# Extraction de relations d'hyperonymie à partir de documents semi-structurés

**Adel Ghamnia**

*Institut de Recherche en Informatique de Toulouse (IRIT)*
*118 Route de Narbonne*
*31062 TOULOUSE*

*Cognition, Langues, Langage, Ergonomie (CLLE)*
*5, allées Antonio-Machado*
*31058 TOULOUSE*

*Adel.Ghamnia@irit.fr*

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

KEYWORDS: *Hypernym extraction, Knowledge Bases, morpho-syntactic patterns*

ENCADREMENT. *Nathalie Aussenac-Gilles (DR-CNRS), Cécile Fabre (PR), Mouna Kamel (MCF) et Cassia Trojahn (MCF)*

## 1. Contexte

Les bases de connaissances jouent aujourd'hui un rôle important dans de nombreuses applications : la recherche d'informations, les systèmes Question/Réponse, la classification de documents, etc. La construction manuelle de ces bases de connaissances est très coûteuse et reste assez restreinte, ce qui a donné naissance à des travaux pour automatiser cette tâche. Plusieurs bases de connaissances ont vu le jour, telles que BabelNet, DBPedia et Yago. Elles sont construites par l'extraction de connaissances à partir de ressources structurées ou semi-structurées disponibles sur le Web : Wikipédia, WordNet, GeoNames, etc. Par exemple, la base de connaissances DBPedia est construite à partir des différents éléments contenus dans les pages Wikipédia (infobox, menus, catégories, etc.). Plusieurs travaux ont montré l'intérêt d'exploiter

aussi le texte rédigé de ces ressources : les travaux de (Mintz *et al.*, 2009) et (Akbik *et al.*, 2012) montrent que le texte rédigé présent sur ces pages apporte des connaissances supplémentaires. Cependant, ces travaux exploitent le texte seul, sans tenir compte de la structure logique des pages, qui est pourtant très présente dans ce type de données textuelles et porteuse de sens ((Navigli *et al.*, 2008), (Fauconnier, 2016)). L'objectif de notre travail est d'exploiter au mieux ces pages semi-structurées, en prenant en compte à la fois le texte rédigé qu'elles contiennent et leur structure logique, pour identifier un type particulier de connaissances, les relations d'hyperonymie (*is-a*, *instance-of*), utilisées dans une ontologie pour typer les entités et les instances. Nous montrons qu'il est ainsi possible d'identifier de nouvelles relations non présentes dans une base de connaissances en exploitant le texte rédigé et en prenant en compte sa structure. Notre expérimentation se base sur l'exploitation des pages Wikipédia pour enrichir la version française de DBPedia. Dans cet article, après un court état de l'art sur l'extraction de relations d'hyperonymie à partir de Wikipédia, nous présenterons notre problématique, les actions réalisées et les actions futures.

## 2. État de l'art

DBPedia est une base de connaissances construite à partir des différents éléments présents dans les pages Wikipédia. Pour cela, (Morsey *et al.*, 2012) ont développé 19 extracteurs dédiés à chaque élément de structure de ces pages : résumé, images, infobox, etc. Ces extracteurs ciblent les différents éléments et extraient les relations qu'ils contiennent. 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 *et al.*, 2007) qui ont exploité la partie Définition, et enfin (Sumida *et al.*, 2008) qui se sont intéressés aux menus. Ces travaux ont été largement exploités pour évaluer des méthodes, mais ils n'ont pas servi pour l'enrichissement d'une base de connaissances. Ainsi, DBPedia est construite uniquement à partir des éléments de structure des pages Wikipédia, ce qui veut dire que la majorité des connaissances présentes dans le texte rédigé de ces pages restent inexploitées. Différentes méthodes ont cependant été définies pour extraire à partir des textes des relations sémantiques. Ces travaux utilisent généralement des extracteurs de termes et des techniques fondées sur l'application de patrons morpho-syntaxiques dans la lignée de (Hearst, 1992), sur des indices de proximité distributionnelle (Lenci *et al.*, 2012), sur des approches statistiques et par apprentissage (Snow *et al.* (2004), Ritter *et al.* (2009)), ou sur l'exploitation de structures textuelles spécifiques, par exemple les définitions (Malaisé *et al.*, 2004) ou les structures énumératives (Fauconnier *et al.*, 2015).

