Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte

Adel Ghamnia, Mouna Kamel, Cassia Trojahn dos Santos, Cécile Fabre, Nathalie Aussenac-Gilles

Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte

Adel Ghamnia[1,2], Mouna Kamel[1], Cassia Trojahn[1], Cécile Fabre[2], Nathalie Aussenac-Gilles[1]

[1] IRIT Institut de Recherche en Informatique de Toulouse
Toulouse, France
{adel.ghamnia,mouna.kamel,cassia.trojahn,nathalie.aussenac-gilles}@irit.fr

[2] Laboratoire CLLE, équipe ERSS, Toulouse, France
cecile.fabre@univ-tlse2.fr

Résumé : Extraire des relations d'hyperonymie à partir des textes est une des étapes clés de la construction automatique d'ontologies et du peuplement de bases de connaissances. Plusieurs types de méthodes (linguistiques, statistiques, combinées) ont été exploités par une variété de propositions dans la littérature. Les apports respectifs et la complémentarité de ces méthodes sont cependant encore mal identifiés pour optimiser leur combinaison. Dans cet article, nous nous intéressons à la complémentarité de deux méthodes de nature différente, l'une basée sur les patrons linguistiques, l'autre sur l'apprentissage supervisé, pour identifier la relation d'hyperonymie à travers différents modes d'expression. Nous avons appliqué ces méthodes à un sous-corpus de Wikipedia en français, composé des pages de désambiguïsation. Ce corpus se prête bien à la mise en œuvre des deux approches retenues car ces textes sont particulièrement riches en relations d'hyperonymie, et contiennent à la fois des formulations rédigées et d'autres syntaxiquement pauvres. Nous avons comparé les résultats des deux méthodes prises indépendamment afin d'établir leurs performances respectives, avec le résultat des deux méthodes appliquées ensemble. Les meilleurs résultats obtenus correspondent à ce dernier cas de figure avec une F-mesure de 0.68. De plus, l'extracteur Wikipedia issu de ce travail permet d'enrichir la ressource sémantique DBPedia en français : 55% des relations exprimées et identifiées par notre extracteur ne sont pas présentes dans DBPedia.
Mots-clés : Extraction de relations d'hyperonymie, Supervision distante, Patrons lexico-syntaxiques, Bases de connaissances

1 Introduction

Dans de nombreux domaines tels que l'intelligence artificielle, le web sémantique, le génie logiciel, la recherche d'information ou l'aide au diagnostic, les applications nécessitent un fort potentiel de raisonnement, basé sur une ressource sémantique qui décrit généralement des concepts liés par des relations. Ces ressources peuvent être construites manuellement. Elles sont alors de bonne qualité, mais du fait de leur élaboration coûteuse, elles n'offrent qu'une couverture restreinte du domaine. Compte tenu du volume sans cesse croissant de textes disponibles dans un format numérique, le traitement automatique de la langue permet d'envisager la construction (semi-)automatique de telles ressources, en exploitant les connaissances véhiculées dans ces textes. Un enjeu majeur est alors d'acquérir ces connaissances, pour ensuite pouvoir les formaliser et les organiser au sein de ressources sémantiques. Dans ce contexte, la tâche d'identification de relations est une étape cruciale, car située en amont d'autres tâches comme l'expansion sémantique en recherche d'information ou encore l'extraction de relations pour la construction de ressources sémantiques (thesaurus, taxonomies, ontologies, ressources termino-ontologiques, etc. (Buitelaar *et al.*, 2005). De nombreux travaux se sont employés à extraire les relations d'hypernonymie car elles constituent l'ossature principale de la plupart de ces types de ressources.

Deux paradigmes organisent ce champ d'étude (Tanev & Magnini, 2008; Granada, 2015) : les approches qualifiées de linguistiques font appel à des patrons pour identifier des indices de relation entre termes (Hearst, 1992) ; les approches statistiques, aujourd'hui dominantes, recourent à des procédures d'apprentissage supervisé (Pantel & Pennacchiotti, 2008) ou non-supervisé (Banko *et al.*, 2007), ou font appel à des indices distributionnels (Lenci & Benotto, 2012). Ces travaux ont appliqué ces différentes méthodes à une langue, un domaine donné (général ou de spécialité), un genre de corpus (encyclopédique, scientifique, journalistique, etc.), selon la nature des sources de connaissances utilisées (documents structurés, semi-structurés ou non structurés), ou selon l'utilisation visée de la taxonomie (intégration dans des ressources plus complexes comme des thésaurus, des termino-ontologies ou des ontologies "riches").

