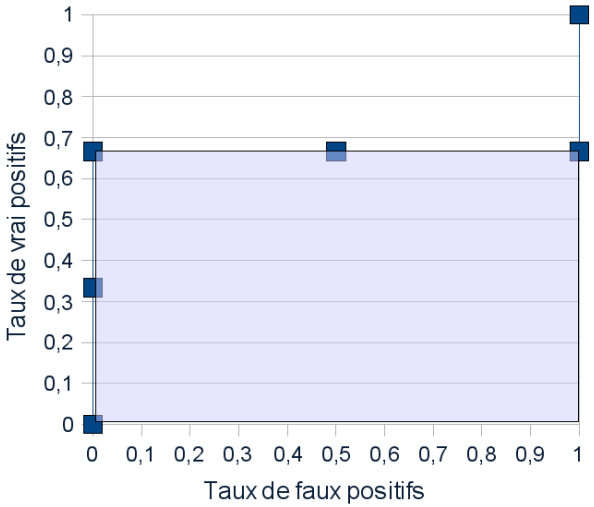

Figure 3. Courbe ROC de l'exemple du tableau 2

La méthode des courbes ROC (Receiver Operating Characteristic), détaillée par (Ferri *et al.* 2002), fut utilisée à l'origine dans le domaine du traitement du signal. Cette méthode est fréquemment employée en médecine afin d'évaluer automatiquement la validité d'un diagnostique de tests. On trouve en abscisse des axes représentant une courbe ROC le taux de faux positifs et l'on trouve en ordonnée le taux de vrais positifs. La surface sous la courbe ROC ainsi créée est appelée AUC (Area Under the Curve). Un des avantages de l'utilisation des courbes ROC réside dans leur résistance à la non parité de la répartition du nombre d'exemples positifs et négatifs.

Une courbe ROC représentée par une diagonale correspond à un système où les relations syntaxiques ont une distribution aléatoire, la progression du taux de vrais positifs est accompagnée par la dégradation du taux de faux positifs. Considérons le cas d'une validation de relations syntaxiques induites. Si toutes les relations sont positives (ou pertinentes), l'AUC vaudrait 1, ce qui signifie avoir toutes les relations pertinentes en début de liste, donc ordonnées de manière optimale.

Le tableau 2 présente un exemple de termes ordonnés avec la seconde combinaison évalués par une validation manuelle pour le concept *Sentiment*. La courbe ROC alors obtenue avec cette validation est présentée dans la figure 3. Nous obtenons finalement une AUC (aire sous la courbe ROC, bleutée sur la figure 3) de 2/3 avec cet exemple.

3.3. Résultats expérimentaux

L'objectif que nous nous fixons dans ces travaux est de réduire la tâche de l'expert en filtrant le nombre de relations syntaxiques induites candidates à un concept. Les expérimentations ci-dessous ont pour but de montrer dans quelle mesure nos approches de validations sont intéressantes. Ainsi nous introduisons un seuil qui n'est autre que le nombre de relations syntaxiques considérées.

Un seuil fixé à 100 indique que l'on ne mesure l'AUC que pour les 100 premières relations syntaxiques. Le tableau 3 présente les AUC obtenues avec les approches validation Web et combinaison 2. La combinaison 2 obtient les meilleures performances. Les résultats des autres approches sont présentés en annexe. Ce tableau compare le protocole d'évaluation manuel avec l'automatique. Pour le protocole manuel, nous ne présentons que les résultats obtenus avec le système de *vote*. Une relation positive est validée si 75% des experts ont attribué la note de 2. Les résultats utilisant la *moyenne*, présentés en annexe, sont assez similaires, et donc non reportés ici.

Avec la seconde combinaison, l'évaluation manuelle donne d'excellents résultats pour les premières relations (AUC jusqu'à 0,83). Les résultats sont de bonne qualité (AUC de 0,70) jusqu'au seuil de 350, pour se dégrader avec la totalité des candidats (AUC proche de l'aléatoire 0,5). Nous ne pouvons ainsi pas fournir à l'expert une liste triée de l'ensemble des candidats mais une liste contenant un sous ensemble. Ainsi, nous privilégions la précision et la qualité de la liste fournie à l'expert en réduisant en contre partie le nombre de candidats disponibles initialement (plus faible rappel).

