Extraction de relations d'hyperonymie à partir de Wikipédia

Adel Ghamnia[1]
(1) IRIT, Avenue de l'étudiant, 31400 Toulouse, France
adel.ghamnia@irit.fr

RÉSUMÉ

Ce travail contribue à montrer l'intérêt d'exploiter la structure des documents accessibles sur le Web pour enrichir des bases de connaissances sémantiques. En effet, ces bases de connaissances jouent un rôle clé dans de nombreuses applications du TAL, Web sémantique, recherche d'information, aide au diagnostic, etc. Dans ce contexte, nous nous sommes intéressés ici à l'identification des relations d'hyperonymie présentes dans les pages de désambiguïsation de Wikipédia. Un extracteur de relations d'hyperonymie dédié à ce type de page et basé sur des patrons lexico-syntaxiques a été conçu, développé et évalué. Les résultats obtenus indiquent une précision de 0.68 et un rappel de 0.75 pour les patrons que nous avons définis, et un taux d'enrichissement de 33% pour les deux ressources sémantiques BabelNet et DBPédia.

ABSTRACT

Hypernym extraction from Wikipédia

The volume of available documents on the Web continues to increase, the texts contained in these documents are rich information describing concepts and relationships between concepts specific to a particular field. In this paper, we propose and exploit an hypernymy extractor based on lexico-syntactic patterns designed for Wikipedia semi-structured pages, especially the *disambiguation pages*, to enrich a knowledge base as BabelNet and DBPedia. The results show a precision of 0.68 and a recall of 0.75 for the patterns that we have defined, and an enrichment rate up to 33% for both BabelNet and DBPédia semantic resources.

MOTS-CLÉS : Extraction de relations d'hyperonymie, Base de connaissances, Patrons morpho-syntaxiques.

KEYWORDS: Hypernym extraction, Knowledge Base, morpho-syntactic patterns.

1 Introduction

L'objectif de notre travail est l'enrichissement de bases de connaissances sémantiques de type BabelNet (Navigli & Ponzetto, 2012) ou DBPédia (Lehmann *et al.*, 2014) à partir des informations contenues dans des documents textuels semi-structurés. Ces bases de connaissances jouent aujourd'hui un rôle clé dans de nombreuses applications du TAL, et leur alimentation constitue donc un enjeu important afin de rendre disponibles des informations lexico-sémantiques multilingues à large échelle. A l'heure actuelle, la construction de ces réseaux se fonde principalement sur des ressources existantes telles WordNet ou sur l'exploitation de la partie structurée des documents encyclopédiques de

Wikipédia [1]. Ainsi, des extracteurs dédiés se focalisent sur les infobox, les catégories, ou les liens définis dans les pages Wikipédia (Morsey *et al.*, 2012; Lehmann *et al.*, 2014). Les contenus textuels des documents, riches en information décrivant des concepts et des relations entre ces concepts, mais plus difficilement accessibles, sont généralement sous-exploités.

Différentes méthodes ont cependant été définies pour extraire à partir des textes des informations (termes et relations sémantiques entre termes) susceptibles d'alimenter ces bases de connaissances. Ces travaux utilisent généralement des extracteurs de termes et font appel à des techniques fondées sur l'application de patrons morpho-syntaxiques dans la lignée de (Hearst, 1992), sur le principe de proximité distributionnelle (Lenci & Benotto, 2012), ou sur l'exploitation de structures textuelles spécifiques, par exemple les définitions (Malaisé *et al.*, 2004) ou les structures énumératives (Fauconnier & Kamel, 2015).

Notre travail de recherche vise à enrichir le Web des données pour le français en mettant en oeuvre de façon combinée plusieurs méthodes d'extraction de termes et de relations entre termes à partir des textes, pour l'acquisition de différents types de relations sémantiques, en premier lieu l'hyperonymie et la méronymie. Comme il a été montré par exemple par (Schropp *et al.*, 2013), la combinaison de plusieurs approches est une piste intéressante pour tirer parti de la multiplicité des indices textuels signalant une relation sémantique, et dépasser les limites identifiées pour chaque méthode. Notre approche s'appuiera sur un corpus issu de Wikipédia en français, dont les articles ont la particularité de combiner différents niveaux de structuration de l'information textuelle.

