Comment valider automatiquement des relations syntaxiques induites

Nicolas Béchet*, Mathieu Roche*, Jacques Chauché*

*LIRMM - UMR 5506, CNRS - Univ. Montpellier 2 - 34392 Montpellier Cedex 5 - France
{bechet,mroche,chauche}@lirmm.fr

Résumé. Nous présentons dans cet article des approches visant à valider des relations syntaxiques induites de type Verbe-Objet. Ainsi, nous proposons d'utiliser dans un premier temps une approche s'appuyant sur des vecteurs sémantiques déterminés à l'aide d'un thésaurus. La seconde approche emploie une validation Web. Nous effectuons des requêtes sur un moteur de recherche associées à des mesures statistiques afin de déterminer la pertinence d'une relation syntaxique. Nous proposons enfin de combiner ces deux méthodes. La qualité de nos approches de validation de relations syntaxiques a été évaluée en utilisant des courbes ROC.

1 Introduction et contexte

L'acquisition de connaissances sémantiques est une importante problématique en Traitement Automatique des Langues (TAL). Ces connaissances peuvent par exemple être utilisées pour extraire des informations dans les textes ou pour la classification de documents. Les connaissances sémantiques peuvent être obtenues par des informations syntaxiques (Fabre et Bourigault (2006)). Comme nous allons le montrer dans cet article, les connaissances sémantiques acquises via la syntaxe permettent de constituer des classes conceptuelles (regroupement de mots ou termes sous forme de concepts). Par exemple, les mots *hangar*, *maison* et *mas* sont regroupés dans un concept *bâtiment*. De plus, ces concepts peuvent être organisés sous forme hiérarchique formant ainsi une classification conceptuelle.
Deux types d'informations syntaxiques peuvent être utilisés pour construire les classes sémantiques : les relations "classiques" issues d'une analyse syntaxique (Lin (1998), Wermter et Hahn (2004)) et les relations dites "induites" à partir des textes. Cet article s'intéresse plus particulièrement à ces dernières. La définition d'une relation induite est présentée ci-dessous.
La méthode d'ASIUM consiste à regrouper les objets des verbes déterminés comme proches par une mesure de qualité (Faure (2000)). D'autres approches utilisent également ce principe, comme le système UPERY (Bourigault (2002)) qui regroupe les termes par des mesures de proximité distributionnelle. Par exemple, dans la figure 1, si les verbes *consommer* et *manger* sont jugés proches, des objets pouvant être obtenus par le biais d'informations syntaxiques sont regroupés (dans notre cas, les objets *essence*, *légume*, *nourriture* et *fruit*). Cependant, en considérant ce groupe d'objets, nous pouvons intuitivement exclure le mot *essence*. Notons que les objets *essence*, *légume* et *nourriture* appartiennent à un même contexte en tant qu'objets

du verbe *consommer*. De plus, les objets *légume* et *nourriture* sont également des objets du verbe *manger* sur la figure 1. Nous appelons dans ce cas l'objet *essence* du verbe *consommer*, un objet **complémentaire** du verbe *manger*. Par opposition, les objets *légume* et *nourriture* sont appelés des objets **communs** au verbe *manger*, car ils sont également objets du verbe *consommer*. La relation syntaxique formée par un verbe et son objet complémentaire est ainsi appelée relation syntaxique **induite** comme les relations *manger essence* et *consommer fruit* de la figure 1. Notons que ces relations syntaxiques induites sont des connaissances nouvelles "apprisent"

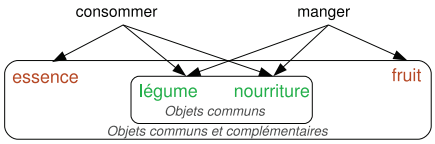

FIG. 1 – *Objets communs et complémentaires des verbes "consommer" et "manger".*

à partir des corpus car elles ne sont pas explicitement présentes dans les données textuelles. Notre objectif dans cet article est de déterminer quels objets complémentaires sont pertinents. Ainsi, nous limitons la tâche de l'expert en proposant de valider uniquement des termes jugés pertinents par nos approches, au lieu de faire valider toutes les relations induites possibles (et donc les éléments des concepts acquis) par un expert. Par exemple, nous souhaitons connaître la pertinence des objets induits de la figure 1 *fruit* et *essence* dans un concept automatiquement constitué des mots *essence*, *légume*, *nourriture* et *fruit*. L'évaluation de la qualité des relations syntaxiques induites *consommer fruit* et *manger essence* nous donnera un indice de qualité des objets complémentaires.
Pour effectuer un telle évaluation, nous proposons de comparer deux approches pour finalement les combiner. La première approche représente une relation syntaxique induite comme une combinaison de concepts issus d'un thésaurus dans un espace vectoriel sémantique. La seconde approche consiste à utiliser la fréquence d'apparition d'une relation induite sur le Web par l'utilisation d'un moteur de recherche. Nous présentons dans la section suivante ces deux approches ainsi que les combinaisons effectuées entre elles. La section 3 présente les expérimentations que nous avons menées en évaluant ces approches.