## 3. Problématique

Notre objectif est précisément d'exploiter et de coordonner la variété des techniques d'extraction de relations que nous venons de citer. Nous nous focalisons dans un premier temps sur la relation sémantique de type hyperonymie entre classes et instances. Notre approche consiste à combiner les différentes techniques d'extraction de relations sémantiques en exploitant à la fois le langage naturel présent dans le texte rédigé ainsi que sa structure. 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. L'intérêt de combiner ces méthodes est de proposer une approche cyclique qui permetterait à une méthode donnée d'améliorer ses performances en utilisant les résultats d'une autre méthode.

## 4. Actions réalisées

Notre première étude porte sur l'exploitation du texte et de la structure des pages de désambigüisation de Wikipédia. Ces pages sont très structurées, très riches en relations d'hyperonymie entre concepts et entre entités nommées, et à notre connaissance, le texte qu'elles contiennent n'est pas encore exploité pour enrichir DBPedia. Dans Wikipédia, une page de désambigüisation (appelée aussi page d'homonymie ou "*disambiguation page*" en anglais) décrit différents sujets et articles partageant un même terme. Exemple : La page de désambigüisation Babel[1] contient dans la rubrique "Patronymes" :

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.

Une page de désambigüisation est structurée en plusieurs rubriques. Parmi ces rubriques, on en retrouve deux qui ont une structure régulière : Patronyme et Toponyme. Ces deux rubriques ont un template proposé par Wikipédia, qui peut être utilisé dans d'autres rubriques comme : Astronomie, Biologie, etc. Ceci nous a conduits à proposer une approche ciblée, et donc un extracteur particulier, pour les rubriques respectant ce template, d'autres approches plus générales devant être utilisées pour les rubriques qui ne le respectent pas.

L'extracteur d'hyperonymie est basé sur la définition de patrons lexico-syntaxiques, constitués d'un ensemble de patrons proposés par (Jacques *et al.*, 2006) et augmentés de patrons spécifiques visant à capter les spécificités des rubriques Patronymes et Toponymes des pages de désambigüisation. Ces rubriques possèdent des caractéristiques semblables : énumération verticale, dont chaque item est numéroté, commençant par

1. https ://fr.wikipedia.org/wiki/Babel

une entité nommée ou un nom de classe suivi par une virgule et un autre nom de classe. 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} ? ')'

Les groupes nominaux (Noun Phrase) correspondent aux arguments de la relation d'hyperonymie. Ils sont identifiés par le recours à une chaîne de prétraitement des textes (étiquetage morphosyntaxique avec *TreeTagger* et extraction de termes avec YaTeA (Aubin *et al.*, 2006)).

Ces patrons nous permettent d'extraire à partir de la rubrique Patronymes de la page Babel, par exemple, les relations d'hypernoymie suivantes qui ne sont pas recensées actuellement dans DBPedia : *Hyp*(Louis Babel, Prêtre-missionnaire), *Hyp*(Isaac Babel, Dramaturge), *Hyp*(Roger Viry-Babel, Universtaire) et *Hyp*(Roger Viry-Babel, Cinéaste).

L'évaluation a été réalisée sur un corpus de 30 pages de désambigüisation contenant 553 relations d'hyperonymie annotées manuellement. Cette évaluation a donné un rappel de 0.75 et une précision de 0.68 pour les patrons proposés. Nous continuons d'affiner les patrons afin d'éliminer certains faux positifs. L'évaluation a indiqué aussi que 33% des relations d'hyperonymie extraites sont absentes de la version actuelle de DBPedia, en considérant les relations de type *rdf :type* et *rdf :subclassOf*. Ce qui confirme que les textes de Wikipédia contiennent aussi des relations d'hyperonymie qui ne sont pas présentes dans les différents éléments (infobox, catégories, etc.).