	Validation Web		***Combinaison 2***	
	Vote	**Auto**	**Vote**	**Auto**
nb relations	***AUC***		***AUC***	
50	0,64	0,59	**0,81**	**0,90**
100	0,50	0,60	**0,83**	**0,87**
150	0,62	0,66	**0,80**	**0,84**
200	0,61	0,65	**0,76**	**0,79**
250	0,56	0,66	**0,71**	**0,75**
300	0,51	0,65	**0,70**	**0,74**
350	0,57	0,67	**0,69**	**0,75**
400	0,59	0,67	**0,67**	**0,74**
450	0,61	0,67	**0,65**	**0,71**
500	0,56	0,68	**0,57**	**0,70**
550	**0,52**	**0,69**	0,52	0,69

Tableau 3. AUC obtenues avec la seconde combinaison pour le protocole manuel et automatique

Nous proposons maintenant de comparer les scores de l'évaluation manuelle avec l'automatique. Nous constatons que les résultats sont du même ordre pour les deux approches. En effet, les résultats de la combinaison 2 sont de très bonne qualité pour les faibles seuils et se dégradent avec la totalité des résultats. Pour la validation Web, les résultats sont assez réguliers de l'ordre de 0,60 pour l'évaluation manuelle et 0,65 pour l'automatique. Nous montrons alors, sur cet échantillon de candidats que notre protocole de validation automatique est de bonne qualité. Il permet en effet de montrer que les premières relations sont les mieux classées avec la seconde combinaison. Il montre également que cette approche fournit les meilleurs classements (Cf. tableaux en annexe).

Néanmoins, les scores obtenus ont tendance à être surévalués avec le protocole automatique. Ces scores s'expliquent notamment par la diversité des tâches effectuées par les deux protocoles. Le protocole manuel cherche à connaître la pertinence d'un terme dans un concept. Le protocole automatique propose de mesurer la cohérence d'une relation syntaxique formée d'un verbe et d'un objet complémentaire. Ces tâches, bien qu'assez proches, ne visent pas les mêmes objectifs. Il est en effet plus difficile de mesurer de manière automatique la qualité d'un candidat potentiel à un concept que la qualité d'une relation syntaxique.

Nous avons présenté précédemment une méthode afin de construire et d'enrichir des classes conceptuelles en utilisant les objets de verbes jugés proches. Nous présentons dans la section suivante une autre méthode d'enrichissement de ces classes utilisant le Web.

4. Enrichissement via le Web

Avec notre précédente méthode d'enrichissement, nous utilisons les informations d'un corpus afin de proposer de nouveaux termes pour enrichir des concepts. Une telle approche utilise des connaissances spécifiques pour enrichir les concepts. Elle est donc propre au corpus. Nos approches fondées sur le Web permettent une validation utilisant des ressources plus générales. Cependant, les termes proposés suite à ces validations sont limités à la thématique du corpus d'où ils sont extraits. Nous présentons dans cette section une autre approche d'enrichissement fondée sur le Web utilisant ainsi des ressources de domaines plus généraux que celles d'un corpus.

4.1. Méthode d'enrichissement

L'objectif de cette méthode est de fournir de nouveaux candidats aux concepts formés tel que décrit section 2. Elle se fonde sur l'énumération de termes sémantiquement proches présents sur le Web. Par exemple, en saisissant dans un moteur de recherche la requête (chaîne de caractères) "lundi, mardi et", nous obtenons d'autres jours de la semaine en résultats.

Afin d'appliquer cette méthode, nous considérons dans un premier temps les objets communs des verbes jugés sémantiquement proches. Ils constituent les *instances de références* des classes ainsi formées. Nous proposons alors d'utiliser le Web afin d'acquérir de nouveaux candidats. Cette méthode présente l'avantage de ne plus se limiter aux termes du corpus dont les classes conceptuelles sont issues.