Nous présentons dans cet article une étape préliminaire de ce travail, qui vise à tester la démarche à partir d'un premier cas d'étude. Notre premier objectif est de déterminer la plus-value potentielle de l'extraction de relations à partir de textes, en évaluant l'apport d'une première expérience d'extraction pour l'alimentation des bases DBPédia et BabelNet. Dans le cadre de cet article, nous nous focalisons sur l'extraction de la relation d'hyperonymie à partir de certains articles de Wikipédia appelés pages de désambiguïsation, et nous nous limitons à la démarche classique d'extraction par patrons lexico-syntaxiques. Nous amorçons notre approche en choisissant un cas de figure favorable, puisque les pages de désambiguïsation de Wikipédia sont riches en entités nommées et en relations d'hyperonymie exprimées dans des structures textuelles contraintes, généralement des énumérations, et normées par les consignes de la charte de rédaction. Cela nous permet de concevoir une liste de patrons lexico-syntaxiques adaptés à ce type de pages, qui couvrent à la fois la relation sémantique et les arguments de cette relation, en tirant parti de la structure particulière de ces pages.

Dans ce qui suit, nous proposons tout d'abord un état de l'art des méthodes d'extraction des relations d'hyperonymie à partir de textes. Nous présentons ensuite notre corpus et les annotations réalisées. La troisième partie décrit la méthode d'extraction de relations, avant une section consacrée à l'évaluation de la méthode, à la fois intrinsèque (performance des patrons) et extrinsèque (apport pour l'alimentation des bases sémantiques).

2 Travaux précédents

Dans cette partie, nous faisons une courte synthèse des travaux réalisés sur l'extraction de la relation d'hyperonymie, et plus particulièrement sur l'extraction de relations à partir de Wikipédia.

1. fr.wikipedia.org

2.1 Extraction de relations d'hyperonymie

De nombreux travaux ont été consacrés à l'extraction automatique de la relation d'hyperonymie. On peut les organiser en deux grands types de démarches. Une première série de travaux est inspirée du travail pionnier de Hearst (1992), qui a montré que la relation d'hyperonymie est directement signalée dans les textes dans des constructions régulières, et peut être extraite par la projection de patrons lexico-syntaxiques reliant deux termes. Ce travail a inspiré de nombreux travaux qui ont progressivement intégré des méthodes d'apprentissage pour limiter le coût de construction des patrons (Morin & Jacquemin, 2004; Snow *et al.*, 2004; Pantel & Pennacchiotti, 2006). Le travail de (Panchenko *et al.*, 2013) a réadapté et évalué les patrons de Hearst sur des gros corpus en français. (Panchenko *et al.*, 2016; Bordea *et al.*, 2015) ont utilisé une méthode d'extraction de relations basée sur des patrons pour reconstruire des taxonomies existantes à partir d'une liste finie de termes. Ces travaux ont montré que l'approche par patrons produit des résultats généralement satisfaisants du point de vue de la précision. Néanmoins, comme ils n'exploitent qu'une partie infime de l'information textuelle, leur couverture est extrêmement réduite.

Un deuxième type d'approche est fondé sur l'hypothèse distributionnelle, qui consiste à établir une relation de proximité entre deux unités lexicales qui présentent des propriétés distributionnelles semblables. On retrouve dans cette catégorie les travaux de (Caraballo, 2001) ou (Van Der Plas *et al.*, 2005). On sait que l'approche distributionnelle a pour effet de mettre au jour une relation de proximité sémantique au sens large, mêlant indifféremment des relations d'hyperonymie, synonymie, co-hyponymie, ou des relations de proximité plus lâche. Il s'agit donc d'identifier au sein des espaces distributionnels des relations de voisinage qui présentent des propriétés spécifiques, par exemple en se fondant, comme le font Lenci & Benotto (2012), sur une hypothèse d'inclusion distributionnelle, les traits distributionnels de l'hyponyme étant inclus dans ceux de l'hyperonyme.

D'autres travaux explorent encore des voies complémentaires, en tirant parti de caractéristiques structurelles ou dispositionnelles des textes, afin de repérer des zones denses en relations sémantiques, en particulier les structures énumératives (Fauconnier & Kamel, 2015). Une technique plus simple, dite d'inclusion lexicale, est également utilisée pour tirer parti de la structuration interne des textes (Lefever *et al.*, 2014), la tête lexicale d'un terme complexe pouvant être considérée comme l'hyperonyme du terme complet.

