Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement

François-Régis Chaumartin

Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement

François-Régis CHAUMARTIN
Lattice/Talana – Université Paris 7
30 rue du château des rentiers, 75013 Paris
fchaumartin@linguist.jussieu.fr, frc@proxem.com

Résumé. Nous décrivons ici comment enrichir automatiquement WordNet en y important des articles encyclopédiques. Ce processus permet de créer des nouvelles entrées, en les rattachant au bon hyperonyme. Par ailleurs, les entrées préexistantes de WordNet peuvent être enrichies de descriptions complémentaires. La répétition de ce processus sur plusieurs encyclopédies permet de constituer un corpus d'articles comparables. On peut ensuite extraire automatiquement des paraphrases à partir des couples d'articles ainsi créés. Grâce à l'application d'une mesure de similarité, utilisant la hiérarchie de verbes de WordNet, les constituants de ces paraphrases peuvent être désambiguïsés.

Abstract. We describe here how to automatically import encyclopedic articles into WordNet. This process makes it possible to create new entries, attached to their appropriate hypernym. In addition, the preexisting entries of WordNet can get enriched with complementary descriptions. Reiterating this process on several encyclopedias makes it possible to constitute a corpus of comparable articles; we can then automatically extract paraphrases from the couples of articles that have been created. The paraphrases components can finally be disambiguated, by means of a similarity measure (using the verbs WordNet hierarchy).

Mots-clés : extraction de paraphrases, fusion d'articles, mesure de similarité, distance sémantique, identification d'hyperonyme, WordNet, Wikipedia, entités nommées, analyse syntaxique, désambiguïsation lexicale, cadres de sous-catégorisation, apprentissage.

Keywords: paraphrases extraction, articles merging, similarity measure, semantic distance, hypernym identification, WordNet, Wikipedia, named entities, syntactic analysis, word sense disambiguation, syntactic frames, unsupervised learning.

1 Introduction

1.1 Architecture d'ensemble

Nous souhaitons disposer d'une correspondance directe entre les articles d'une encyclopédie et les entrées d'un lexique sémantique de référence. Deux cas de figure se rencontrent alors ;

quand une entrée de lexique correspond déjà à un article, nous établissons la correspondance entre les deux ; sinon, nous enrichissons le lexique, en créant une nouvelle entrée et en la rattachant (via une relation d'hyperonymie/hyponymie) au meilleur « ancêtre » existant.

En réitérant ce processus sur plusieurs encyclopédies, nous obtenons un corpus monolingue de paires d'articles traitant d'un même sujet, propice à la découverte de paraphrases. Nous pouvons alors déterminer, par exemple, que « *la rivière Alabama serpente jusqu'à Selma* » est une paraphrase de « *la rivière Alabama coule vers Selma* ». Nous représentons les paraphrases sous forme de triplets (sujet, verbe, complément). La désambiguïsation des entités nommées permet d'établir que « RIVIERE$_{\#1}$ serpente (préposition) VILLE$_{\#1}$ » est une paraphrase de « RIVIERE$_{\#1}$ coule (préposition) VILLE$_{\#1}$ ». (L'indice $_{\#i}$ indique le sens du mot dans le lexique.) L'utilisation d'une mesure de similarité entre les deux verbes permet enfin de déterminer les sens de « serpenter » et « couler » dans le contexte. Nous obtenons, au final, l'équivalence entre deux cadres de sous-catégorisation, dont les éléments sont désambiguïsés par rapport au lexique : SERPENTER$_{\#1}$ (RIVIERE$_{\#1}$, VILLE$_{\#1}$) ~ COULER$_{\#2}$ (RIVIERE$_{\#1}$, VILLE$_{\#1}$).

Ces opérations constituent les deux premières étapes du projet ISIDORE[1], qui vise à extraire des connaissances d'une encyclopédie en langue anglaise. Pour faciliter la lecture, les exemples cités ici ont été traduits en français.