Considérons alors les N concepts $C_{i \in \{1,N\}}$ et leurs instances respectives $I_j(C_i)$. Pour chaque concept C_i nous soumettons alors à un moteur de recherche les requêtes suivantes :

"$I_{jA}(C_i)$, $I_{jB}(C_i)$ et" *et* "$I_{jA}(C_i)$, $I_{jB}(C_i)$ ou"

avec jA et jB$\in\{1, NbInstanceC_i\}$ et jA $\neq$ jB. Plus concrètement avec l'exemple de la figure 1, nous fournissons au moteur de recherche les requêtes : "*drapeau, fleur et*", "*drapeau, fleur ou*", "*fleur, drapeau et*", "*fleur, drapeau ou*".

Le moteur de recherche nous retourne alors un ensemble de résultats desquels nous extrayons de nouveaux candidats à un concept. Après avoir identifié la requête dans nos résultats, le terme qui suit notre requête constitue une nouvelle instance du concept, tel qu'illustré dans l'exemple suivant.

Considérons la requête : "*drapeau, fleur et*", le moteur nous retourne alors :

"Tu joues version normale (Carreau, pique, coeur et trèfle) ou version bourbi... heu... suisse-allemande (Gland, **Drapeau, Fleur et** Grelot)".

Après avoir identifié notre requête dans le résultat retourné (en gras sur notre exemple), nous ajoutons au concept le terme suivant directement la requête (ici, le terme *Grelot*).

4.2. Protocole et résultats expérimentaux

Nous avons expérimenté cette seconde méthode d'acquisition de termes afin d'enrichir nos cinq concepts déjà expérimentés dans la section 3.3. Nous utilisons pour nos expérimentations l'API du moteur de recherche *Yahoo!* afin d'obtenir nos nouveaux termes.

Trente cinq nouveaux termes ont été obtenus, sur lesquelles nous avons appliqué différents traitements. Tout d'abord, nous appliquons un filtrage grammatical afin de ne conserver que des noms, puis nous appliquons un élagage en supprimant les termes génériques, tels que *même, chose, avoir, etc.* Il nous reste alors trente termes.

Nous faisons alors évaluer à 3 experts les différents termes en leurs demandant si un terme peut être considéré comme pertinent (respectivement, non pertinent) dans un concept. Nous obtenons alors trois évaluations. Nous calculons pour chacune le taux de positifs défini par le nombre de termes pertinents sur le nombre total de termes. Notons que dans notre cas, le taux de positifs n'est autre que la précision. Finalement, nous effectuons une moyenne de la précision qui atteint **0,70** dans nos expérimentations.

Ce résultat est assez encourageant bien que pouvant être amélioré. En effet, la thématique même du concept présenté aux experts est discutable car elle est établie manuellement. Alors la subjectivité de l'évaluation humaine joue un rôle non négligeable dans cette évaluation.

Citons par exemple le concept « *manifestation de protestation* ». La question posée est alors : le terme "*adhésion*" est-il une instance correcte de ce concept ? Une définition du terme adhésion (provenant du TLFI) est : "Reconnaissance implicite ou explicite de l'autorité d'une loi, d'un gouvernement, etc.". D'une manière triviale en se fondant sur cette définition, on aurait plutôt tendance à dire que ce terme appartiendrait à un concept opposé à celui-ci. Mais si l'on considère une adhésion comme un engagement politique ou associatif s'opposant aux règles ou aux lois établies, il peut être perçu comme un moyen de protestation. Cette subjectivité humaine pose là une question importante au niveau de la qualité de l'évaluation humaine pour des systèmes de fouilles de textes.

5. Synthèse

Nous proposons dans cette section de présenter une synthèse des deux approches d'enrichissement des concepts.

Nous allons nous appuyer sur le concept *Sentiment*. Ce concept a été formé par les objets communs de deux verbes jugés proches : *exprimer* et *manifester*. Les instances de ce concept sont : *désir, souhait, mécontentement, déception, désaccord*. Nous proposons alors d'utiliser les deux méthodes présentées dans cet article afin d'enrichir ce concept.