L'étude que nous menons, et dont nous présentons ici une première évaluation, vise à montrer l'intérêt d'appliquer différentes approches sur un même corpus pour identifier la relation d'hyperonymie, à travers ses différents modes d'expression, selon qu'elle est formulée dans une section structurée, semi-structurée ou non structurée du texte. En effet, la relation d'hyperonymie peut être exprimée par le lexique et la structure syntaxique comme dans *Le sable est une roche sédimentaire meuble*, par une inclusion lexicale comme dans *pigeon domestique* (sous-entendu *le pigeon domestique est un pigeon*), ou encore à l'aide d'éléments de ponctuation ou de mise en forme qui se substituent aux marqueurs lexicaux comme la virgule dans *Le cheval de Troie, un mythe grec* ou encore la disposition dans les structures énumératives.

Pour ce faire, nous analysons la complémentarité de deux approches, l'une linguistique basée sur des patrons lexico-syntaxiques, et l'autre statistique basée sur de l'apprentissage supervisé. Nous les avons appliquées sur un corpus constitué des pages de désambiguïsation de Wikipedia. En effet, ces pages offrent un premier cas de figure favorable : très riches en relations d'hyperonymie, elles comportent du texte rédigé (assez minoritairement), et, pour l'essentiel, du texte peu rédigé (structure syntaxique incomplète) usant de mise en forme matérielle variée comme la ponctuation, diverses polices de caractère ou la disposition.

Ce travail s'inscrit dans le cadre du projet SemPedia [1] dont l'objectif est d'enrichir la ressource sémantique DBPedia pour le français (les ressources pour la langue française faisant défaut aujourd'hui encore), en spécifiant et implémentant un ensemble de nouveaux extracteurs Wikipedia dédiés à l'extraction de la relation d'hyperonymie.

L'article est organisé de la façon suivante. La section 2 rappelle les principaux travaux connexes à notre proposition. La section 3 présente le matériel et les méthodes mis en œuvre, à savoir la description des corpus d'apprentissage et de référence, leurs pré-traitements, et les approches d'extraction retenues. Les résultats obtenus sont présentés et discutés dans la section 4. Enfin la section 5 permet de conclure et d'indiquer les perspectives envisagées.

2 Travaux connexes

En matière d'extraction de relations, le travail pionnier des méthodes linguistiques est celui de (Hearst, 1992) qui a défini un ensemble de patrons lexico-syntaxiques spécifiques à l'hyperonymie pour l'anglais. Ce travail a été repris en français pour identifier différents types de relations (Séguéla & Aussenac-Gilles, 1999), des relations d'hyperonymie entre termes (Morin & Jacquemin, 2004), des relations de méronymie (Berland & Charniak, 1999), en intégrant

1. http ://www.irit.fr/Sempedia

progressivement des techniques d'apprentissage. Dans la deuxième famille de méthodes, les travaux de Snow *et al.* (2004) et Bunescu & Mooney (2005) appliquent des techniques d'apprentissage supervisé à un ensemble d'exemples annotés à la main. On sait que le coût de l'annotation manuelle constitue la principale limite de l'apprentissage supervisé. Une manière de pallier ce problème est l'apprentissage par supervision distante qui consiste à construire l'ensemble d'exemples à l'aide d'une ressource externe (Mintz *et al.*, 2009). Cette approche requière cependant de disposer d'une ressource sémantique offrant un taux de couverture correct du corpus. D'autres manières concernent l'apprentissage semi-supervisé ou non supervisé. Brin (1998) a utilisé une sélection de patrons pour construire l'ensemble d'exemples, mettant en œuvre une méthode fondée sur l'apprentissage semi-supervisé appelée aussi bootstrapping. (Agichtein & Gravano, 2000) et (Etzioni *et al.*, 2004) ont repris cette méthode en ajoutant des traits sémantiques pour identifier des relations entre entités nommées. L'apprentissage non supervisé, basé sur des techniques de clustering, a été mis en œuvre par (Yates *et al.*, 2007) et (Fader *et al.*, 2011) qui ont utilisé des traits syntaxiques pour entraîner leurs classifieurs et identifier des relations entre entités nommées.

Au-delà de ces travaux, qui proposent et évaluent des approches une à une, peu de résultats existent sur les apports respectifs et la complémentarité de plusieurs méthodes en vue d'en optimiser la combinaison. Granada (2015) a comparé les performances de différents méthodes (par patrons, par inclusion lexicale, distributionnelles) pour la tâche d'extraction de la relation d'hyperonymie dans différentes langues, en définissant plusieurs métriques telles que, outre les mesures classiques d'évaluation (rappel et précision), la densité et la profondeur des hiérarchies. L'évaluation a été menée sur différents types de corpus mais ne prend pas en compte les approches par apprentissage.