Ces approches sont complémentaires : les unes s'intéressent aux segments dans lesquels les deux termes cooccurrent dans des contextes spécifiques, d'autres au contraire tirent parti de l'ensemble des occurrences des termes dans le corpus, ou de leurs propriétés formelles. La combinaison des différents types d'approches s'avère de fait prometteuse, comme démontré par (Schropp *et al.*, 2013).

2.2 Extraction de relations à partir de Wikipédia

DBPédia est une base de connaissances qui contient des concepts et relations extraits de Wikipédia. Morsey *et al.* (2012) ont développé 19 extracteurs, chacun étant dédié au traitement d'un type particulier d'information structurée au sein des pages Wikipédia : infobox, catégorie, lien, image, etc. La figure 1 montre la façon dont les pages Wikipédia font coexister des structures directement accessibles par ce type d'extracteurs (par exemple, les liens entre entités), et du contenu textuel, dont l'essentiel n'est pas exploité.

Ainsi, les travaux d'extraction de relations d'hyperonymie à partir de Wikipédia ciblent particuliè-

FIGURE 1 – Une page Wikipédia

rement les parties structurées. Par exemple, Suchanek *et al.* (2007) ont utilisé les catégories, qui constituent le système de classement thématique de Wikipédia ; Kazama & Torisawa (2007) ont exploité la partie définition ; enfin Sumida & Torisawa (2008) se sont intéressés aux menus, qui offrent un moyen d'accéder à la hiérarchie des concepts.

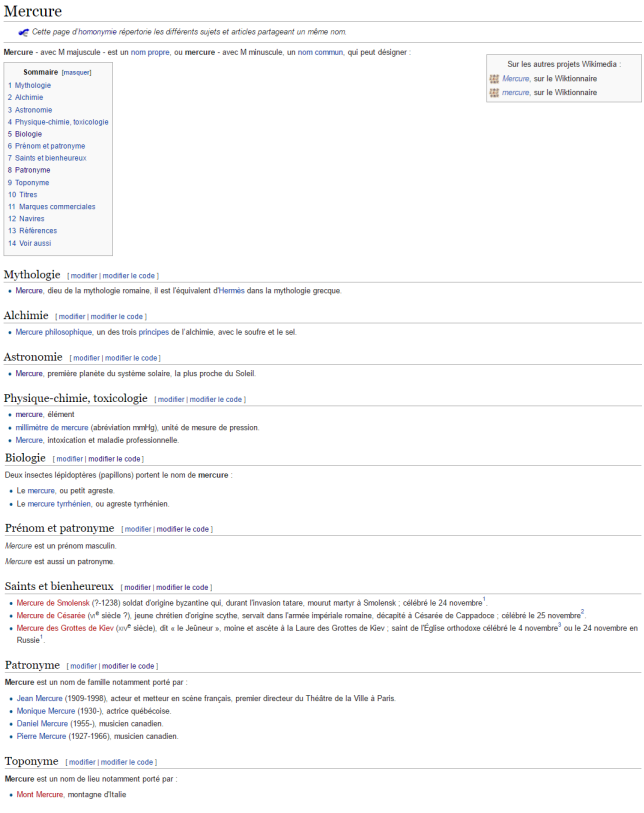
3 Corpus et données d'annotation

Nous avons constitué un corpus à partir du *dump* de la version française de l'encyclopédie Wikipédia. Nous avons choisi de nous focaliser dans ce travail sur un type particulier de pages, appelées pages de désambiguïsation (appelée aussi page d'homonymie). Ces pages listent les articles dont le titre est ambigu, et donnent une définition de toutes les acceptions recensées. Ces pages sont riches en relations d'hyperonymie et ont une structure régulière. Par exemple, la page d'homonymie *Mercure* cite plusieurs articles, parmi lesquels :

— le dieu romain Mercure ;
— la planète Mercure ;
— l'élément chimique mercure ;

Une page de désambiguïsation est structurée en plusieurs rubriques correspondant à différents types d'homonymie. Deux d'entre elles, consacrées au recensement des patronymes et des toponymes, ont une structure régulière, comme illustré dans la figure 2. Ces deux rubriques doivent être rédigées selon un *template* proposé par Wikipédia. Cette normalisation est généralement adoptée par le rédacteur, ce qui nous a permis de définir des patrons spécifiques pour le repérage des relations d'hyperonymie.

