Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles

Mouna Kamel, Cassia Trojahn dos Santos

Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles

Mouna Kamel, Cassia Trojahn

IRIT UMR 5505 Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel, cassia.trojahn}@irit.fr

Résumé : L'acquisition automatique de connaissances à partir de textes est un enjeu majeur pour la construction de ressources sémantiques. L'une des tâches cruciales concerne l'identification des relations sémantiques. Cette tâche peut être déclinée en trois phases : l'identification de relations dites candidates, la validation de ces relations et l'intégration de ces relations au sein d'une ressource existante ou en cours de construction. Nous intéressons ici à la validation automatique de relations candidates extraites de structures textuelles spécifiques, les structures énumératives parallèles, en tirant profit de leurs propriétés discursives. L'approche proposée repose sur l'exploitation combinée d'un réseau sémantico-lexicale et une ressource distributionnelle. Les résultats obtenus montrent une exactitude comprise entre 0.5 et 0.67 selon les conditions expérimentales.

Mots-clés : relations sémantiques, validation, structures discursives, réseau lexico-sémantique, ressource distributionnelle

1 Introduction

Le processus d'extraction de relations sémantiques à partir de texte est une tâche cruciale car en amont des processus de construction de ressources sémantiques. Ce processus se fait généralement en trois étapes : repérer les relations candidates dans le texte, valider ces relations, et, pour celles qui ont été validées, les intégrer dans une ressource existante ou en cours de construction. La première étape a fait l'objet de nombreux travaux et de nombreuses approches ont été proposées (approches linguistiques, statistiques, mixtes, avec ou sans apprentissage). Des erreurs peuvent cependant se produire, dues à la stratégie mise en œuvre (patrons lexico-syntaxiques insuffisamment contraints, exactitude des techniques d'apprentissage inférieure à 100% , etc.), ou encore aux différents outils de traitement automatique des langues (ou TAL) appliqués successivement dans les phases de pré-traitement. Aussi, la deuxième étape relative à la validation des relations précédemment identifiées s'avère indispensable. La troisième étape consiste à intégrer les relations validées au sein d'une ressource existante, en s'appuyant, sur des approches d'alignement terminologiques, structurelles ou sémantiques.

Dans cet article nous nous intéressons à la tâche de validation de relations candidates, qui consiste à confirmer ou non ces relations candidates. Plusieurs approches existent, des approches manuelles qui font appel à une expertise humaine, et des approches automatiques qui consistent soit à s'appuyer sur des ressources sémantiques externes, soit à procéder par renforcement si l'on dispose de gros corpus, à l'échelle du web. L'approche que nous proposons s'inscrit dans la continuité de travaux visant à extraire les relations sémantiques, en l'occurrence les relations d'hyperonymie, portées par les structures énumératives parallèles (appelées par la

suite SEP) (Fauconnier & Kamel, 2015). Le choix de s'intéresser aux SEP est motivé par les raisons suivantes : (1) elles sont fréquentes en corpus, notamment dans les textes scientifiques ou encyclopédiques qui sont des textes appropriés pour la construction de ressources ; dans ce type de texte, les SEP sont très souvent exprimées à l'aide de moyens de mise en forme (caractères typographiques et dispositionnels) pour faciliter l'effort cognitif du lecteur - ces SEP sont alors hors de portée des outils classiques de TAL, (2) elles sont souvent porteuses de relations hiérarchiques, relations qui constituent l'ossature des ressources sémantiques (Buitelaar *et al.*, 2005), (3) elles possèdent des propriétés discursives bien établies qui leur confèrent une unité sémantique (Virbel, 1989; Pascual, 1991).

L'approche que nous présentons ici, bien qu'elle s'appuie aussi sur des ressources externes, est originale dans la mesure où elle exploite les propriétés discursives de la SEP, à l'aide de deux types de ressources externes complémentaires, un réseau lexico-sémantique et une ressource distributionnelle. Le réseau lexico-sémantique permet alors de valider les relations spécifiées dans ce réseau, alors que la ressource distributionnelle favorise la spécification de nouvelles relations. De plus, l'unité sémantique dont bénéficie la SEP permet de désambiguïser les relations.

Outre le fait que l'approche que nous proposons a pour objet de valider des relations, elle peut également être utilisée pour valider les systèmes d'extraction de relations à partir de SEP. Par ailleurs, les nouvelles relations validées à l'aide de la ressource distributionnelle constituent alors une source d'enrichissement pour le réseau lexico-sémantique. Enfin, bien qu'implémentée et évaluée pour la langue française, elle reste reproductible pour toute autre langue.