2 Différentes approches de validation

Cette section résume les différentes approches utilisées pour valider les relations syntaxiques induites. Nous précisons dans un premier temps la manière dont les relations induites sont obtenues.

2.1 Les relations induites

Cette section présente une manière de déterminer les relations induites à partir de relations syntaxiques issues d'un corpus. Nous utilisons dans un premier temps l'analyseur morphosyntaxique SYGFRAN pour extraire des relations syntaxiques d'un corpus. Nous calculons ensuite

la proximité sémantique des verbes issus des relations extraites, en nous appuyant sur la mesure d'ASIUM décrite dans (Faure (2000)). Notons que la mesure d'ASIUM considère comme sémantiquement proches deux verbes partageant un certain nombre d'objets en communs. Ainsi nous ne considérons que les couples de verbes jugés proches comme par exemple les verbes *consommer* et *manger* de la figure 1. La suite de l'article s'intéresse plus particulièrement à l'évaluation automatique des relations induites (par exemple *consommer fruit* et *manger essence*).

2.2 Les vecteurs sémantiques

Nous présentons dans cette section notre approche utilisant les vecteurs sémantiques afin de mesurer la proximité sémantique entre le verbe et l'objet des relations syntaxiques. Nous indiquons dans un premier temps, comment sont utilisées des approches similaires dans la littérature. Wilks (1998) discute du fait que les différents descripteurs d'un thésaurus comme le Roget peuvent être bénéfiques pour des tâche relatives au TAL. Des approches dites *à la Roget* sont employées dans divers domaines du TAL comme la désambiguïsation sémantique, la recherche d'information, la cohésion textuelle, ou comme mesure de similarité entre termes.
Jarmasz et Szpakowicz (2003) utilisent la taxonomie de la structure du thésaurus Roget pour déterminer la proximité sémantique entre deux termes. Ils obtiennent des résultats de bonne qualité pour les tests du TOEFL, ESL et Reader's Digest. Notre approche utilise quant à elle une approche *à la Roget* dans un contexte différent. Nous proposons de 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 le verbe et l'objet de la relation originale. Concrètement avec l'exemple de la figure 1, il s'agit de mesurer la proximité sémantique des relations *manger fruit* (relation originale) et *consommer fruit* (relation induite). L'objectif est d'attribuer un score à chaque relation induite, afin de les classer par pertinence. Nous avons choisi de représenter une relation induite par un vecteur sémantique. Il 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, "*consommer*" est associé à "*fin, nutrition, accomplissement, usage, dépense et repas*" et "*nourriture*" à "*nutrition, éducation, repas et pain*", chaque élément se voyant attribuer un poids de 1. La représentation vectorielle de la relation syntaxique *consommer nourriture* se traduit par un vecteur de dimension 873, illustrée par la figure 2. Ce vecteur résulte d'une combinaison linéaire de la représentation de *consommer* et de *nourriture*, dont les coefficients prennent en compte la structure syntaxique (dans notre cas, un verbe et son objet) (Chauché (1990)). Ainsi, les composantes du vecteur de la relation *consommer nourriture* sont toutes nulles sauf celles associées aux concepts 58, 337, 415, 538, 567, 835, 855 et 857. Afin de mesurer la pertinence

FIG. 2 – *Vecteur sémantique de "consommer nourriture".*

d'une relation induite, nous évaluons si cette relation partage les mêmes concepts que la relation syntaxique dont elle est issue. Pour mesurer une telle proximité, nous calculons le cosinus de l'angle formé par les deux vecteurs sémantiques, le cosinus étant défini comme la division du produit scalaire des vecteurs par le produit de leurs normes. Notre objectif est d'obtenir un classement des différentes relations syntaxiques par cette mesure de cosinus, afin de valider celles apparaissant en tête du classement. Un exemple de classement de relations induites avec la méthode des vecteurs sémantiques est présenté en section 2.5.
Une autre manière d'attribuer un score aux relations induites qui sera décrite dans la section suivante, consiste à mesurer la dépendance entre le verbe et l'objet constituant la relation.