Première méthode : Induction d'informations provenant du corpus.

Nous *fabriquons* une liste d'objet dits *complémentaires* à partir du corpus. Ces objets vont alors être ordonnés avec la seconde combinaison des approches validation Web et vecteurs sémantiques. Nous les soumettons alors à un expert qui va sélectionner les plus pertinentes.

Les objets retenus par l'expert sont : sympathie, regrets, doute, crainte, exaspération, satisfaction, sensibilité, espoir, indignation, dédain, joie, amertume, désarroi, solidarité, confiance, colère.

Seconde méthode : Enrichissement via le Web.

La seconde méthode propose d'utiliser des ressources extérieures *généralisant* ainsi notre concept original. Elle se fonde sur l'envoi de requête à un moteur de recherche. Les candidats obtenus par le Web après validation de l'expert sont : *fatalisme, stabilité, angoisse, inconscience*.

6. Conclusion

Cet article a proposé de joindre les ressources d'un corpus à celles du Web afin d'enrichir des classes conceptuelles. De telles classes ont été formées par des objets communs de verbes jugés sémantiquement proches. Nous avons alors proposé deux approches d'enrichissement afin d'extraire de nouveaux termes. Une première utilisant les objets des verbes dits complémentaires. Nous avons alors validé ces approches avec différents protocoles, dont un totalement automatique, se révélant assez efficace. Cependant, un tel enrichissement limite la construction des classes à un domaine spécialisé, à l'image du corpus dont les termes sont extraits. Nous avons alors présenté une autre méthode, se fondant non plus sur le corpus mais sur le Web. Cette méthode s'est avérée pertinente. Mais une évaluation plus poussée doit être menée afin de confirmer sa qualité.

Nous envisageons comme futurs travaux pour la première méthode, d'introduire le contexte dans nos différentes approches de validation. Pour la validation Web, il s'agit introduire le contexte dans la requête fournie au moteur de recherche. Pour les vecteurs sémantiques, il s'agit d'utiliser des vecteurs dit *contextualisés* prenant en compte la structure morphosyntaxique de la phrase dont le terme à valider est issu. De plus, nous pensons intégrer la notion de *nominalisation* afin d'acquérir une plus grande quantité de relations syntaxiques. Citons par exemple "consommer un fruit", plus fréquemment rencontré comme suit "consommation de fruits".

La seconde approche mérite quant à elle d'être approfondie, en effectuant notamment des expérimentations sur des corpus d'autres thématiques.

Nous souhaitons également pouvoir exécuter à plusieurs reprises l'acquisition de termes par le Web. Nous devrons alors faire sélectionner les termes par un expert ou bien avec une méthode automatique qui devra être proposée. Ces termes vont effet être utilisés afin de former la requête Web d'où la nécessité de les sélectionner rigoureusement.

Les travaux présentés dans cet article posent la question ouverte de la qualité de l'évaluation des ressources terminologiques, sémantiques ou autres. Bien que la plupart de ces ressources furent conçues avec la validation d'experts des domaines traités, la subjectivité humaine peut mettre en doute la qualité de certaines de ces ressources ainsi que les évaluations telles que celles présentées dans ce papier. Par exemple, un domaine est-il défini par toutes personnes de manière analogue ? Ou encore un concept a-t-il toujours la même caractérisation ?

Bibliographie

Béchet, N. *Comment valider automatiquement des relations syntaxiques induites*. In EvalECD'09, acte des ateliers de EGC'09, p. A5-25 - A5-33, 2009a.

Béchet, N., Roche M., Chauché J. Towards the Selection of Induced Syntactic Relations. In ECIR'09, poster proceedings, to appear, 2009b.

Bourigault D. et Fabre C. *Approche linguistique pour l'analyse syntaxique de corpus*. Cahiers de Grammaires, 25, 131–151., 2000.

Bourigault D., Lame G. *Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit*, in TAL, 43-51, 2002.