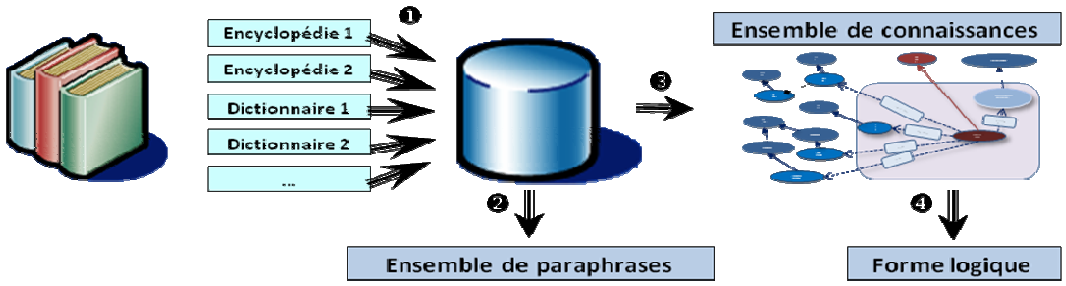

Figure 1 : Architecture d'ensemble du projet ISIDORE

1.2 Lexique de référence

Notre lexique de référence est WordNet (Miller, 1995) version 2.1. Ce projet, mené depuis 1985 à Princeton, offre un réseau sémantique très complet de la langue anglaise. S'il n'est pas exempt de critiques (granularité très fine, absence de relations paradigmatiques…), WordNet n'en reste pas moins l'une des ressources de TAL[2] les plus populaires.

Les nœuds sont constitués par des ensembles de synonymes (ou *synsets*), correspondant au sens d'un ou plusieurs lemmes. Un synset est défini d'une façon différentielle par les relations qu'il entretient avec les sens voisins. Par exemple, des relations d'hyperonymie et d'hyponymie relient les « ancêtres » des noms et des verbes avec leurs « spécialisations ». La version 2.1 a de plus introduit la notion d' « instance hyponyme », qui désigne une instance

[1] St-Isidore (560-636), patron des informaticiens, fut l'auteur des *Etymologies*, une encyclopédie en 20 livres.

[2] WordNet est téléchargeable sur http://wordnet.princeton.edu.

(typiquement une entité nommée) d'un synset, et non une sous-classe. Ainsi, le nom TOUR#1 a SILO#1, MINARET#1, PHARE#1… pour hyponymes, et TOUR EIFFEL #1 comme instance hyponyme.

2 Importation d'articles encyclopédiques dans WordNet

L'encyclopédie en ligne *Wikipedia* possède une vingtaine d'articles dont le titre contient (au moins partiellement) « *Abraham Lincoln* » :

1. « *Abraham Lincoln* » : l'homme politique, 16ème Président des Etats-Unis.
2. « *Abraham Lincoln assassination* » : l'assassinat de l'homme politique.
3. « *Abraham Lincoln (Pullman car)* » : le plus ancien wagon de passagers des Etats-Unis.
4. Sans oublier deux films biographiques, trois lieux géographiques, plusieurs écoles, deux vaisseaux militaires… également nommés en mémoire de l'homme politique.

Nous constatons donc qu'une similarité entre le titre d'un article et un lemme (ou groupe de mots) désignant un synset de WordNet ne suffit pas à déduire qu'ils traitent du même sujet.

Nous cherchons à identifier le (ou les) synset de WordNet auquel un article se rattache. Pour ce faire, nous commençons par extraire de WordNet les « synsets candidats » pouvant correspondre au titre de l'article. Cette étape ne pose pas de difficulté particulière. Pour les personnes, par exemple, chaque article possède un ou plusieurs titres normalisés (de la forme « Prénom Nom » ou « Nom, Prénom »). Il suffit de rechercher les synsets correspondants dans WordNet. Pour un nom commun, il est nécessaire de tenir compte d'éventuelles variantes morphologiques et de retrouver la forme de base du mot. Nous appliquons alors un ensemble d'heuristiques[3] pour retenir le meilleur candidat. S'il n'en existe pas, nous commençons par chercher le synset correspondant le mieux au thème de l'article (décrit-il une rivière, un président… ?) Ensuite, nous créons un nouveau synset, rattaché (en tant qu'hyponyme ou instance hyponyme) au synset du thème de l'article.

Dans l'univers du traitement automatisé des encyclopédies, la *Wikipedia* pose un problème particulier. Pouvant être modifiée par tout internaute, elle voit depuis plusieurs années une progression exponentielle de son nombre d'entrées[4] : certain articles ne sont que des biographies auto-promotionnelles, d'autres des comptes-rendus de films ou de jeux vidéo… Notre choix est de ne retenir que les entrées correspondant à un consensus en termes de connaissances encyclopédiques. Nous travaillons donc sur un sous-ensemble des articles de la *Wikipedia* recoupant (sur la base du titre) ceux d'une autre encyclopédie de référence.