Afin d'évaluer notre approche, nous avons constitué un sous-corpus composé de 30 pages de désambiguïsation, tirées aléatoirement du corpus. Le tableau 1 détaille les 30 pages utilisées et le nombre



Mercure

Cette page d'homonymie répertorie les différents sujets et articles partageant un même nom.

Mercure - avec M majuscule - est un nom propre, ou **mercure** - avec M minuscule, un nom commun, qui peut désigner :

Sommaire [masquer]
1 Mythologie
2 Alchimie
3 Astronomie
4 Physique-chimie, toxicologie
5 Biologie
6 Prénom et patronyme
7 Saints et bienheureux
8 Patronyme
9 Toponyme
10 Titres
11 Marques commerciales
12 Navires
13 Références
14 Voir aussi

Sur les autres projets Wikimedia :
Mercure, sur le Wiktionnaire
mercure, sur le Wiktionnaire

Mythologie [modifier | modifier le code]
- Mercure, dieu de la mythologie romaine, il est l'équivalent d'Hermès dans la mythologie grecque.

Alchimie [modifier | modifier le code]
- Mercure philosophique, un des trois principes de l'alchimie, avec le soufre et le sel.

Astronomie [modifier | modifier le code]
- Mercure, première planète du système solaire, la plus proche du Soleil.

Physique-chimie, toxicologie [modifier | modifier le code]
- mercure, élément
- millimètre de mercure (abréviation mmHg), unité de mesure de pression.
- Mercure, intoxication et maladie professionnelle.

Biologie [modifier | modifier le code]
Deux insectes lépidoptères (papillons) portent le nom de **mercure** :
- Le mercure, ou petit agreste.
- Le mercure tyrrhénien, ou agreste tyrrhénien.

Prénom et patronyme [modifier | modifier le code]
Mercure est un prénom masculin.
Mercure est aussi un patronyme.

Saints et bienheureux [modifier | modifier le code]
- Mercure de Smolensk (?-1238) soldat d'origine byzantine qui, durant l'invasion tatare, mourut martyr à Smolensk ; célébré le 24 novembre[1].
- Mercure de Césarée (vi^e siècle ?), jeune chrétien d'origine scythe, servait dans l'armée impériale romaine, décapité à Césarée de Cappadoce ; célébré le 25 novembre[2].
- Mercure des Grottes de Kiev (xv^e siècle), dit « le Jeûneur », moine et ascète à la Laure des Grottes de Kiev ; saint de l'Église orthodoxe célébré le 4 novembre[3] ou le 24 novembre en Russie[4].

Patronyme [modifier | modifier le code]
Mercure est un nom de famille notamment porté par :
- Jean Mercure (1909-1998), acteur et metteur en scène français, premier directeur du Théâtre de la Ville à Paris.
- Monique Mercure (1936-), actrice québécoise.
- Daniel Mercure (1955-), musicien canadien.
- Pierre Mercure (1927-1966), musicien canadien.

Toponyme [modifier | modifier le code]
Mercure est un nom de lieu notamment porté par :
- Mont Mercure, montagne d'Italie

FIGURE 2 – Une page de désambiguïsation

de relations d'hyperonymie qu'elles contiennent. Le corpus a subi une phase de nettoyage pour extraire le texte à partir de la version XML. Nous avons ensuite utilisé TreeTagger [2] pour l'étiquetage morpho-syntaxique. Enfin, nous avons utilisé l'extracteur de termes YaTeA (Aubin & Hamon, 2006) pour identifier les syntagmes nominaux qui occupent les positions d'arguments dans la relation d'hyperonymie. La dernière étape du processus consiste à transformer le format de sortie de TreeTagger au format requis par l'environnement d'ingéniere linguistique Gate, afin d'exécuter les patrons lexico-syntaxiques.