L'article est organisé de la façon suivante. La section 2 fait état des méthodes de validation de relations généralement utilisées. La section 3 définit la SEP, énonce ses propriétés, et donne une représentation de son schéma discursif selon la Rhetorical Structure Theory. La section 4 présente les principes sur lesquels se base l'approche adoptée, ainsi que les algorithmes mis en œuvre. L'application et les résultats de l'évaluation sont décrits dans la section 5, suivis d'une discussion. Nous concluons et proposons quelques perspectives à ce travail en section 6.

2 Etat de l'art

Valider une relation identifiée en corpus consiste à confirmer le sens véhiculé par cette relation dans le domaine de connaissances considéré. Le processus de validation dépend alors de la stratégie d'identification préalablement employée.

Le processus d'identification des relations peut ne nécessiter aucune étape de validation à proprement parler. C'est notamment le cas des approches manuelles qui ont recours à une expertise humaine, comme par exemple Terminae (Biebow & Szulman, 1999) qui confie à un expert la tâche (outillée) de sélectionner, dans une liste de termes candidats, ceux qui dénotent des concepts du domaine, puis de relier ces concepts pour former un réseau de termino-concepts. D'autres approches ont recours à des ressources sémantiques externes qui guident l'annotation des termes dénotant les concepts ou des termes dénotant les relations. Par exemple, l'exploitation conjointe du thesaurus UMLS [1] et de la Gene Ontology [2] en bioNLP permet de caractériser

1. https ://www.nlm.nih.gov/research/umls/
2. http ://geneontology.org/

les interactions entre gènes (Mcdonald *et al.*, 2004). Certaines approches adjoignent des patrons pour caractériser les relations ciblées (Embarek & Ferret, 2007).

L'étape de validation peut constituer une étape à part entière, suite au processus d'identification des relations. Il s'agit dans ce cas, soit de comparer les relations identifiées avec celles existant dans des ressources essentiellement lexicales (Wordnet et Eurowordnet sont largement utilisés à cet effet) (He & Da-You, 2004), soit de comparer les relations identifiées avec celles produites par un groupe d'annotateurs ayant obtenu un taux d'accord correct (> à 0.6 (Carletta, 1996)) suite à une campagne d'annotation (Mukherjee *et al.*, 2014), soit encore de comparer les relations avec un modèle de référence (certains sont disponibles par le biais des campagnes d'évaluation telles que BioNLP Shared Task 2011 (Kim *et al.*, 2011)). Dans ces contextes, les évaluations se font à l'aide de mesures comme *Taxonomy Overlap* (Cimiano *et al.*, 2005), qui compare le chevauchement entre deux taxonomies, ou celle qui mesure les correspondances en termes de *Coverage* (nombre de paires communes à la taxonomie générée et à la ressource correspondante), *Novelty* (nombre de relations correctes mais qui ne sont pas présentes dans la ressource) et *ExtraCoverage* (nombre de relations correctes mais sans correspondances dans la ressource sur le nombre total de relations dans la ressource)(Ponzetto & Strube, 2011).

Un autre type de validation concerne les approches endogènes, celles qui n'ont recours à aucune ressource externe. Une première approche, dite « par renforcement » ou « consolidation » nécessite de gros volumes de données, à l'échelle du web. La validation se base sur la redondance des informations issues de différentes sources, pour pouvoir confronter les résultats produits par différents extracteurs. Ces extracteurs peuvent être basés sur une approche symbolique (extracteur d'infobox, extracteur de tableau, etc.) dans Yago (Suchanek *et al.*, 2007), ou sur des techniques d'apprentissage (extracteur de table, extracteur de texte) dans Nell (Carlson *et al.*, 2010), ou plus précisément sur des classifieurs comme dans Google Knowledge Vault (Dong *et al.*, 2014). Ces approches de consolidation peuvent également mettre en œuvre des règles d'inférence pour détecter les similitudes et les inconsistances logiques (Suchanek *et al.*, 2009). Une autre approche, dite collaborative, permet d'attribuer un taux de confiance aux relations à travers des jeux sérieux (Lafourcade & Zampa, 2009) (voir par exemple « jeu de mots » à http ://www.jeuxdemots.org/).