Dans (Yap & Baldwin, 2009), les auteurs étudient l'impact du choix du corpus, la taille des exemples d'entraînement sur la performance de méthodes de même nature (méthodes supervisées), sur plusieurs types de relation (hyperonymie, synonymie et antonymie), tandis que dans (Ben Abacha & Zweigenbaum, 2011), une approche hybride permet d'extraire des relations entre entités spécifiques (maladie et traitement) dans des corpus biomédicaux. Cette approche repose sur des patrons, basés sur l'expertise humaine, et sur une méthode d'apprentissage statistique basée sur le classifieur SVM. L'approche calcule automatiquement les poids pour les deux différentes méthodes et ces poids sont ensuite appliqués pour intégrer la sortie de chaque méthode. Dans cette même ligne, nous exploitons des méthodes de nature différente, mais nous nous concentrons sur un type de relation spécifique.

En ce qui concerne l'enrichissement de la base de connaissances DBPedia, plusieurs applications, appelées "extracteurs" ont été développées pour analyser les différents éléments présents dans les pages Wikipédia. Ainsi Morsey *et al.* (2012) ont développé 19 extracteurs qui extraient des entités et des relations entre entités identifiés au sein de chaque élément de structure de ces pages : résumé, images, infobox, etc. D'autres travaux ont été proposés pour l'extraction de relations à partir de ces différents éléments, notamment pour les relations d'hyperonymie. Citons Suchanek *et al.* (2007) qui ont exploité la partie Catégorie dans les pages Wikipédia pour construire la base de connaissances Yago, Kazama & Torisawa (2007) qui ont exploité la partie Définition, et enfin Sumida & Torisawa (2008) qui se sont intéressés aux menus. On constate donc que la base de connaissances DBPedia est construite uniquement à partir des éléments de structure des pages Wikipédia : les travaux visant l'extraction des relations à partir de textes

n'ont pas été mis à profit pour l'alimentation de ces ressources, ce qui veut dire que la majorité des connaissances présentes dans ces pages restent inexploitées.

Nous proposons ici d'analyser la complémentarité de deux méthodes d'extraction de relations d'hypéronymie, de nature différente, afin de mieux exploiter les différentes sections du texte présentant des types de rédaction et des niveaux de structuration variables. Ces méthodes ont été appliquées sur un corpus de pages de désambiguïsation Wikipédia. Ces travaux sont décrits dans la section suivante.

3 Données et Méthodes

Cette section est consacrée à la description des corpus utilisés, des pré-traitements effectués sur ces corpus, et les méthodes d'extraction de relations que nous avons retenues.

3.1 Corpus

Des pages de nature différente peuvent être identifiés au sein de l'encyclopédie Wikipedia. Parmi elles, les pages d'homonymie (appelées aussi *pages de désambiguïsation*) listent les articles dont le titre est polysémique, et donnent une définition de toutes les acceptions recensées pour ce titre, qui renvoient à autant d'entités. Grâce aux consignes de rédaction de Wikipedia, qui recommandent l'usage de templates (*Toponymes*, *Patronymes*, etc.), ces pages présentent des régularités aussi bien rédactionnelles que de mise en forme pour présenter les différentes acceptions du terme (comportant un ou plusieurs mots) faisant l'objet de la page. De fait, pour chaque acception, une définition de ce sens et un lien vers la page correspondante sont fournis. Or les définitions sont des objets textuels au sein desquels la relation d'hyperonymie est souvent présente (Malaisé *et al.*, 2004) (Rebeyrolle & Tanguy, 2000).

Enfin, sur ces pages, les définitions prennent des formes variées mais prévisibles. Elles comportent pour l'essentiel du texte peu rédigé (structure syntaxique incomplète), mais aussi une mise en forme matérielle riche, utilisant la ponctuation, diverses polices de caractère ou la disposition.

Par exemple, la page d'homonymie *Mercure* cite plusieurs articles, et donne pour chacun une définition (Figure 1). Chaque définition présente plusieurs relations d'hyperonymie, exprimées par le lexique (*le mercure est un élément chimique*), à l'aide de caractères de ponctuation (la virgule dans *le Mercure, un fleuve du sud de l'Italie*), ou de la disposition comme dans les structures énumératives verticales (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.)