L'évaluation des systèmes d'extraction de la relation d'hyperonymie se fonde généralement sur des ressources lexicales telles que WordNet. En l'absence d'une ressource aussi complète pour le français, nous avons opté pour la création d'une annotation de référence, en marquant toutes les occurrences de la relation d'hyperonymie dans notre corpus. Ce sous-corpus utilisé contient 30 pages de désambiguïsation, 9718 tokens et 553 relations d'hyperonymie annotées manuellement.

2. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger

Nom de la page	Nbre de relations d'hyperonymie annotées
Timbre	20
Souris	26
Samurai	4
Renaissance	26
Produit	37
Prairial	6
Pluton	9
Pentagone	8
Opera	19
Morville	12
Montreuil	41
Matrice	33
Magma	9
Lumen	10
Louis Philippe	13
Lincoln	61
Kikai	5
Henri Giraud	8
Gaulois	8
Gandhi	25
Divergence	10
Coulombs	3
Cornet	30
Colombier	34
Calypso	23
Champollion	3
Apache	14
Analyse	8
Ampoule	10
Columbia	38

TABLE 1 – Le sous-corpus utilisé

4 L'extracteur d'hyperonymie

L'extracteur d'hyperonymie que nous proposons pour les pages d'homonymie est basé sur la définition de patrons lexico-syntaxiques, constitués d'un ensemble d'une dizaine de patrons proposés par (Jacques & Aussenac-Gilles, 2006) et augmentés de patrons spécifiques visant à capter les spécificités des rubriques patronymes et toponymes des pages de désambiguïsation. Voici à titre d'exemple un extrait de la rubrique *Patronymes* de la page de désambiguïsation *Babel* :

1. Louis Babel, prêtre-missionnaire oblat et explorateur du Nouveau-Québec (1826-1912).
2. Isaac Babel, écrivain et dramaturge russe (1894-1940).
3. Ryan Babel, joueur de football batave (1986-).
4. Roger Viry-Babel (1945-2006), universitaire et cinéaste français.

Le tableau 2 regroupe les relations qui peuvent être extraites de cette rubrique et les place en regard de celles qui existent actuellement dans DBPédia pour les entités correspondantes, à savoir *Louis Babel*, *Isaac Babel*, *Ryan Babel*, et *Roger Viry-Babel*. L'ontologie de DBPédia modélise les relations d'hyperonymie par la propriété RDF *rdf :type* pour les relations entre une instance et sa classe, et la propriété RDF *rdf :subClassOf* pour les relations entre classes. Cette comparaison permet de constater que 5 des 12 relations décrites dans le texte ne sont actuellement pas recensées dans DBPédia.

Phrase	Relations	Relations dans DBPédia
(1)	*Hyp*(Louis Babel, Prêtre-missionnaire) *Hyp*(Louis Babel, Explorateur)	*rdf :type*(Louis Babel, Agent) *rdf :type*(Louis Babel, Personne)
(2)	*Hyp*(Issac Babel, Ecrivain) *Hyp*(Issac Babel, Dramaturge)	*rdf :type*(Issac Babel, Agent) *rdf :type*(Issac Babel, Artiste) *rdf :type*(Issac Babel, Personne) *rdf :type*(Issac Babel, Ecrivain)
(3)	*Hyp*(Ryan Babel, Joueur de football)	*rdf :type*(Ryan Babel, Agent) *rdf :type*(Ryan Babel, Athlete) *rdf :type*(Ryan Babel, Personne) *rdf :type*(Ryan Babel, Joueur de football)
(4)	*Hyp*(Roger Viry-Babel, Universitaire) *Hyp*(Roger Viry-Babel,Cinéaste)	*rdf :type*(Roger Viry-Babel, Agent) *rdf :type*(Roger Viry-Babel, Personne)

TABLE 2 – Exemples de relations et comparaison avec DBPédia

Des caractéristiques semblables concernent également la rubrique des toponymes, ce qui nous a amené à proposer des patrons pour extraire plus spécifiquement les relations d'hyperonymie présentes dans ces deux rubriques. Ces patrons sont décrits à l'aide des Expressions Régulières (ER) suivantes :

1. NP '(' { ({NUM - (NUM | ' ?')} | NUM) } ')' , NP {, NP}* {(et|ou) NP} ?
2. NP (,| :) NP {, NP}* {(et|ou) NP} ?
3. NP '(' NP {, NP}* {(et|ou) NP} ? ')'