2.3 La validation Web

La validation Web utilisée pour mesurer la dépendance entre verbe et objet d'une relation induite, est proche de l'approche de Turney (2001), utilisant le Web pour définir une fonction de rang. L'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) décrit par Turney (2001) utilise le moteur de recherche AltaVista pour déterminer les synonymes appropriés à une requête donnée. D'autres approches utilisent le Web dans la littérature comme les travaux de Cilibrasi et Vitanyi (2007) qui proposent de mesurer la similarité de termes en utilisant, entre autres, le moteur de recherche Google.
Dans notre cas, nous proposons de mesurer la dépendance entre un verbe et un objet d'une relation induite afin d'établir un classement par pertinence des relations. Pour cela, nous interrogeons le Web en fournissant à un moteur de recherche une requête, dans notre cas une relation syntaxique, sous forme de chaîne de caractères (par exemple, "consommer un fruit"). Cette approche présente la particularité de refléter la popularité d'une relation syntaxique sur le Web, s'adaptant ainsi à une époque ou une certaine mode d'écriture. La requête fournie au moteur de recherche est donnée par la fonction $nb(x)$ qui représente le nombre de pages de résultats retournés par la soumission d'une chaîne de caractères x au moteur de recherche de Yahoo en utilisant une API (*http ://api.search.yahoo.com*). En français, langue sur laquelle nous nous appuyons dans nos travaux, un verbe est couramment séparé d'un objet par un article. Ainsi, nous considérons cinq articles fréquents : *un*, *une*, *le*, *la*, *l'* pour composer notre chaîne de caractères représentant notre relation induite. $nb(v, o)$ est alors défini comme le nombre de pages retournées pour la relation syntaxique Verbe-Objet (v,o) avec respectivement v et o, le verbe et l'objet de cette relation. La formule suivante décrit le calcul effectué afin d'obtenir la fréquence d'apparition d'une relation syntaxique :
$nb(v, o) = nb(v$ **un** $o) + nb(v$ **une** o$) + nb(v$ **le** o$) + nb(v$ **la** o$) + nb(v$ **l'** o$)$
$nb(v$ **un** $o)$ est la valeur retournée par le moteur de recherche avec la chaîne de caractères v **un** o.
Nous utilisons alors une mesure statistique afin d'évaluer la dépendance entre verbe v et objet o d'une relation induite. Des travaux dans la littérature comme (Vivaldi et al. (2001), Daille (1994)) utilisant des fonctions de rang obtiennent le plus souvent de meilleurs résultats avec l'Information Mutuelle au Cube (IM^3). Nos travaux expérimentaux ont confirmé ces résultats (Roche et Prince (2007)). Nous n'utiliserons donc que la fonction de rang IM^3.
L'IM^3 est une mesure empirique fondée sur l'Information Mutuelle (Church et Hanks (1990)). Elle permet de valoriser les co-occurrences fréquentes, ce qui n'est pas le cas avec l'IM (Daille (1994)). Adaptée à notre problématique de relations syntaxiques, pour un verbe v, un objet o et la fonction nb désignant le nombre de pages retournées par le moteur de recherche Yahoo,

nous définissons l'IM^3 par la formule suivante :

$$IM^3(v,o) \quad = \quad \log \frac{nb(v,o)^3}{nb(v)nb(o)} \tag{1}$$

Finalement, nous obtenons un classement des relations syntaxiques induites en mesurant pour chacune leur IM^3 ainsi définie. Les relations induites les plus pertinentes selon les deux approches (vecteurs sémantiques et validation Web) doivent se retrouver en tête de liste. Notons que nos expérimentations effectuées sur un ensemble de 60000 relations syntaxiques (Cf section 3) nécessitent 420 000 requêtes en utilisant l'IM^3 (1 requête pour le verbe, 1 pour l'objet et 5 pour les couples verbe-objet séparés par cinq articles : $60000 \times 7 = 420000$). Nous n'avons par conséquent pas pris en compte les différentes variantes morphologiques du au grand nombre de requêtes déjà nécessaires. Un exemple de classement de relations induites avec la méthode de la validation Web est présentée en section 2.5.

2.4 L'hybridation

Pour optimiser nos deux approches précédemment présentées, les vecteurs sémantiques (VS) et la validation Web (VW), nous proposons de combiner ces deux techniques.