Toutes ces approches souffrent cependant de limites : la vérification par rapport à des ressources externes se limite aux relations lexicales exprimées dans ces ressources, le recours aux experts ou aux annotateurs est très coûteux, il n'existe pas toujours de modèle de référence dans le domaine considéré, le corpus n'est pas toujours suffisamment volumineux. L'approche que nous proposons, bien qu'actuellement confinée aux SEP, pallie certains de ces inconvénients en n'exigeant aucune expertise humaine, en étant indépendante de la taille du corpus, et en offrant la possibilité de valider des relations qui ne sont pas totalement spécifiées dans les ressources utilisées. Mais avant de présenter l'approche proposée, nous rappelons dans la section suivante les caractéristiques et propriétés des SEP.

3 Structures énumératives parallèles

La structure énumérative (SE) est une structure textuelle ayant la propriété d'exprimer des connaissances hiérarchiques au travers de différents composants : une amorce, une liste d'items (au moins deux) constituant l'énumération, et éventuellement une clôture qui, lorsqu'elle existe, synthétise les différentes propositions exprimées à travers les items. Sur le plan sémantique, la

SE forme un tout. Sur le plan de la mise en forme, elle peut être exprimée selon différents modes, allant d'une forme linéaire sans mise en forme (Figure 1(a)) ou usant de caractères typographiques (Figure 1(b)), à une forme non linéaire usant de dispositifs typographiques et dispositionnels (Figure 1(c)).

(a) Le dromadaire a été répertorié dans 35 pays, tels que l'Inde, la Turquie, le Kenya, le Pakistan, la corne de l'Afrique et bien d'autres encore.

(b) Certaines plantes (palmiers, bambous. . .) produisent des tissus lignifiés.

(c) Une chaussure se compose principalement :
- du semelage, partie qui protège la plante des pieds
- de la tige, partie supérieure qui enveloppe le pied

FIGURE 1 – Différentes formes d'expression de la structure énumérative parallèle.

Plusieurs définitions de l'énumération existent, qui s'accordent sur l'égalité d'importance entre les différentes entités énumérées, par rapport à un critère de recensement (Dammame-Gilbert, 1989; Pascual, 1991). C'est toutefois la définition de Virbel (Virbel, 1989) qui semble le mieux prendre en compte à la fois les phénomènes architecturaux du texte et l'intention de l'auteur : "énumérer mobilise deux actes : un acte mental d'identification des éléments d'une réalité du monde dont on vise un recensement, et où on établit une relation d'égalité d'importance par rapport au motif de recensement ; et un acte textuel qui consiste à transposer textuellement la coénumérabilité des entités recensées, par la coénumérabilité des segments linguistiques qui les décrivent.".

La SE a également fait l'objet de nombreuses études au cours desquelles différentes typologies ont pu être proposées. Les SE linéaires ont été essentiellement analysées dans le cadre de l'analyse du discours. Elles ont donné lieu à des typologies comme celle de (Vergez-Couret *et al.*, 2008) où les SE à un temps ont été opposées aux SE à deux temps, ou encore comme celle de (Ho-Dac *et al.*, 2010) où les SE ont été classifiées selon leur niveau de granularité (SE dont les items sont des titres, SE en tant que listes formatées, SE multi-paragraphiques sans marque visuelle, SE intra-paragraphiques). Les SE usant de dispositifs typo-dispositionnels (dites verticales) ont quant à elles été notamment analysées dans le cadre de la génération de texte. (Hovy & Arens, 1998) distinguent les listes d'items (ensemble de composants de même niveau) des listes énumérées (pour lesquelles l'ordre des composants est pris en compte), alors que (Christophe, 2000) propose une typologie qui oppose les SE parallèles (paradigmatiques, homogènes visuellement et isolées) aux SE non parallèles. Cette dernière typologie est basée sur la composition du modèle rhétorique de la Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) et du Modèle d'Architecture Textuelle (MAT) (Virbel, 1989).

Nous nous intéressons ici aux relations portées par les structures énumératives parallèles (SEP) au sens de Luc (2000), c'est à dire les SE pour lesquelles les items de l'énumération sont tous fonctionnellement équivalents (du point de vue syntaxique et rhétorique) et indépendants dans un contexte donné. Les exemples donnés ci-dessus correspondent à des SEP. D'un point de vue discursif, si la segmentation est faite en fonction des composants de la SE, c'est à dire que l'amorce et chacun des items correspondent à des unités de discours (UD), alors les UD relatives aux items sont successivement reliées par une relation rhétorique multi-nucléaire (ou

coordination), et l'UD relative à l'amorce est reliée à l'UD relative au premier item par une relation de type noyau-satellite (ou subordination). La figure 2 décrit le schéma discursif de la SEP selon la RST (Rhetorical Structure Theory).