NP (pour *Noun Phrase*) correspond aux traces linguistiques des arguments de la relation d'hyperonymie (l'hyperonyme et l'hyponyme), Nous avons utilisé l'extracteur de terme YaTeA pour annoter les

syntagmes nominaux en prenant en compte la proposition la plus longue que propose cet extracteur. Par exemple pour la phrase *Une vache ayant quatre sabots est un animal*, YaTeA propose 2 termes : *une vache* et *une vache ayant quatre sabots*. On considère comme argument la seconde proposition correspondant au terme le plus long.

Le tableau 3 présente des exemples des relations extraites avec ces patrons.

Patron	Texte reconnu	Relation extraite
(1)	Sonia Gandhi (1946), présidente du Parti du Congrès	*Hyp*(Sonia Gandhi, présidente du Parti du Congrès)
(2)	Gopalkrishna Gandhi, homme politique	*Hyp*(Gopalkrishna Gandhi, homme politique)
(3)	atelier (local, espace)	*Hyp*(atelier, local)
Hearst	Gandhi est un film	*Hyp*(Gandhi, film)

TABLE 3 – Exemples de relations extraites

Les relations d'hyperonymie présentes dans les pages d'homonymie relèvent de deux types du point de vue de l'ontologie : relation de typage (*is-a* ou *type-of*) et relation d'instance (*instance-of*). Les rubriques Patronymes et Toponymes contiennent généralement des entités nommées pour les lieux et les personnes, donc des relations d'instance. Par ailleurs, les autres rubriques sont riches en relations de typage ; par exemple *Hyp*(Souris, Dispositif informatique) extraite de la page *Souris* est une relation de typage. Le type des arguments pourrait donc permettre de distinguer ces deux types de relation, ce que nous n'avons pas fait jusqu'à présent. Les patrons proposés ont été définis en analysant 20 pages de désambiguïsation de Wikipédia. Dans la section suivante nous présentons leur évaluation sur un corpus de 30 pages différentes que nous avons présenté dans la section 3.

5 Evaluation et discussion

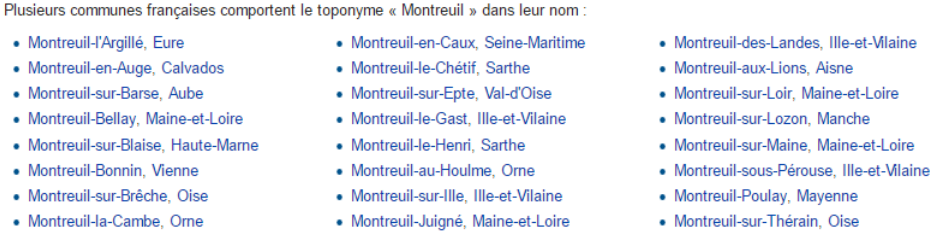
A titre indicatif, nous avons calculé le rappel et la précision par rapport à notre corpus annoté manuellement (tableau 4).

	Rappel	Précision
Corpus	0.75	0.68

TABLE 4 – Evaluation des patrons

Nous remarquons que la précision est inférieure au rappel, ce qui est atypique dans le cas des approches basées sur patrons. Ce résultat confirme que la relation d'hyperonymie s'exprime de façon très régulière dans le corpus particulier que nous avons traité. Néanmoins, certains patrons pourraient être affinés pour éliminer certains faux positifs. C'est en particulier le cas du patron (2), qui génère une quantité importante de bruit malgré la restriction à des termes situés en début de phrase. Par exemple la phrase *Dans la mythologie grecque, Calypso est une nymphe* contient la relation

Hyp(Calypso, nymphe). Cette relation est correctement retrouvée par un patron générique défini par Hearst. Par contre, le patron (2) extrait la relation *Hyp*(mythologie grecque, Calypso), qui est erronée. La présence de listes énumératives verticales dans la rubrique *Patronyme* n'est pas prise en compte par ces patrons. Par exemple, la page de désambiguïsation *Montreuil* contient une grande quantité de relations d'hyperonymie présentes dans une liste énumérative verticale, telle qu'illustrée par la figure 3. Les patrons que nous proposons ne sont pas adaptés au traitement de tels exemples, et devraient donc être complétés. Nous envisageons de définir un modèle d'apprentissage des patrons basé sur la structure de texte en nous appuyant sur l'approche de (Fauconnier & Kamel, 2015) pour le traitement des structures énumératives verticales.