2.4.1 Combinaison 1 : Une combinaison pondérée par un scalaire

La première combinaison entre ces approches consiste à introduire un paramètre $k \in [0,1]$ pour donner un poids supplémentaire à l'une ou l'autre des approches. Nous normalisons au préalable les résultats donnés par les deux approches à combiner. Alors, pour une relation syntaxique r, nous combinons les approches avec le calcul suivant :

$$combine\_score_r = k \times VS + (1-k) \times VW \tag{2}$$

2.4.2 Combinaison 2 : Un système hybride adaptatif

Nous présentons une seconde approche combinant VS et VW, l'**hybridation adaptative**. Le principe de cette combinaison est de classer dans un premier temps la totalité des relations syntaxiques par l'approche VS. Nous retenons et plaçons en tête les n premières relations syntaxiques. Ensuite, l'approche VW effectue le classement des n relations retenues par la méthode VS. Ainsi avec notre approche adaptative, VS effectue une sélection globale sur la base des connaissances sémantiques et VW affine la sélection préalablement effectuée. Notons que si n correspond au nombre total de relations syntaxiques, ceci revient à appliquer une validation Web "classique".

2.5 Exemple de classement avec cinq relations induites

Cette section présente un exemple de classements de relations syntaxiques induites avec les différentes approches présentées : les vecteurs sémantiques, la validation Web et les deux approches de combinaisons.

Nous calculons tout d'abord les scores résultant de l'approche des vecteurs sémantiques. Les

Relations Verbe-Objet		Cosinus
Induites	Originales	
poursuivre réforme	demander réforme	0,60
dépasser recherche	faire recherche	0,52
réussir évaluation	faire évaluation	0,41
dire croisade	poursuivre croisade	0,37
lancer recherche	mener recherche	0,27

TAB. 1 – *Résultats avec les vecteurs sémantiques.*

Relations Verbe-Objet	nb(Verbe)	nb(Objet)	nb(Verbe, Objet)	IM³
lancer recherche	82 700 000	863 000 000	2 299 288	0,71
poursuivre réforme	46 200 000	39 000 000	45 914	0,49
dire croisade	370 000 000	4 120 000	72	0,41
réussir évaluation	27 600 000	57 900 000	1 366	0,35
dépasser recherche	15 900 000	863 000	363	0,28

TAB. 2 – *Résultats avec la validation Web.*

cinq relations syntaxiques induites et celles existantes sont présentées dans le tableau 1. Nous les représentons dans un premier temps sous forme de vecteurs sémantiques avec SYGFRAN. Nous pouvons alors calculer le cosinus entre les relations syntaxiques induites et celles existantes (Cf section 2.2). Les résultats obtenus pour les cinq relations testées sont présentés dans le tableau 1.

Relations Verbe-Objet	VS	VW	Combinaison 1	Combinaison 2
lancer recherche	0,60	0,49	0,55	1,49
poursuivre réforme	0,41	0,35	0,38	1,35
réussir évaluation	0,52	0,33	0,43	1,33
dépasser recherche	0,37	0,13	0,25	0,37
dire croisade	0,27	0,71	0,49	0,27

TAB. 3 – *Relations syntaxiques triées avec l'ensemble des approches.*

Nous calculons ensuite les scores résultant de la validation Web. Nous effectuons pour cela des requêtes sur le Web avec les objets, verbes et relations syntaxiques induites afin de déterminer le nombre de pages retournées par le moteur de recherche (fonction nb). Alors nous pouvons calculer l'IM^3 pour les cinq relations induites. Les scores obtenus sont présentés dans le tableau 2. Nous appliquons alors la première méthode de combinaison présentée en section 2.4.1. Nous fixons à titre d'exemple le scalaire k à $0,5$ pour donner le même poids à chacune des approches (VS et VW). De plus, dans cet exemple, nous fixons la constante n de la mesure hybride adaptative à 3. Rappelons que la mesure adaptative consiste à classer dans un premier temps les relations syntaxiques induites avec les vecteurs sémantiques, pour ensuite classer dans notre cas les 3 meilleures avec la validation Web. Les résultats obtenus pour les

VS	VW	Combinaison 1	Combinaison 2
poursuivre réforme	lancer recherche	poursuivre réforme	poursuivre réforme
dépasser recherche	poursuivre réforme	lancer recherche	réussir évaluation
réussir évaluation	réussir évaluation	dépasser recherche	dépasser recherche
dire croisade	dépasser recherche	réussir évaluation	dire croisade
lancer recherche	dire croisade	dire croisade	lancer recherche

TAB. 4 – *Classement obtenu des relations syntaxiques.*

deux combinaisons[1] sont présentés dans le tableau 3. Une fois l'ensemble des scores obtenus pour chacune des approches, nous pouvons ordonner les relations syntaxiques. Le classement ainsi obtenu est donné dans le tableau 4.