Ces pages d'homonymie étant pertinentes pour notre étude, nous avons constitué un corpus regroupant toutes celles de Wikipedia en français (dump 2016 de la version XML), soit 5924 pages de désambiguïsation. De ce corpus ont été extraits deux sous-corpus.

— 20 pages de désambiguïsation choisies aléatoirement forment le *corpus de référence*. Dans ces pages, les relations d'hyperonymie ont été annotées manuellement, en marquant les mots renvoyant aux entités en relation et la zone de texte signifiant la relation. Ce sous-corpus sera utilisé pour évaluer l'enrichissement potentiel de la ressource DBPedia ;

— les pages restantes (5904) forment un sous-corpus que nous appelons *corpus d'apprentissage*, destiné à entraîner et à évaluer notre modèle d'apprentissage (section 3.3.2).

Physique et chimie [modifier | modifier le code]

- Le mercure (symbole Hg) est un élément chimique.
- Le terme mercure rouge désignait au XIXe siècle l'iodure de mercure. Dans la dernière partie du XXe siècle il a été appliqué à une substance imaginaire, présentée comme un matériau stratégique rentrant dans la construction des armes nucléaires.
- Le millimètre de mercure (symbole mmHg), ou torr, est unité de mesure de pression.
- Plusieurs appareils de mesure ou méthodes physiques font référence au mercure, dont notamment :
  - la diode à vapeur de mercure,
  - la pile au mercure,
  - la pompe à mercure,
  - le porosimètre à mercure (en),
  - le thermomètre à mercure.

Toponyme et hydronyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :

- Mercure, une station du métro de Lille Métropole ;
- le Mercure, un fleuve du sud de l'Italie ;
- les îles Mercure, un archipel néo-zélandais, au large de la péninsule de Coromandel.
- le lac Mercure, un lac de l'île principale de l'archipel des Kerguelen, dans les Terres australes et antarctiques françaises ;
- le monastère Saint-Mercure, un important monastère féminin copte orthodoxe, situé dans le vieux Caire (Égypte) ;
- le mont Mercure, une montagne d'Italie ;
- Saint-Michel-Mont-Mercure, une ancienne commune française située dans le département de la Vendée, en région Pays-de-la-Loire
- la Vallée du Mercure, un grand bassin fluvial italien situé dans le sud de la Basilicate et le nord de la Calabre, et qui fut occupé par u lac au Pliocène.

FIGURE 1 – Extrait de la page de désambiguïsation *Mercure*

3.2 Pré-traitements

Chaque corpus est étiqueté morpho-syntaxiquement par TreeTagger [2]. Pour identifier l'expression de relations sémantiques, le texte est aussi annoté à l'aide de termes, à savoir des syntagmes, le plus souvent nominaux, pouvant désigner des entités ou des classes conceptuelles. Par exemple, *Mercure, système solaire, planète* sont quelques-uns des termes présents sur la figure 1. Les termes peuvent donc être inclus les uns dans les autres (*système* dans *système solaire*). Plutôt que d'utiliser un extracteur de termes, nous avons choisi de construire a priori deux listes de termes :

— LBabel comporte les labels en français des concepts présents dans la ressource sémantique BabelNet [3] ; elle servira à annoter le corpus d'apprentissage ;
— LCorpus, composée des termes annotés manuellement dans le corpus de référence ; elle sera utilisée pour l'évaluation de l'approche sur le corpus de référence.

L'annotation du corpus d'apprentissage par des termes provenant d'une source sémantique partagée assure une plus grande validité du modèle d'apprentissage. Cela évite aussi que l'identification des termes vienne biaiser l'extraction des relations.

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

3.3 Approches retenues

Nous avons retenu deux approches, souvent opposées par le coût de leur mise en œuvre et par les résultats de précision et de rappel qu'elles fournissent : l'utilisation de patrons lexico-syntaxiques, et une approche statistique basée sur de l'apprentissage supervisé mettant en œuvre le principe de supervision distante. Alors que les patrons représentent des schémas langagiers récurrents faisant appel au lexique, à la syntaxe et à la ponctuation, l'apprentissage automatique permet de combiner de nombreux indices du corpus, de différentes natures (morphologiques, syntaxiques, sémantiques ou encore de mise en forme, ...) et de capter ainsi de façon plus globale les propriétés des contextes.