Plusieurs communes françaises comportent le toponyme « Montreuil » dans leur nom :

- Montreuil-l'Argillé, Eure
- Montreuil-en-Auge, Calvados
- Montreuil-sur-Barse, Aube
- Montreuil-Bellay, Maine-et-Loire
- Montreuil-sur-Blaise, Haute-Marne
- Montreuil-Bonnin, Vienne
- Montreuil-sur-Brêche, Oise
- Montreuil-la-Cambe, Orne
- Montreuil-en-Caux, Seine-Maritime
- Montreuil-le-Chétif, Sarthe
- Montreuil-sur-Epte, Val-d'Oise
- Montreuil-le-Gast, Ille-et-Vilaine
- Montreuil-le-Henri, Sarthe
- Montreuil-au-Houlme, Orne
- Montreuil-sur-Ille, Ille-et-Vilaine
- Montreuil-Juigné, Maine-et-Loire
- Montreuil-des-Landes, Ille-et-Vilaine
- Montreuil-aux-Lions, Aisne
- Montreuil-sur-Loir, Maine-et-Loire
- Montreuil-sur-Lozon, Manche
- Montreuil-sur-Maine, Maine-et-Loire
- Montreuil-sous-Pérouse, Ille-et-Vilaine
- Montreuil-Poulay, Mayenne
- Montreuil-sur-Thérain, Oise

FIGURE 3 – Exemple de liste énumérative verticale

Rappelons que l'évalution des patrons eux-mêmes n'était pas l'objectif principal de cette première étape du travail. Notre intention était de commencer à mesurer la capacité de notre démarche à enrichir les ressources sémantiques DBPédia et BabelNet. Pour cela, nous avons comparé le nombre de relations valides repérées par notre approche aux relations présentes dans ces deux ressources et impliquant les mêmes entités. Cette évaluation s'est effectuée en trois étapes :

1. pour chaque relation trouvée, interroger DBPédia et BabelNet pour vérifier l'existence de l'entité hyponyme ;
2. si cet hyponyme existe, récupérer tous les hyperonymes identifiés dans DBPédia et BabelNet ;
3. comparer les hyperonymes trouvés dans DBPédia et BabelNet avec les hyperonymes trouvés par notre approche.

Le tableau 5 montre le pourcentage des relations trouvées avec notre approche et qui ne sont pas présentes dans DBPédia (v19.01.2015) et BabelNet (v3.6). Nous précisons que ces pourcentages concernent les relations valides, donc après élimination des relations fausses trouvées par notre extracteur :

DBPédia	BabelNet
33.49%	33.01%

TABLE 5 – Proportion des relations absentes des ressources

On constate donc que l'application des patrons permettrait de compléter de façon conséquente les ressources disponibles dans DBPédia et BabelNet. Les pourcentages de relations manquantes dans DBPédia et BabelNet sont equivalents. De plus, les relations absentes dans les deux cas sont les

mêmes. On peut faire l'hypothèse que cette proximité tient en partie au fait que les deux ressources sont partiellement construites sur les mêmes principes, à savoir l'exploitation des éléments structurés de Wikipédia. Ce qui expliquerait pourquoi les relations absentes dans les deux ressources sont les mêmes. Ci-dessous quelques exemples de relations d'hyperonymie extraites par notre approche et qui n'existent pas dans DBPédia et BabelNet :

— la souris est le muscle charnu qui tient à l'os du manche d'un gigot : *Hyp*(Souris, Muscle) ;
— Cornet à dés, gobelet servant à mélanger puis jeter les dés : *Hyp*(Cornet à dés, gobelet) ;
— Le Cornet, goguette fondée en 1896 par Georges Courteline : *Hyp*(Le Cornet, goguette) ;
— Charles Joseph Cornet (1879-1914), explorateur et écrivain français : *Hyp*(Charles Joseph Cornet, explorateur), *Hyp*(Charles Joseph Cornet, écrivain).