3 Expérimentations

3.1 Protocole expérimental

Nous extrayons d'un premier corpus les relations induites, que nous ordonnons qualitativement par nos différentes approches. Ce premier corpus écrit en français est extrait du site Web d'informations de Yahoo (*http ://fr.news.yahoo.com/*). Il contient 8 948 articles (16,5 Mo). Ce corpus est utilisé comme corpus de test, il sera nommé *corpus T*. Nous avons obtenu à partir de ce corpus, 60 000 relations syntaxiques induites. Afin de mesurer la qualité de nos relations induites, nous utilisons un second corpus (*corpus V*) écrit également en français, de taille plus conséquente (125 Mo). Celui-ci contient plus de 60 000 articles issus du corpus du quotidien Le Monde. Par ailleurs, les deux corpus sont du même domaine, actualités avec un style journalistique. Nous jugeons comme pertinentes des relations induites créées à partir du *corpus T* si celles-ci sont présentes dans le *corpus V*. Concrètement, si une relation induite est retrouvée dans le *corpus V*, on la qualifiera de **positive**. Dans le cas contraire, elle sera jugée non pertinente et sera donc qualifiée de **négative**. Nous avons opté pour cette validation afin de pouvoir mesurer la qualité de nos approches, sur un grand nombre de relations, de manière automatique (qu'un ou des expert(s) humain(s) ne pourrai(en)t évaluer, faute de temps). Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le *corpus V* n'est pas pour autant non pertinente. Nous proposons d'évaluer nos différentes approches en utilisant les courbes ROC.
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 (dans notre cas, le taux de relations syntaxiques induites non pertinentes, soit les relations non retrouvées dans le *corpus V*) et l'on trouve en ordonnée le taux de vrais positifs (dans notre cas les relations pertinentes, soit celles existantes dans le *corpus V*). La surface sous la courbe ROC ainsi créée est appelée AUC (*Area Under the Curve*).

[1] Afin de représenter sous forme de scores l'application de l'approche adaptative, pour les 3 meilleures relations classées par VS, nous reportons leurs scores respectifs obtenus avec VW, auxquelles nous ajoutons 1, ce qui place ces 3 relations automatiquement en tête de liste (car les scores sont normalisés).

Seuil	*VW*	*VS*	*Seuil*	*VW*	*VS*
5000	0,61	0,51	*35000*	0,75	0,55
10000	0,65	0,52	*40000*	0,76	0,56
15000	0,68	0,54	*45000*	0,78	0,55
20000	0,70	0,54	*50000*	0,79	0,54
25000	0,72	0,55	*55000*	0,80	0,54
30000	0,74	0,55	*60000*	0,81	0,54

TAB. 5 – *AUC obtenues avec les approches Validation Web et Vecteurs Sémantiques.*

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.

3.2 Résultats

Nous présentons dans cette section, les résultats de l'évaluation de nos approches présentées en section 2, pour différents seuils donnés. Concrètement, nous avons produit 60 000 relations syntaxiques induites avec le corpus de test ordonnées avec nos approches en appliquant un seuil. Ce dernier permet de mesurer l'impact de nos approches, en fonction du nombre de relations syntaxiques induites à considérer. Ainsi, un seuil fixé à 5 000 relations signifie que nous calculons l'AUC sur les 5 000 premières relations ainsi classées.
Le tableau 5 présente les AUC obtenues avec l'utilisation des vecteurs sémantiques et la validation Web. Les AUC obtenues avec les vecteurs sémantiques ne sont pas très satisfaisantes