3.3.1 Patrons lexico-syntaxiques

Un patron lexico-syntaxique décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte conformes à cette expression. Dans le cas de la recherche de relations, le patron caractérise un ensemble de formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes (Rebeyrolle & Tanguy, 2000). Un patron est d'autant plus efficace qu'il est adapté au corpus. Toutefois, la mise au point étant coûteuse, il est classique d'implémenter des patrons génériques comme ceux de (Hearst, 1992).

Nous avons mené deux expériences séparées, en utilisant tout d'abord **PatronsG**, une liste de 30 patrons génériques issus des travaux de (Jacques & Aussenac-Gilles, 2006), puis **PatronsGS**, cette même liste augmentée de patrons spécifiques à notre corpus et qui ont été définis manuellement [4] (Ghamnia, 2016).

3.3.2 Approche par apprentissage supervisé

Nous avons choisi de reprendre le principe de la supervision distante proposée par (Mintz *et al.*, 2009). Comme pour tout apprentissage supervisé, il convient de créer un ensemble d'exemples, d'entraîner un modèle statistique sur ces exemples, et d'évaluer le modèle sur un ensemble de test ou par validation croisée. L'originalité de la supervision distante réside dans le fait de construire les exemples automatiquement à partir d'une ressource externe. Dans le cadre de l'extraction de relations, la construction d'un exemple consiste à extraire une paire de termes d'une unité textuelle, à associer un ensemble de valeurs de traits préalablement définis, et à associer une classe. La classe correspond à la relation (si elle existe) qui lie, dans la ressource externe, deux concepts ayant pour labels les deux termes extraits du texte. Une fois entraîné à partir des exemples classés, un algorithme de classification multi-classes permet d'associer une classe à chaque exemple d'un nouveau corpus.

Nous avons adapté cette méthode en nous focalisant sur la relation d'hypéronymie, et en procédant à une classification binaire. Pour chaque exemple, nous avons défini un ensemble de propriétés exploitées par le système d'apprentissage : i) une paire de termes (ci-après Terme1 et Terme2) extraits d'une même phrase, ii) un contexte (ou fenêtre) de taille n formé par la séquence {n tokens précédant Terme1, Terme1, tokens séparant Terme1 et Terme2, Terme2, n

4. Une implémentation en JAPE de ces patrons est visible sur le site : https ://github.com/aghamnia/SemPediaPatterns

tokens suivant Terme2}, iii) un ensemble de valeurs de traits ayant des niveaux de granularité différents (voir Table 1), et iv) la classe (positif ou négatif) à laquelle appartient l'exemple. Un exemple est positif (resp. négatif) si les deux termes dénotent deux concepts qui existent dans la ressource sémantique, et si la relation d'hyperonymie entre ces deux concepts est représentée (resp. n'est pas présente) dans cette ressource. Dans tous les autres cas, la paire de termes ne constitue pas un exemple d'apprentissage.

Niveau de granularité	Trait	Signification	Type
Token	POS	Part Of Speech	chaîne de caractères
	LEMME	Forme lemmatisée du token	chaîne de caractères
Fenêtre	distT1	Nombre de tokens entre le mot et Terme1	entier
	distT2	Nombre de mots entre le token et Terme2	entier
	distT1T2	Nombre de tokens entre Terme1 et Terme2	entier
	nbMotsFenêtre	Nombre de tokens dans la fenêtre	entier
Phrase	nbMotsPhrase	Nombre de tokens dans la phrase	entier
	presVerbe	Présence d'une forme verbale	booléen

TABLE 1 – Ensemble des traits associés à un exemple.

Les mauvaises performances fournies par les analyseurs syntaxiques lorsqu'ils sont appliqués sur du texte peu rédigé nous a conduit à ne pas prendre en compte les dépendances syntaxiques.

Nous décrivons ci-dessous la construction d'un exemple à partir de la phrase *"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"*

La projection de la liste des termes LBabel conduit à annoter la phrase par les termes `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Considérons le couple <`Lime`, `fruit`> choisi aléatoirement par le système : Terme1=`Lime` et Terme2=`fruit`. Pour une fenêtre égale à 3 tokens [5], le système extrait alors de la phrase :

```
Terme1 ou citron vert, le Terme2 des limettiers :
```

où les mots correspondant aux termes ont été remplacés par Terme1 et Terme2. L'annotation par Tree-Tagger permet de remplacer les formes exactes des tokens par leur lemme précédé de leur catégorie syntaxique :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
```

Enfin, des fonctions de traits associent à chaque token les distances relatives (en nombre de tokens) de ce token à Terme1 et à Terme2 sous forme de couple de valeurs, le nombre de tokens entre Terme1 et Terme2 (en l'occurrence 5) et le nombre de tokens dans la phrase (ici 16). Le dernier trait indique l'absence de verbe dans la phrase.