6 Conclusion

Dans cet article, nous avons présenté notre approche d'extraction de relations d'hyperonymie à partir de Wikipédia pour enrichir DBPédia, en particulier à partir des pages de désambiguïsation. L'évaluation de notre travail a montré que les textes de Wikipédia contiennent aussi des relations d'hyperonymie qui ne sont pas présentes dans les parties structurées. Cette première expérience est une toute première étape par rapport à l'objectif général que nous nous sommes fixé. Rappelons que celui-ci consiste à combiner un ensemble de méthodes d'extraction de relations sémantiques afin d'alimenter des ressources sémantiques du Web des données en français. Nous avons testé notre démarche sur un type de textes spécifique, les pages de désambiguïsation de Wikipédia, en nous limitant à une approche par patrons. Ce premier travail nous a cependant permis de constituer une première chaîne de traitement consistant à prétraiter le *dump* du corpus Wikipédia (étiquetage et extraction de termes), à produire un corpus annoté, à projeter un ensemble de patrons morpho-syntaxiques, et à connecter nos résultats aux ressources BabelNet et DBPédia afin de s'assurer de l'apport de ressources complémentaires issues directement des textes de Wikipédia, et non plus seulement des informations structurées contenues dans ces articles. L'étape suivante va consister à mettre en oeuvre à plus grande échelle l'approche par patrons, en faisant appel à des techniques d'apprentissage, et en considérant cette fois l'intégralité du corpus Wikipédia.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, p. 380–387. Springer.

BORDEA G., BUITELAAR P., FARALLI S. & NAVIGLI R. (2015). SemEval-2015 task 17 : Taxonomy Extraction Evaluation (TExEval). *SemEval-2015*, **452**(465), 902.

CARABALLO S. (2001). *Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University*. PhD thesis.

FAUCONNIER J.-P. & KAMEL M. (2015). Discovering Hypernymy Relations using Text Layout. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, p. 249–258, Denver, Colorado : Association for Computational Linguistics.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

JACQUES M.-P. & AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. volume 47, p. 11–32.

KAZAMA J. & TORISAWA K. (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 698–707.

LEFEVER E., VAN DE KAUTER M. & HOSTE V. (2014). Evaluation of automatic hypernym extraction from technical corpora in English and Dutch. In *9th International Conference on Language Resources and Evaluation (LREC)*, p. 490–497 : European Language Resources Association (ELRA).

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. & BIZER C. (2014). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

LENCI A. & BENOTTO G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, p. 75–79 : Association for Computational Linguistics.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Detecting semantic relations between terms in definitions. In *COLING*, p. 55–62.

MORIN E. & JACQUEMIN C. (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, **38**(4), 363–396.

MORSEY M., LEHMANN J., AUER S., STADLER C. & HELLMANN S. (2012). DBpedia and the live extraction of structured data from Wikipedia. *Program*, **46**(2), 157–181.

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.

PANCHENKO A., FARALLI S., RUPPERT E., REMUS S., NAETS H., FAIRON C., PONZETTO S. P. & BIEMANN C. (2016). Taxi : a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

PANCHENKO A., NAETS H., BROUWERS L., ROMANOV P. & FAIRON C. (2013). Recherche et visualisation de mots sémantiquement liés. *TALN-RÉCITAL 2013*, p. 747–754.

PANTEL P. & PENNACCHIOTTI M. (2006). Espresso : Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 113–120 : Association for Computational Linguistics.

SCHROPP G., LEFEVER E. & HOSTE V. (2013). A Combined Pattern-based and Distributional Approach for Automatic Hypernym Detection in Dutch. In G. ANGELOVA, K. BONTCHEVA & R. MITKOV, Eds., *RANLP*, p. 593–600 : RANLP 2013 Organising Committee / ACL.

SNOW R., JURAFSKY D. & NG A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

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

SUMIDA A. & TORISAWA K. (2008). Hacking wikipedia for Hyponymy Relation Acquisition. In *IJCNLP*, volume 8, p. 883–888.

VAN DER PLAS L., BOUMA G. & MUR J. (2005). Automatic acquisition of lexico-semantic knowledge for QA. In *Proceedings of the IJCNLP workshop on Ontologies and Lexical Resources*, p. 76–84.