Seuil	VW (k=0)	*k* = 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	VS (k=1)
5000	0,61	0,61	0,61	0,62	0,62	0,63	0,63	**0,66**	0,55	0,56	0,51
10000	0,65	0,66	0,66	0,66	0,67	0,67	**0,67**	0,67	0,64	0,57	0,52
15000	0,68	0,68	0,68	0,68	0,69	0,70	0,70	**0,71**	0,65	0,56	0,54
20000	0,70	0,70	0,71	0,72	0,73	**0,74**	0,73	0,71	0,66	0,58	0,54
25000	0,72	0,73	0,75	0,75	0,75	**0,76**	0,75	0,73	0,69	0,62	0,55
30000	0,74	0,76	0,77	0,78	0,78	**0,78**	0,77	0,75	0,71	0,63	0,55
35000	0,75	0,78	0,79	**0,79**	0,79	0,78	0,77	0,75	0,72	0,64	0,55
40000	0,76	0,79	**0,79**	0,79	0,79	0,78	0,77	0,75	0,71	0,65	0,56
45000	0,78	**0,79**	0,79	0,79	0,79	0,79	0,78	0,76	0,73	0,67	0,55
50000	0,79	**0,80**	0,80	0,79	0,78	0,75	0,74	0,72	0,69	0,64	0,54
55000	0,80	**0,80**	0,79	0,78	0,75	0,73	0,71	0,69	0,66	0,62	0,54
60000	**0,81**	0,79	0,78	0,76	0,74	0,72	0,70	0,68	0,65	0,61	0,54

TAB. 6 – *AUC obtenues avec la première combinaison.*

avec des scores proches d'une distribution aléatoire ($AUC = 0,5$). La structure des vecteurs sémantiques peut expliquer ces résultats. En effet, les 873 concepts définissant ces vecteurs ne sont pas assez discriminants, notre corpus utilisant un vocabulaire assez homogène (corpus d'actualités). Cette faible dimension des vecteurs sémantiques ne permet donc pas de classer assez finement nos relations syntaxiques induites. La validation Web donne quant à elle de meilleures AUC. Pour un seuil réduit, inférieur à 20 000, les résultats restent néanmoins assez faibles (moins de $0,70$). Cela signifie que pour l'ensemble des relations syntaxiques évaluées,

cette approche est meilleure (AUC de $0,81$) mais que le classement des premières relations reste difficile. Ceci peut s'expliquer par le fait que les relations syntaxiques les plus populaires ne sont pas nécessairement les plus pertinentes. Nous proposons alors d'appliquer la première approche (combinaison 1) effectuant une combinaison des vecteurs sémantiques et de la validation Web. Nous faisons varier le paramètre k de 0 à 1 par intervalle de $0,1$. Le tableau

Seuil	*VW*	*Comb. 1*	*Comb. 2*	*Seuil*	*VW*	*Comb. 1*	*Comb. 2*
5000	0,61	**0,66**	**0,83**	*35000*	0,75	0,75	**0,83**
10000	0,65	**0,67**	**0,82**	*40000*	0,76	0,75	**0,83**
15000	0,68	**0,71**	**0,83**	*45000*	0,78	0,76	**0,83**
20000	0,70	**0,71**	**0,83**	*50000*	0,79	0,72	**0,82**
25000	0,72	**0,73**	**0,83**	*55000*	0,80	0,69	**0,82**
30000	0,74	**0,75**	**0,83**	*60000*	0,81	0,68	**0,81**

TAB. 7 – *AUC obtenues avec la seconde combinaison.*

6 montre les résultats obtenus avec cette approche. Intuitivement, cette combinaison devrait donner des résultats pertinents pour des valeurs de k faibles, ce qui privilégie l'approche VW. Néanmoins, cette supposition ne se vérifie pas pour tous les seuils. En effet, nous obtenons, pour la première moitié des seuils considérés, de meilleures AUC que celles obtenues pour la validation Web, avec des valeurs de k autour de $0,7$. Ces améliorations sont cependant peu significatives, de l'ordre de 3%. Pour une valeur de k supérieure à $0,5$, nous obtenons des AUC assez proches de celles obtenues avec la validation Web. Ces résultats nous amènent à penser que l'approche utilisant les vecteurs sémantiques peut se révéler pertinente pour certaines relations. Ainsi, nous allons évaluer la seconde approche qui selectionne dans un premier temps et de manière globale des relations retenues par les vecteurs sémantiques. Le tableau 7 présente la seconde combinaison aux meilleurs scores obtenus pour de faibles seuils (les plus difficiles à améliorer) pour les approches précédentes : la validation Web et la première combinaison avec $k = 0,7$. La figure 3 présente quant à elle, les courbes ROC correspondantes au seuil 5000.
Pour la seconde combinaison, nous allons classer les 60 000 relations avec les vecteurs sémantiques. Puis, nous classons les n premières avec la validation Web (la valeur de n propre au paramètre de la seconde combinaison correspond ici au seuil). Nous obtenons avec la seconde combinaison de meilleures AUC que les approches précédentes, quelque soit le seuil testé. Les améliorations sont encore plus significatives pour les premiers seuils. En effet, pour un seuil de 5 000, l'AUC passe de $0,61$ avec la validation Web à $0,83$ avec la seconde combinaison. Cette combinaison est l'approche fournissant de meilleurs résultats afin de répondre à notre problématique, la validation automatique des relations syntaxiques induites. Notons également que le score obtenu par la seconde combinaison n'est pas dépendant du choix du seuil car les AUC restent relativement constantes (AUC variant de $0,81$ à $0,83$).