```
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2)
(-9,-3) 5 16 false
```

5. Nous avons évalué des fenêtres de dimension 1, 3 et 5, l'optimum étant obtenu pour la dimension 3

Voici l'exemple dans son intégralité :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2)
(-9,-3) 5 16 false
```

Cet exemple est positif car les termes "lime" et "fruit" renvoient à des ressources en relation d'hyperonymie dans BabelNet.

Nous avons ainsi produit automatiquement ~8000 exemples, et avons conservé 6000 exemples (3000 positifs et 3000 négatifs). L'ensemble d'entraînement est composé de 4000 exemples pris aléatoirement parmi les 6000, en maintenant une quasi-parité positifs / négatifs (~2000/~2000), et l'ensemble de test comporte les 2000 exemples restants. Nous avons entraîné un algorithme de régression logistique binaire, MaxEnt (Berger *et al.*, 1996) sur l'ensemble d'entraînement. Appliqué à l'ensemble de test, MaxEnt a fourni un rappel de 0.63 et une précision de 0.71.

4 Expériences, résultats et discussion

A partir du corpus de référence, 688 exemples vrais positifs (VP) et 267 exemples vrais négatifs (VN) ont été identifiés manuellement. Nous interprétons les résultats obtenus afin de juger de la complémentarité des méthodes expérimentées, et de l'intérêt de leur utilisation conjointe.

4.1 Évaluation quantitative

Nous avons mis en œuvre les deux approches indépendamment, l'approche par patrons avec **PatronsG** puis avec **PatronsGS** (voir section 3.3.1), et l'approche par apprentissage supervisé (MaxEnt). Nous avons évalué leur complémentarité à l'aide de l'union et de l'intersection de leurs résultats. Le tableau 2 fournit les différentes valeurs de la précision, du rappel, de la F-mesure et de l'exactitude.

	PatronsG	PatronGS	MaxEnt	PatronsG union MaxEnt	PatronsGS union MaxEnt
Précision	0.96	0.81	0.71	0.72	0.73
Rappel	0.04	0.46	0.63	0.65	0.77
F-mesure	0.07	0.59	0.67	0.68	0.75
Exactitude	0.31	0.54	0.55	0.56	0.63

TABLE 2 – Évaluation des approches.

Confirmant l'état de l'art (Hearst, 1992) (Malaisé *et al.*, 2004), les patrons obtiennent un fort taux de précision, au détriment d'un faible rappel. Et comme attendu, les patrons génériques augmentés des patrons spécifiques donnent de meilleurs résultats que les patrons génériques seuls : bien que la précision baisse à 0.81, le rappel passe à 0.46, pour une F-mesure de 0.59. En revanche, l'approche par apprentissage supervisé, moins bonne en précision, produit un meilleur taux de rappel. En effet, le corpus présente de fortes régularités (aussi bien syntaxiques

que de mise en forme), ce qui permet de renforcer l'apprentissage. Ces résultats corroborent les résultats reportés dans la littérature pour d'autres types de corpus.

Nous constatons que l'application des deux approches fournit une meilleure F-mesure que les approches prises indépendamment, et que l'union des résultats de PatronsGS et de MaxEnt permet d'obtenir la meilleure F-mesure. Nous avons également pu analyser la complémentarité des PatronsGS et de MaxEnt à travers les résultats donnés dans la Table 3.

	PatronsGS ou MaxEnt	PatronsGS et MaxEnt	PatronsGS seul	MaxEnt seul	Aucune des 2 méthodes
Nombre VP	527	221	96	210	161

TABLE 3 – Parmi les 688 VP du corpus de référence, nombre de VP trouvés par les approches.

Nous avons pu ainsi constater que parmi les 306 VP qui ne sont identifiés que par une seule des deux méthodes, PatronsGS en identifie 32%, et MaxEnt 68%. Ce résultat confirme la complémentarité de ces deux approches.

4.2 Evaluation qualitative

Nous avons constaté que parmi les 221 relations trouvées par PatronsGS et MaxEnt, 9 relations concernent des relations exprimées par le verbe *être*, comme entre les termes *macédoine* et *salade de fruits* dans la phrase "La macédoine est une salade de fruits ou de légumes". Quasiment toutes les autres relations correspondent au schéma "X, Y" comme dans "Le cheval de Troie, un mythe grec". On remarque toutefois que les patrons retrouvent des relations entre des noms communs, alors que MaxEnt trouve essentiellement des relations entre entités.