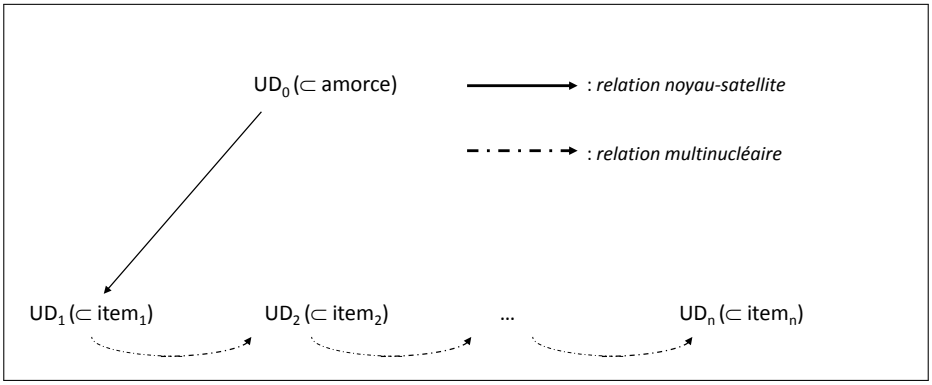

FIGURE 2 – Représentation discursive de la SE parallèle selon la RST (UD = Unité de Discours)

Selon les théories du discours (Asher, 1993), si " UD_j est subordonnée à UD_i, alors toute UD_k coordonnée à UD_j est subordonnée à UD_i". Ainsi, *N* relations de type noyau-satellite entre UD_0 et UD_i, pour *i=1,...,N* (si *N* est le nombre d'items dans la SE) peuvent être inférées. Ces *N* relations peuvent alors être spécialisées en *N* relations sémantiques $R(H, h_i)_{i=1,...,N}$ **de même nature**, où H correspond à un terme de UD_0, et h_i à un terme de UD_i.

Une SEP peut cependant porter plus de relations hiérarchiques, notamment lorsqu'un item est lui-même composé d'une énumération. Il est à noter qu'une SEP peut également porter des relations non hiérarchiques, dans ce cas généralement exprimées au sein de l'amorce ou d'un item. L'exemple de la Figure 3 illustre ces cas.

Les principaux actes cultuels sont :
- le sacrifice, la libation, l'offrande et l'éducation ;
- la prière (invocation, louange, demande, etc.) ;
- le chant et la musique ;
- la lecture de textes sacrés ;
- la prédication qui a un rôle important dans les religions abrahamiques ;
- les pèlerinages, processions.

FIGURE 3 – Exemple de SE où les items comportent des énumérations.

Cette SEP, outre le fait qu'elle intègre des SE sous diverses formes, montre que plusieurs relations hiérarchiques peuvent exister entre l'amorce et le premier item par exemple (*acte cultuel* et *sacrifice*, *acte cultuel* et *libation*, etc.), ainsi qu'une relation syntagmatique exprimée dans l'avant dernier item (*prédication* et *religions abrahamiques*).

Le travail proposé ici s'inscrit dans la continuité des travaux présentés dans (Fauconnier & Kamel, 2015), qui consistent à repérer les potentielles relations d'hyperonymie $R(H, h_i)_{i=1,...,N}$

au sein d'une SEP, à l'aide d'un système d'apprentissage supervisé. Dans ce contexte, la structure discursive de la SEP a été exploitée pour faire valoir la présence d'une relation hiérarchique entre un terme présent dans l'amorce et des termes présents dans les items (un terme par item). L'approche de validation que nous proposons ici exploite également la structure discursive des SEP pour faire valoir cette fois la proximité sémantique, notamment en termes de cohésion lexicale, entre les h_i (*i=1,...,N*) liés par une même relation à une même entité H.

4 Approche proposée

Les erreurs d'identification de relations par les systèmes d'extraction automatique de relations à partir de texte sont généralement dues soit à une mauvaise interprétation de la nature de la relation (difficulté pour les systèmes à résoudre certains phénomènes linguistiques comme les anaphores, les ellipses, etc.), soit à la mauvaise identification des arguments (difficultés à extraire les bons termes notamment). Une autre cause peut être la mauvaise formalisation des connaissances de l'auteur du texte. Il s'agit alors de ne valider que les relations qui portent sens.