3.3 Discussions

L'aire sous la courbe ROC (AUC) est une bonne indication de la qualité d'une mesure en permettant une évaluation globale des fonctions de rang. Nous proposons d'étudier plus finement la pertinence des premières relations en calculant la précision, car se sont les premières relations qui pourront être prises en compte par l'expert. Nous proposons alors de calculer pour un faible seuil, soit les 1 000 premières relations induites, la précision des approches VW et

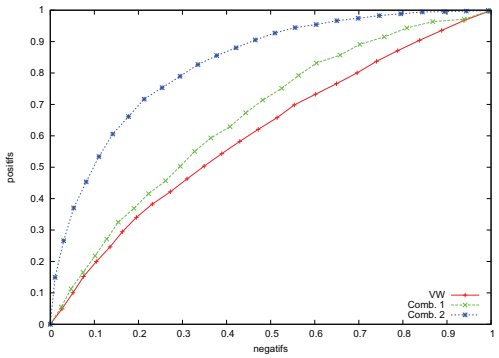

FIG. 3 – *Courbes ROC obtenues avec la validation Web (VW), la première combinaison (Comb. 1) et la seconde combinaison (Comb. 2) au seuil 5 000.*

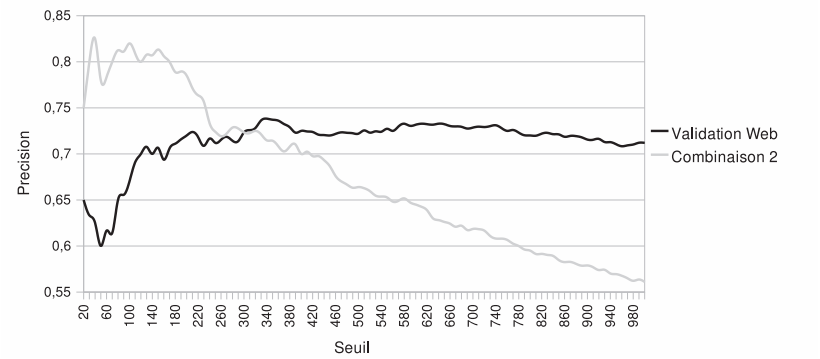

FIG. 4 – *Courbe lift comparant le classement de la VW et de la seconde combinaison.*

la seconde combinaison. La précision qui calcule la proportion de relations induites correctes est définie comme le nombre de relations syntaxiques induites positives divisé par le nombre de relations syntaxiques induites. La figure 4 montre les *courbes d'élévation* ou *courbes lift* (précision en fonction du nombre de relations syntaxiques) des 1 000 premières relations. Une telle courbe permet d'avoir une vue globale de la précision. La figure 4 montre que celle-ci est nettement meilleure avec la seconde combinaison pour les 300 premières relations. Ce résultat signifie que les relations pertinentes sont bien ordonnées en début de liste par cette combinaison et confirme donc les résultats obtenus avec les courbes ROC. Nous avons alors cherché à savoir si les relations syntaxiques placées en tête de liste avec la seconde combinaison étaient les mêmes que celles classées en tête avec la validation Web. Autrement dit, est-ce que les relations privilégiées par la seconde combinaison sont les relations les plus populaires sur le Web ?

Nos expérimentations ont montré que les relations en tête de liste avec la seconde combinaison

ne sont pas les mêmes que celles en tête du classement effectué avec la validation Web. Par exemple, sur les *300* meilleures relations issues de la combinaison 2, avec VW, *37 (12 %)* sont placées entre $[0, 300[$, *39 (13 %)* entre $[300, 600[$, *39 (13 %)* entre $[600, 900[$ et *186 (62 %)* au rang plus élevé. Citons par exemple la relation syntaxique *détenir_arme*, classée 150^{ieme} avec la seconde combinaison qui est classée 1159^{ieme} avec la validation Web. Cela nous indique que la seconde combinaison permet de déterminer des relations moins fréquentes sur le Web, n'étant pas en tête de liste du classement avec la validation Web. Ainsi, nous déterminons et validons des pépites de connaissances, pouvant être plus discriminantes et plus intéressantes que des relations fréquentes qui n'apportent pas d'informations nouvelles.