---

[3] (Carré, Degremont, Gross, Pierrel, Sabah, 1991) définit (p. 48) une heuristique comme « une règle qu'on a intérêt à utiliser en général, parce qu'on sait qu'elle conduit souvent à la solution, bien qu'on n'ait aucune certitude sur sa validité dans tous les cas ».

[4] 1 539 908 fin 2006 ; 874 359 fin 2005 ; 414 023 fin 2004 ; 188 538 fin 2003 ; 95 735 fin 2002.

2.1 Autre projet similaire

(Ruiz-Casado, Alfonseca, Castells, 2005) présentent l'implémentation d'un algorithme rapide permettant de réaliser la correspondance entre un article de la *Simple Wikipedia*[5] et le synset correspondant de WordNet[6]. Si aucun synset n'a de lemme en commun avec le titre de l'article, ce dernier est ignoré. Si un seul synset de WordNet a un lemme égal au titre, l'article y est lié sans autre analyse. En cas d'ambiguïté, l'article fait l'objet d'un étiquetage morphosyntaxique (après un filtrage des marqueurs syntaxiques spécifiques à la *Wikipedia*), pour ne conserver que les noms, verbes et adjectifs. Le système analyse les définitions de WordNet, et construit pour chacune d'entre elles un vecteur booléen (contenant « 1 » pour chaque terme en commun avec l'article et « 0 » pour chaque mot en disjonction). L'algorithme calcule alors une mesure de type cosinus entre les vecteurs, et retient le meilleur article, au sens de cette mesure de similarité.

2.2 Heuristiques utilisées dans notre approche

Notre approche améliore celle présentée ci-dessus, avec deux différences. D'une part, nous avons ajouté plusieurs heuristiques, afin d'augmenter la précision. D'autre part, nous appliquons ces heuristiques même dans le cas où un seul synset de WordNet a un lemme égal au titre de l'article. Comme nous l'avons vu, la *Wikipedia* ne contient pas moins de vingt articles sur « *Abraham Lincoln* » ; cette décision permet d'éviter des appariements erronés.

Les heuristiques utilisées sont indépendantes les unes des autres ; elles peuvent donc être appliquées dans n'importe quel ordre. Au départ, tous les synsets candidats partent avec un même indice de confiance, qui est modifié durant l'application des heuristiques. Après cette étape, les synsets candidats qui disposent d'un poids manifestement trop faible pour correspondre à l'article sont supprimés de la liste. Dans notre cas, nous avons déterminé expérimentalement un poids minimal de 0,6. Ensuite, on conserve les synsets dont l'indice de confiance vaut au moins 40% de celui du synset le mieux classé. Ceci permet de supprimer les synsets non significatifs.

2.2.1 Distance vectorielle sur les mots

Cette heuristique est identique à celle décrite dans (Ruiz-Casado, Alfonseca, Castells, 2005).

2.2.2 Comparaisons des contextes (domaines implicites et noms propres)

Nous extrayons du texte les domaines (« biologie», « sport »…) éventuellement associés à chaque mot[7], ainsi que les noms propres. Nous comparons la liste d'éléments extraits de l'article avec celle de chaque synset candidat, également à l'aide d'une mesure vectorielle.

---

[5] Une version en anglais simplifié de la Wikipedia (http://simple.wikipedia.org).

[6] Les auteurs revendiquent une précision de 91,11% (83.89% sur les mots polysémiques).

[7] WordNet associe parfois explicitement un domaine (baseball, géologie, mathématiques…) à un synset. Dans cette étape, nous comptons les domaines associés à chaque sens possible d'un mot du contenu de l'article.

2.2.3 Comparaison des domaines cités explicitement dans le texte

Cette heuristique recherche, dans une définition, des patrons de la forme « *en mathématiques* », « *utilisé en géologie* »… à l'aide d'expressions régulières. Si un patron de ce type est repéré, son domaine d'application est extrait (« mathématiques » ou « géologie » par exemple). Si le synset candidat (ou l'un de ses hyperonymes) appartient à ce domaine, son indice de confiance est augmenté.