4.1 Principe général

Le principe de validation que nous mettons en œuvre exploite les propriétés discursives de la SEP pour valider conjointement (et non indépendamment les unes des autres) les relations $R(H, h_i)$ (i=1,..,N) issues d'une même SEP, et où R correspond à la relation d'hyperonymie, H à l'hyperonyme, et h_i à l'hyponyme. Nous utilisons pour cela deux ressources sémantiques : un réseau lexico-sémantique qui fournit en général une bonne précision mais dont le taux de couverture est variable, et une ressource distributionnelle qui ne spécifie pas les relations entre termes mais dont la couverture est très large. Le principe de validation, appliqué à une SEP, se déroule en deux étapes :

1. valider toutes les relations $R(H, h_i)$ qui sont exprimées dans le réseau sémantique, avec un coefficient de 1.
2. valider toutes les relations $R(H, h_k)$ qui ne sont pas exprimées dans le réseau sémantique en évaluant le coefficient de proximité distributionnelle qu'entretient h_k avec tous les h_i appartenant aux relations validées à l'étape 1.

Nous décrivons ci-dessous le principe algorithmique de façon plus détaillée.

4.2 Algorithme

Nous donnons les définitions suivantes :
— RS le réseau sémantique, et RD la ressource distributionnelle ;
— R_i une relation candidate d'hyperonymie issue d'une SEP et liant l'hyperonyme H détecté dans l'amorce, à l'hyponyme h_i identifié dans le i^{me} item (*i=1,..., N* si *N* est le nombre d'items présents dans la SEP) ;
— ***Lexicalisation(C)*** l'ensemble des termes associés au concept C au sein de RS
— $Synset(H) = \{C \in RS / H \in Lexicalisation(C)\}$;
— $Hyperonymes_{RS}(h_i)$ l'ensemble des hyperonymes directs de h_i dans RS

— $SuperHyperonymes_{RS}(h_i)$ l'ensemble des hyperonymes de h_i existant dans RS et liés à h_i par un chemin de longueur inférieur ou égal à un seuil S fixé de façon empirique selon le réseau sémantique utilisé :
$SuperHyperonymes_{RS}(h_i) = \bigcup_{k=1}^{S} SuperHyperonymes_{RS}^{k}(h_i)$ où
$SuperHyperonymes_{RS}^{k}(h_i)$ est l'ensemble des hyperonymes de h_i de rang k (k étant la longueur maximale du chemin reliant h_i à un de ses hyperonymes dans RS), et où
$SuperHyperonymes_{RS}^{k}(h_i)$ est défini récursivement par :
$SuperHyperonymes_{RS}^{1}(h_i) = \{hyperonymes_{RS}(h_i)\}$;
$SuperHyperonymes_{RS}^{k}(h_i) = SuperHyperonymes_{RS}^{k-1}(h_i) \bigcup \bigcup_{h \in SuperHyperonymes_{RS}^{k-1}(h_i)} hyperonymes_{RS}(h)$.

Nous donnons la description complète de l'algorithme ci-dessous (Algorithm 1).

Algorithm 1 Principe de validation des relations issues d'une SEP

V est l'ensemble des relations validées
$\overline{V}$ est l'ensemble des relations non validées R_i
% au départ, toutes les relations ont le statut de "non validée" %
$V \leftarrow \emptyset$
$\overline{V} \leftarrow \bigcup_{i=0}^{N} R_i$
Pour chaque relation $R_i(H, h_i) \in \overline{V}$ **Faire**
  Si $SuperHyperonymes_{RS}(h_i) \cap Synset(H) \neq \emptyset$ **alors**
    $Valider(R_i(H, h_i)) \leftarrow 1$
    $V = V \bigcup \{R_i\}$
    $\overline{V} = \overline{V} - \{R_i\}$
  Fin Si
Fin Pour
% si au moins une des relations a obtenu le statut de "validée" (ce qui confirme l'hyperonymie), et si au moins une relation possède toujours le statut de "non validée" (ce qui amène à exploiter la ressource distributionnelle)%
Si $V \neq \emptyset$ *et* $\overline{V} \neq \emptyset$ **alors**
  Pour chaque relation $R_k(H, h_k) \in \overline{V}$ **Faire**
    $valider(R_k) = \frac{\sum_{R_i \in V} p(h_i, h_k)}{|V|}$ *% où p(h_i,h_k) correspond à la proximité sémantique fournie par RD %*
    % calcul de la somme des mesures de proximité entre l'hyponyme de la relation non validée h_k et les hyponymes h_i des relations validées %
  Fin Pour
Fin Si