Chauché, J. *Un outil multidimensionnel de l'analyse du discours*. In Proceedings of Coling, Standford University, California, p. 11–15, 1984.

Chauché, J. *Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance*. In TA Information, pp. 17–24, 1990.

Church, K. W. et Hanks P. *Word association norms, mutual information, and lexicography*. In Computational Linguistics, Volume 16, pp. 22–29, 1990.

Desrosiers-Sabbath *Comment enseigner les concepts* - Sillery: Presses de l'Université du Québec, 1984.

Faure, D. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph. D. thesis, Université Paris-Sud, 2000.

Ferri, C., Flach P., et Hernandez-Orallo J.. *Learning decision trees using the area under the ROC curve*. In Proceedings of ICML'02, pp. 139–146., 2002.

Harris Z. *Mathematical Structures of Language*, New-York, John Wiley & Sons, 1968.

Larousse, T. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Ed.Larousse, Paris, 1992.

L'Homme M. -C. *Le statut du verbe en langue de spécialité et sa description lexicographique*. Cahiers de Lexicologie. 73. 61-84, 1998.

Nazarenko A., Zweigenbaum P., Habert B, Bouaud J. *Corpus-based Extension of a Terminological Semantic Lexicon*. In Recent Advances in Computational Terminology, 327-351, 2001.

Smadja, F., McKeown K. R., et Hatzivassiloglou V. *Translating collocations for bilingual lexicons : A statistical approach*. Computational Linguistics 22(1), 1–38, 1996.

Roche C. *Terminologie et ontologie*. Revue Langages n°157, pp. 48-62, Éditions Larousse, mars 2005.

Turney, P. Mining the Web for synonyms : PMI– R versus LSA on TOEFL. Proc of ECML, LNCS, 2167, 491–502, 2001.

A propos des auteurs

Équipe TAL - LIRMM
UMR 5506, CNRS, Univ. Montpellier 2
34392 Montpellier Cedex 5 - France
{bechet,mroche,*chauche*}*@lirmm.fr*
http://www.lirmm.fr/~{bechet,mroche,chauche}

Annexes

Automatique	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,54	0,59	0,65	**0,90**
100	0,54	0,60	0,65	**0,87**
150	0,55	0,66	0,73	**0,84**
200	0,48	0,65	**0,80**	0,79
250	0,52	0,66	0,66	**0,75**
300	0,47	0,65	0,64	**0,74**
350	0,50	0,67	0,62	**0,75**
400	0,51	0,67	0,64	**0,74**
450	0,50	0,67	0,62	**0,71**
500	0,53	0,68	0,62	**0,70**
550	0,55	**0,69**	0,62	**0,69**

Tab. 4: AUC obtenues avec le protocole automatique.

Vote	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,50	0,64	0,78	**0,81**
100	0,66	0,50	0,54	**0,83**
150	0,55	0,62	0,69	**0,80**
200	0,57	0,61	0,74	**0,76**
250	0,53	0,56	0,58	**0,71**
300	0,35	0,51	0,55	**0,70**
350	0,42	0,57	0,53	**0,69**
400	0,46	0,59	0,53	**0,67**
450	0,46	0,61	0,53	**0,65**
500	0,41	0,56	0,46	**0,57**
550	0,39	**0,52**	0,43	**0,52**

Tab. 5: AUC obtenues avec le protocole manuel, en utilisant le *vote*.

Moyenne	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,54	0,62	**0,79**	0,74
100	0,57	0,53	0,54	**0,82**
150	0,52	0,58	0,69	**0,79**
200	0,50	0,61	**0,75**	0,75
250	0,42	0,57	0,56	**0,65**
300	0,34	0,52	0,49	**0,68**
350	0,42	0,56	0,51	**0,67**
400	0,47	0,58	0,53	**0,66**
450	0,46	0,60	0,52	**0,63**
500	0,41	0,57	0,47	**0,57**
550	0,39	**0,53**	0,43	**0,53**

Tab. 6: *AUC obtenues avec le protocole manuel, en utilisant la moyenne (score positif pour une moyenne supérieur ou égale à 1,75).*