Pour les 96 relations trouvées par PatronsGS et n'ayant pas été identifiées par MaxEnt, on trouve des relations (19) exprimées par le verbe *être*, mais lorsque la relation n'est pas exprimée en début de phrase, comme la relation entre *Babel fish* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire". Quasiment toutes les autres relations trouvées seulement par les patrons correspondent au schéma "X,Y" comme décrit ci-dessus. Nous n'avons pas encore identifié la cause du silence de MaxEnt à ce niveau.

Parmi les 210 relations trouvées par MaxEnt et non identifiées par PatronsGS, on trouve les cas d'inclusion lexicale (85), comme la relation entre *gare de Paris Bastille* et *gare*, issue du groupe nominal "gare de Paris Bastille". MaxEnt permet également de trouver des relations exprimées par d'autres verbes d'état (8) comme la relation entre *aigle* et *oiseaux rapaces* dans "Aigle désigne en français certains grands oiseaux rapaces". MaxEnt identifie aussi des relations exprimées dans des unités textuelles comportant des coordinations, comme la relation entre *poisson Babel* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire", ou encore entre *Louis Babel* et *explorateur* dans "Louis Babel, prêtre-missionnaire oblat et explorateur". Enfin MaxEnt identifie très bien les relations au sein de texte usant de mise en forme comme la relation entre *arête* et *barbe de l'épi* dans "Arête, "barbe de l'épi"", ou entre *Aigle* et *chasseur de mines* dans "Aigle (M647), chasseur de mines".

Finalement, parmi les 161 relations qui n'ont été trouvées ni par PatronsGS ni par MaxEnt, 55 correspondent à des inclusions lexicales (nous n'avons pas non plus identifié la cause du silence de MaxEnt), 64 possèdent une incise entre les deux termes comme dans "Un Appelant

(jansénisme) est, au XVIIIe siècle, un ecclésiastique qui ... " ou "Giuseppe Cesare Abba (1838-1910), écrivain". Les 42 cas restants concernent des formes d'expression non prises en charge par les patrons et trop peu fréquentes pour être apprises par MaxEnt, comme "X tel que Y".

Cette analyse confirme l'intérêt de mettre en œuvre sur un même corpus des approches complémentaires. Tout d'abord, nous avons pu constater que les inclusions lexicales sont identifiées par MaxEnt seul. Ensuite, les différentes occurrences de relations au sein d'une même phrase sont identifiées par les deux méthodes, comme on l'a vu ci-dessus à travers l'exemple "Le poisson Babel ou Babel fish est une espèce imaginaire". Enfin, MaxEnt permet d'identifier les relations exprimées selon différentes variantes d'un même schéma (incises, usage de mise en forme, etc.), dès lors que ces structures sont récurrentes. Par ailleurs, nous avons également pu observer que PatronsGS et MaxEnt sont complémentaires dans une proportion de ~1/3 vs. 2/3.

4.3 Enrichissement de la ressource DBPedia

Nous avons évalué l'enrichissement de la ressource DBPedia par les relations extraites par PatronsGS et/ou par MaxEnt. Pour cela, nous avons manuellement vérifié la présence dans DBPedia des 688 relations VP, en interrogeant la ressource DBPedia pour vérifier si des entités aux labels proches de *Terme1* et *Terme2* y sont liées par un chemin formé de relations *rdf :type* et *rdf :subclassOf*. Nous avons empiriquement fixé à 3 la longueur maximale de ce chemin. Cette vérification manuelle se justifie par le fait que les termes de LCorpus peuvent différer des labels de DBPedia.

Parmi ces 688 relations, 199 relations ne sont pas exprimées dans DBPedia. 103 de ces 199 relations ont été identifiées par MaxEnt et 42 d'entre elles ont été trouvées par PatronsGS. En considérant l'union des résultats des deux approches, 125 relations identifiées (20 relations étant présentes dans l'intersection des résultats individuels) ne sont pas dans DBPedia . La Table 4 présente le taux d'enrichissement de la ressource par rapport aux relations identifiées par chaque méthode. Ces résultats confirment que les textes de Wikipedia contiennent des relations d'hyperonymie non encore exploitées par les extracteurs Wikipédia.

	PatronsGS	MaxEnt	Union PatronsGS et MaxEnt
Taux d'enrichissement	0.21%	0.51%	0.63%

TABLE 4 – Taux d'enrichissement de DBPedia.