5 Application et évaluation

5.1 Jeu de données

Le jeu de données que nous avons utilisé pour notre expérimentation est constitué de 262 relations candidates d'hyperonymie fournies par le système d'extraction de relations décrit dans (Fauconnier & Kamel, 2015) (comme dit précédemment, l'approche de validation que nous proposons s'inscrit dans la suite de ces travaux). Ce système extrait par apprentissage supervisé des relations d'hyperonymie à partir de SEP, ces SEP ayant été elles-mêmes extraites automatiquement de pages Wikipedia. En effet, les articles Wikipédia sont rédigés selon le guide "the Manual of Style" (`http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`) qui préconise l'utilisation de SEP et recommande pour ces structures d'utiliser la même forme grammaticale pour tous les items. Une validation de ces structures énumératives en tant que parallèles avait été menée manuellement.

Les 262 relations composant notre jeu de données proviennent donc de 67 de ces SEP. Ces relations ont fait l'objet d'une validation manuelle effectuée par deux annotateurs en double aveugle, à la suite de laquelle 27 conflits inter-annotateurs ont été identifiés et résolus (i.e., les deux annotateurs se sont *a posteriori* mis d'accord). Les désaccords étaient principalement liés au fait que certaines relations pouvaient être caractérisées d'hyperonymie ou d'holonymie. Ainsi 206 relations ont été évalués comme étant correctes et 56 comme incorrectes. Ces relations constituent alors un ensemble de relations de référence.

5.2 Ressources utilisées

Nous avons utilisé le réseau lexico-sémantique *BabelNet* (Navigli & Ponzetto, 2012) et la ressource distributionnelle *Voisins de Wikipédia*[3] (Adam *et al.*, 2013). Le réseau *BabelNet* a été construit automatiquement à partir de l'intégration de plusieurs ressources (WordNet, Open Multilingual WordNet, Wikipedia, GeoNames, WoNef, etc.). Il est composé d'environ 14 millions d'entrées, incluant concepts et entités nommées, chaque entrée définissant un *BabelSynset*. Chaque *BabelSynset* correspond à un sens donné (*BabelSense*) et regroupe tous les synonymes dans 271 langues différentes, dont la langue française. La ressource *Voisins de Wikipédia* a éte construite à partir d'un corpus de 262 millions de mots, selon les principes décrits par (Bourigault, 2002) à partir d'un modèle structuré (Baroni & Lenci, 2010).

Ces ressources ont été choisies d'une part car elles expriment des connaissances en langue française, et d'autre part car elles ont pour origine le même corpus que celui dont est issu le jeu de données.

5.3 Conditions de l'expérimentation

Nous avons fixé de façon empirique la longueur maximale du chemin reliant un hyponyme à l'un de ses hyperonymes à k=3. Par ailleurs, les mesures de confiance accordées aux relations validées par le système ont été calculées de la façon suivante :

— nous avons considéré que toute relation entre un hyperonyme et un de ses hyponymes exprimée dans *BabelNet* a une mesure de confiance égale à 1.

3. http ://redac.univ-tlse2.fr/applications/vdw.html

— les mesures de confiance entre deux hyponymes sont celles données par la ressource distributionnelle *Voisins de Wikipedia*. Ces mesures correspondant à des scores de Lin (Lin, 1998) compris entre 0,1 et 0,29 pour 97% des entrées (Adam *et al.*, 2013).

5.4 Résultats

Nous avons expérimenté et évalué notre approche sur deux ensembles de relations candidates :

— E l'ensemble des relations du *gold standard* (206 relations)

— E_{BN} l'ensemble des relations dont l'hyperonyme possède une entrée dans *BabelNet* (116 relations)

L'évaluation a été menée en termes de précision, rappel et exactitude. Ces résultats sont présentés dans la Table 1.

	Précision	Rappel	F-Measure	Exactitude
E	.97	.37	.54	.50
E_{BN}	.97	.66	.78	.67

TABLE 1 – Résultats de la validation des relations candidates.

Nous avons obtenu le même taux de précision pour les deux ensembles E et E_{BN} : 76 relations parmi les 78 relations automatiquement validées sont correctes. Comme attendu, nous avons obtenu de moins bons résultats en termes de rappel. Pour l'ensemble E, 76 relations ont été validées sur les 206 correctes. Les résultats pour l'ensemble E_{BN} sont meilleurs, 76 relations ont été validées sur 116 correctes. En termes d'exactitude, pour l'ensemble E, parmi les 262 relations (impliquant les relations annotées comme correctes et incorrectes par les annotateurs), 130 ont été correctement validées par le système. Pour l'ensemble E_{BN}, 88 relations ont été correctement confirmées, sur un total de 131.