4 Conclusion

Nous présentons dans cet article des solutions permettant de réduire l'implication humaine dans la validation de relations syntaxiques, qui sont dans notre cas, des relations dites induites. Ces relations ne sont originalement pas présentes dans le corpus, et peuvent permettre par exemple d'enrichir des ontologies. Nous les déterminons en considérant la proximité des verbes et les objets respectifs de ces verbes. Dès lors, nous proposons plusieurs approches afin de proposer à l'expert les relations induites les plus pertinentes. La première approche mesure la proximité sémantique entre le verbe et l'objet de la relation induite en utilisant des vecteurs sémantiques (VS). De tels vecteurs représentent une relation syntaxique comme une combinaison de concepts d'un thésaurus. Une mesure de cosinus est utilisée afin de mesurer la proximité sémantique entre une relation induite et la relation originale dont elle est issue, établissant un classement des relations. Une seconde approche, la validation Web (VW), consiste à interroger un moteur de recherche afin de mesurer la dépendance entre le verbe et l'objet d'une relation induite. Nous utilisons l'*Information Mutuelle au Cube* afin d'ordonner les relations par pertinence. Enfin, nous proposons deux combinaisons entre les approches VS et VW.

Nous avons évalué la qualité de nos méthodes de validation en utilisant des courbes ROC. Nous obtenons de meilleures AUC avec VW plutôt que VS. La seconde combinaison donne de meilleurs résultats que VW, quelque soit le seuil considéré. La mesure de la précision, effectuée pour cette approche, confirme nos AUC plaçant les relations syntaxiques les plus pertinentes en début de liste.

Nous envisageons dans de futurs travaux d'étendre nos approches aux relations syntaxiques *originales* et d'appliquer d'autres mesures de proximité entre un verbe et son objet. Nous envisageons également, afin de mesurer la qualité et la cohérence des concepts formés par nos approches, de les soumettre à un expert. Ces travaux vont de plus être utilisés afin d'améliorer la qualité de l'approche ExpLSA (Béchet et al. (2008)). Cette approche propose d'enrichir un corpus original en effectuant une expansion de contextes par des termes jugés proches sémantiquement. Cet enrichissement se fonde sur une analyse syntaxique préalable. Le fait d'ordonner qualitativement les relations syntaxiques extraites permettra d'améliorer l'approche ExpLSA.

Références

Béchet, N., M. Roche, et J. Chauché (2008). How the ExpLSA approach impacts the document classification tasks. In *Proceedings of the International Conference on Digital Information Management, ICDIM'08*, University of East London, London, United Kingdom,.

Bourigault, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN, Nancy*, pp. 75–84.

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

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

Cilibrasi, R. et P. M. B. Vitanyi (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering 19*, 370.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.

Fabre, C. et D. Bourigault (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *TALN'06, 10-13 avril 2006*, pp. 121–129.

Faure, D. (2000). *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.

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

Jarmasz, M. et S. Szpakowicz (2003). Roget's thesaurus and semantic similarity. In *Conference on Recent Advances in Natural Language Processing*, pp. 212–219.

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

Lin, D. (1998). Extracting collocations from text corpora. In *In First Workshop on Computational Terminology*, pp. 57–63.

Roche, M. et V. Prince (2007). *AcroDef* : A quality measure for discriminating expansions of ambiguous acronyms. In *Proc of CONTEXT, LNCS*, pp. 411–424.

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

Vivaldi, J., L. Màrquez, et H. Rodríguez (2001). Improving term extraction by system combination using boosting. *Proc of ECML, LNCS, 2167*, 515–526.

Wermter, J. et U. Hahn (2004). Collocation extraction based on modifiability statistics. In *COLING '04*, Morristown, NJ, USA, pp. 980. Association for Computational Linguistics.

Wilks, Y. (1998). Language processing and the thesaurus. In *National Language Research Institute*.

Summary

We propose in this paper to use NLP approaches to extract and validate induced syntactical relations (Verb-Object). We first propose a Semantic Vector approach which is a Roget-based approach, computing a syntactic relation as a vector. The second approach is the Web Validation which uses a search engine to determine the relevance of a syntactic relation. We finally propose to combine both methods. We evaluate these approaches by using ROC curves.