## 5. Actions futures

Ce premier travail nous a permis d'établir une chaîne de traitement allant du prétraitement des pages Wikipédia jusqu'à la connexion avec DBPedia en français pour vérifier que les résultats serviront à enrichir cette base de connaissances. L'étape suivante va consister à combiner la méthode par patrons avec des techniques d'apprentissage en considérant en premier lieu la totalité des pages de désambiguïsation, et en second lieu le reste des pages Wikipédia pour identifier l'impact de la structure textuelle. Nous envisageons dans un troisième temps de diversifier les méthodes utilisées, en intégrant en particulier des indices de type distributionnel et des informations relatives à la structure textuelle (titres, énumérations, etc.). La dernière étape va consister à intégrer toutes ces méthodes dans une chaîne cyclique de façon à ce qu'une méthode donnée soit capable d'améliorer ses performances en utilisant les résultats d'une autre méthode. Enfin, nous nous consacrerons à l'intégration des nouvelles connaissances extraites dans la base de connaissances DBPedia (enrichir l'ontologie DBPedia, garantir sa cohérence, attacher chaque triplet au bon concept, etc.)

## 6. Bibliographie


Akbik A., Visengeriyeva L., Herger P., Hemsen H., Löser A. et al., « Unsupervised Discovery of Relations and Discriminative Extraction Patterns. », *COLING*, p. 17–32, 2012.

Aubin S., Hamon T., « Improving term extraction with terminological resources », *Advances in Natural Language Processing*, Springer, p. 380–387, 2006.

Fauconnier J.-P., Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes : exploitation des structures énumératives, Thèse de doctorat, Université de Toulouse, Toulouse, France, janvier, 2016.

Fauconnier J.-P., Kamel M., « Discovering Hypernymy Relations using Text Layout », *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, Association for Computational Linguistics, Denver, Colorado, p. 249–258, June, 2015.

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

Jacques M.-P., Aussenac-Gilles N., « Variabilité des performances des outils de TAL et genre textuel », *Traitement automatique des langues*, vol. 47, n° 1, p. 11–32, 2006.

Kazama J., Torisawa K., « Exploiting Wikipedia as external knowledge for named entity recognition », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 698–707, 2007.

Lenci A., Benotto G., « Identifying hypernyms in distributional semantic spaces », *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*, Association for Computational Linguistics, p. 75–79, 2012.

Malaisé V., Zweigenbaum P., Bachimont B., « Detecting semantic relations between terms in definitions », *COLING*, p. 55–62, 2004.

Mintz M., Bills S., Snow R., Jurafsky D., « Distant supervision for relation extraction without labeled data », *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*, Association for Computational Linguistics, p. 1003–1011, 2009.

Morsey M., Lehmann J., Auer S., Stadler C., Hellmann S., « Dbpedia and the live extraction of structured data from wikipedia », *Program*, vol. 46, n° 2, p. 157–181, 2012.

Navigli R., Velardi P., « From glossaries to ontologies : Extracting semantic structure from textual definitions », *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*p. 71–87, 2008.

Ritter A., Soderland S., Etzioni O., « What Is This, Anyway : Automatic Hypernym Discovery. », *AAAI Spring Symposium : Learning by Reading and Learning to Read*, p. 88–93, 2009.

Schropp G., Lefever E., Hoste V., « A Combined Pattern-based and Distributional Approach for Automatic Hypernym Detection in Dutch. », *RANLP*, RANLP 2013 Organising Committee / ACL, p. 593-600, 2013.

Snow R., Jurafsky D., Ng A. Y., « Learning syntactic patterns for automatic hypernym discovery », *Advances in Neural Information Processing Systems 17*, 2004.

Suchanek F. M., Kasneci G., Weikum G., « Yago : A Core of Semantic Knowledge Unifying WordNet and Wikipedia », *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706, 2007.

Sumida A., Torisawa K., « Hacking Wikipedia for Hyponymy Relation Acquisition. », *IJCNLP*, vol. 8, Citeseer, p. 883–888, 2008.