Il est à noter que 12 relations parmi les 76 validées par le système, ont été validées par la ressource distributionnelle, et sont correctes. Dans ce contexte, la ressource distributionnelle a validé nos relations avec une précision de 1.0, et a permis d'améliorer les performances de notre système à hauteur de 15%.

5.5 Discussion

Bien que la précision soit très haute, nous avons pu identifier dans quel cas notre système validait une relation fausse : cela est dû au fait d'utiliser des BabelSynsets qui regroupent des termes de sens proche. Par exemple, lors de l'évaluation de la relation candidate $R(pays, Corne\ de\ l'Afrique)$, le BabelSynset $bn\ :\ 00028934n$ composé par les BabelSenses {terre, sol, terre ferme, contrée, pays} appartient à l'intersection des ensembles $SuperHyperonymes^{3}_{BN}(Corne\ de\ l'Afrique)$ et $Synset(pays)$.

Pour ce qui concerne le rappel, nous avons identifié deux causes au silence. La première cause est une conséquence de l'absence de l'hyperonyme comme entrée dans *BabelNet* (e.g., 'feuillus colonisateur' ou 'poisson sauvage'). Dans ce cas, aucune relation de la SEP porteuse de cet hyperonyme n'est validée. Cela concerne 62 relations pour l'ensemble E. La deuxième

cause provient du fait que la valeur fixée à 3 pour la profondeur du chemin de recherche des hyperonymes de l'hyponyme h_i n'est quelquefois pas suffisante. Augmenter alors la valeur de k permettrait d'améliorer le rappel, mais nécessiterait de définir des heuristiques pour conserver une complexité algorithmique acceptable.

La ressource distributionnelle permet d'identifier des relations non exprimées dans le réseau lexico-syntaxique. Par exemple, la relation *R(anomalie chromosomique,insertion)* absente de *BabelNet* a pu être validée, par le fait que *R(anomalie chromosomique,délétion)* est présente dans *BabelNet*, et que les termes *insertion* et *délétion* sont sémantiquement proches dans la ressource distributionnelle. Bien que les entrées de cette ressource ne correspondent très majoritairement qu'à des mots simples, et non à des termes, alors que 40% des hyponymes de relations que nous avions à valider correspondent à des termes composés de plus d'un mot, les performances par rapport à l'exploitation du réseau lexico-syntaxique seul ont pu être améliorées de 15%. Une ressource distributionnelle intégrant les termes permet d'envisager d'améliorer encore ces résultats.

6 Conclusion et Perspectives

Ce travail, qui s'inscrit dans la continuité de travaux entrepris sur l'identification de relations d'hyperonymie portées par les structures énumératives parallèles, a pour objet de proposer une méthode de validation automatique de ces relations. L'originalité de l'approche proposée tient au fait qu'elle exploite la structure discursive de la structure énumérative parallèle, et qu'elle met l'accent sur la complémentarité entre deux ressources de nature différente, un réseau lexico-sémantique et une ressource distributionnelle. L'approche a été évaluée sur les SEP, avec une exactitude comprise entre 0.5 et 0.67 selon les conditions expérimentales, et avec une amélioration de 15% des performances due à l'exploitation conjointe de la ressource distributionnelle. Cette approche vaut pour tout objet textuel ayant même schéma discursif que la SEP, comme les titres et sous-titres, les champs de formulaire, etc.

Une des premières suites que nous prévoyons à ce travail est de voir comment l'exploitation des ressources externes peut conduire à améliorer notre système. Plusieurs pistes sont envisagées : ressources distributionnelles intégrant les termes (idéalement du domaine), analyser le compromis entre étendue des recherches au sein du réseau sémantique vs. complexité algorithmique, d'autres réseaux sémantiques voire utiliser conjointement plusieurs réseaux lexico-syntaxiques et plusieurs ressources distributionnelles.

Une autre suite envisagée concerne l'adaptation et l'évaluation de notre approche pour des relations lexicales autres que l'hyperonymie, comme la méronymie, la synonymie et l'antonymie.

Remerciements

Cassia Trojahn est partiellement financée par le projet FUI SparkinData.

Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *TAL*, **54**(1), 71–97.