5 Conclusion et perspectives

Ces premières expériences nous ont permis de mettre en place des données et un protocole pour la comparaison de deux méthodes d'extraction des relations, de manière à en analyser finement la complémentarité. Les premiers résultats sont encourageants et convergent avec les travaux de (Malaisé *et al.*, 2004) (Granada, 2015) (Buitelaar *et al.*, 2005). Nous envisageons de les pousser plus loin dans plusieurs directions. Nous souhaitons en priorité intégrer d'autres techniques pour prendre en compte d'autres éléments textuels, par exemple le système qui traite les structures énumératives verticales et régulières (Kamel & Trojahn, 2016) ou encore les outils développés dans (Granada, 2015). Pour améliorer les performances de chaque méthode, outre un meilleur encodage des patrons, nous prévoyons de rajouter des nouveaux traits pour

l'apprentissage automatique. Bien sûr, la méthode devra être testée sur un autre corpus incluant d'autres types de pages Wikipedia.

A terme, notre ambition est de croiser les méthodes de manière à ce que les résultats des unes servent d'entrées plus riches aux autres, et améliorent ainsi leurs performances. La première piste envisagée dans ce sens serait d'utiliser les patrons pour annoter le corpus permettant d'indiquer qu'un patron a été (ou non) reconnu dans le contexte de deux termes, ce qui serait un signe fort de présence de la relation marquée par ce patron. Ce type de trait permettrait d'entraîner le classifieur à reconnaître plusieurs types de relations en plus de l'hyperonymie.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, p. 85–94 : ACM.

BANKO M., CAFARELLA M. J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, p. 2670–2676.

BEN ABACHA A. & ZWEIGENBAUM P. (2011). *A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts*, In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing : 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II*, p. 139–150. Springer Berlin Heidelberg.

BERGER A. L., PIETRA V. J. D. & PIETRA S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, **22**(1), 39–71.

BERLAND M. & CHARNIAK E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 57–64 : Association for Computational Linguistics.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, p. 172–183 : Springer.

BUITELAAR P., CIMIANO P. & MAGNINI B. (2005). Ontology learning from text : An overview. In *Ontology Learning from Text : Methods, Evaluation and Applications*, p. 3–12 : IOS Press.

BUNESCU R. C. & MOONEY R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 724–731 : Association for Computational Linguistics.

ETZIONI O., CAFARELLA M., DOWNEY D., KOK S., POPESCU A.-M., SHAKED T., SODERLAND S., WELD D. S. & YATES A. (2004). Web-scale information extraction in knowitall :(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, p. 100–110 : ACM.

FADER A., SODERLAND S. & ETZIONI O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1535–1545 : Association for Computational Linguistics.

GHAMNIA A. (2016). Extraction de relations d'hyperonymie à partir de wikipédia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*.

GRANADA R. L. (2015). *Evaluation of methods for taxonomic relation extraction from text*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, p. 539–545, Stroudsburg, PA, USA : Association for Computational Linguistics.

JACQUES M.-P. & AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thématique*, **47**(1), (en ligne).

KAMEL M. & TROJAHN C. (2016). Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles. In *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 27th French Knowledge Engineering Conference), Montpellier, France, June 6-10, 2016.*, p. 111–122.

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.

LENCI A. & BENOTTO G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 75–79 : Association for Computational Linguistics.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Detecting semantic relations between terms in definitions. In S. ANANADIOU & P. ZWEIGENBAUM, Eds., *COLING 2004 CompuTerm 2004 : 3rd International Workshop on Computational Terminology*, p. 55–62, Geneva, Switzerland : COLING.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, p. 1003–1011 : Association for Computational Linguistics.

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.

PANTEL P. & PENNACCHIOTTI M. (2008). Automatically harvesting and ontologizing semantic relations. *Ontology learning and population : Bridging the gap between text and knowledge*, p. 171–198.

REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, **25**, 153–174.

SÉGUÉLA P. & AUSSENAC-GILLES N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, p. 79–88.

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.

SUMIDA A. & TORISAWA K. (2008). Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8, p. 883–888 : Citeseer.

TANEV H. & MAGNINI B. (2008). Weakly supervised approaches for ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, p. 129–143 : Citeseer.

YAP W. & BALDWIN T. (2009). Experiments on pattern-based relation learning. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1657–1660 : ACM.

YATES A., CAFARELLA M., BANKO M., ETZIONI O., BROADHEAD M. & SODERLAND S. (2007). Textrunner : open information extraction on the web. In *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, p. 25–26 : Association for Computational Linguistics.