ASHER N. (1993). *Reference to Abstract Objects in Discourse : A Philosophical Semantics for Natural Language Metaphysics*, volume 50 of *SLAP*. http ://www.wkap.nl/ : Kluwer.

BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Comput. Linguist.*, **36**(4), 673–721.

BIEBOW B. & SZULMAN S. (1999). Terminae : A linguistic-based tool for the building of a domain ontology. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '99, p. 49–66, London, UK, UK : Springer-Verlag.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, TALN'2002, p. 75–84.

BUITELAAR P., CIMIANO P. & MAGNINI B. (2005). *Ontology Learning from Text : An Overview*, In P. BUITELAAR, P. CIMIANO & B. MAGNINI, Eds., *Ontology Learning from Text : Methods, Evaluation and Applications*, volume 123. IOS Press.

CARLETTA J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Comput. Linguist.*, **22**(2), 249–254.

CARLSON A., BETTERIDGE J., KISIEL B., SETTLES B., JR. E. R. H. & MITCHELL T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

CHRISTOPHE L. (2000). *Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés*. PhD thesis.

CIMIANO P., HOTHO A. & STAAB S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.*, **24**(1), 305–339.

DAMMAME-GILBERT B. (1989). *La série énumérative : étude linguistique et stylistique s' appuyant sur dix romans français publiés entre 1945 et 1975*, volume 19. Librairie Droz.

DONG X., GABRILOVICH E., HEITZ G., HORN W., LAO N., MURPHY K., STROHMANN T., SUN S. & ZHANG W. (2014). Knowledge vault : A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, p. 601–610, New York, NY, USA : ACM.

EMBAREK M. & FERRET O. (2007). Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical. In *Proceedings of 14ème Conférence sur le Traitement automatique des langues naturelles (TALN 2007)*, p. 37–46.

FAUCONNIER J.-P. & KAMEL M. (2015). Discovering Hypernymy Relations using Text Layout (regular paper). In *Joint Conference on Lexical and Computational Semantics (SEM), Denver, Colorado, 04/06/15-05/06/15*, p. 249–258, http ://www.aclweb.org : Association for Computational Linguistics (ACL).

HE H. & DA-YOU L. (2004). Learning owl ontologies from free texts. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on (Volume :2)*, p. 1233–1237 : IEEE Conference Publications.

HO-DAC L.-M., PÉRY-WOODLEY M.-P. & TANGUY L. (2010). Anatomie des Structures Énumératives. In *Traitement Automatique des Langues Naturelles*, p. (publication numérique), Montréal, Canada.

HOVY E. H. & ARENS Y. (1998). Readings in intelligent user interfaces. chapter Automatic Generation of Formatted Text, p. 256–262. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

KIM J.-D., PYYSALO S., OHTA T., BOSSY R., NGUYEN N. & TSUJII J. (2011). Overview of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, p. 1–6, Stroudsburg, PA, USA : Association for Computational Linguistics.

LAFOURCADE M. & ZAMPA V. (2009). Pticlic : a game for vocabulary assessment combining jeux-demots and lsa. In *In proc of CICLing (Conference on Intelligent text processing and Comptational*

Linguistics). Mexico : Marsh 1-7.

LIN D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, p. 296–304, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

MCDONALD D. M., CHEN H., SU H. & MARSHALL B. B. (2004). Extracting gene pathway relations using a hybrid grammar : the arizona relation parser. *Bioinformatics*, p. 3378.

MUKHERJEE S., AJMERA J. & JOSHI S. (2014). Domain cartridge : Unsupervised framework for shallow domain ontology construction from corpus. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, p. 929–938.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

PASCUAL E. (1991). *Représentation de lárchitecture textuelle et génération de texte*. PhD thesis.

PONZETTO S. & STRUBE M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. volume 9 of *175*, p. 1737–1756.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706, New York, NY, USA : ACM.

SUCHANEK F. M., SOZIO M. & WEIKUM G. (2009). Sofie : a self-organizing framework for information extraction. In J. QUEMADA, G. LEÓN, Y. S. MAAREK & W. NEJDL, Eds., *WWW*, p. 631–640 : ACM.

VERGEZ-COURET M., PRÉVOT L. & BRAS M. (2008). Interleaved discourse, the case of two-step enumerative structures. p. 85–94 : Proceedings of Contraints In Discourse III, Postdam.

VIRBEL J. (1989). Structured documents. chapter The Contribution of Linguistic Knowledge to the Interpretation of Text Structures, p. 161–180. New York, NY, USA : Cambridge University Press.