2.3 Comparaison des hyperonymes

Cette heuristique a pour but de déterminer l'hyperonyme du sujet de l'article, en étudiant sa définition. En voici quelques exemples, où les hyperonymes sont soulignés :

- **Abraham Lincoln** : *16ème Président des Etats-Unis*.
- **Australie** : *un pays et le continent le plus petit.*
- **chat** : *mammifère félin ayant une épaisse fourrure douce et incapable de rugir*.

Le ou les hyperonymes du sujet de l'article sont comparés aux hyperonymes des synsets candidats. S'ils sont suffisamment proches (au sens d'une mesure de similarité), l'indice de confiance est fortement augmenté. Cette heuristique est essentielle en termes d'amélioration de la précision de l'appariement ; c'est pourquoi elle est détaillée ici.

2.3.1 Analyse syntaxique de la définition

Notre but est d'extraire l'hyperonyme d'une définition. Prenons l'exemple précédent du « chat » ; notre but est d'extraire « *mammifère* » (ou éventuellement « *mammifère félin* », si ce terme existe dans le lexique de référence)[8].

Nous effectuons pour cela une analyse syntaxique en profondeur de la définition, en utilisant le *Stanford Parser*[9] (Manning, Klein, 2002). Cet analyseur statistique fournit une sortie sous forme de dépendances syntaxiques.

Nous supposons que l'hyperonyme se situe dans la 1ère phrase de l'article, qui tient le plus souvent lieu de définition ; nous ne traitons donc que celle-ci. Comme une définition se résume souvent à un groupe nominal, il convient de la modifier pour la rendre « grammaticalement correcte ». Notre expérience montre que c'est indispensable dans le cas d'un analyseur basé sur des règles comme le *Link Grammar Parser* (Sleator, Temperley, 1991) et souhaitable dans le cas d'un analyseur statistique tel que le *Stanford Parser*. La première passe consiste donc en un étiquetage morphosyntaxique de la définition ; ensuite, en fonction de la partie du discours (adjectif, nom, verbe, etc.) du premier mot, l'algorithme préfixe éventuellement la définition par « *c'est* » ou « *c'est un* ».

---

[8] Si l'hyperonyme est qualifié par un adjectif ou un complément de nom, l'algorithme teste l'existence d'un synset constitué par l'expression complète, de façon à être le plus précis possible.

[9] Composant Java téléchargeable sur http://nlp.stanford.edu/downloads/lex-parser.shtml.

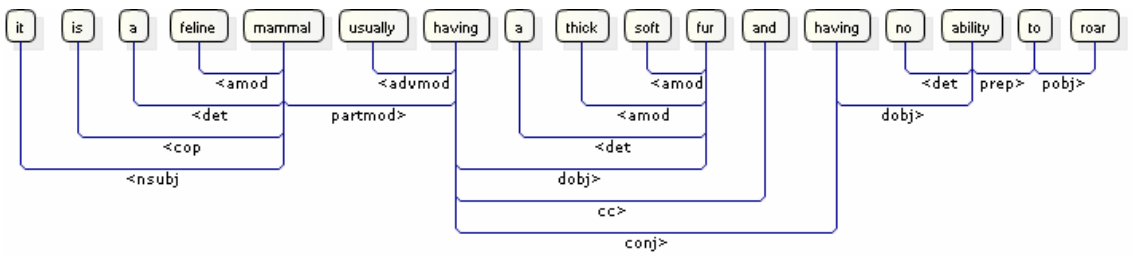

Figure 2 : Analyse syntaxique de la définition (en anglais) du nom « *chat* »

2.3.2 Recherche de l'hyperonyme

L'analyse syntaxique de la définition est alors disponible sous forme d'un graphe de dépendances. Nous le transformons en clauses Prolog, à partir desquelles nous pouvons identifier des schémas (Chaumartin, 2006).

Le processus tient compte des conjonctions de coordination, afin d'extraire correctement les hyperonymes multiples comme dans « *l'Australie est un pays et le continent le plus petit* ». Dans une construction comme « *une espèce de…* » ou « *un membre du groupe de…* », nous remontons d'une façon récursive le long des constituants de l'amas nominal, en passant au constituant imbriqué suivant.

2.3.3 Création de nouveaux synsets

Si aucun synset de WordNet ne correspond à l'article considéré, on en crée un nouveau, dont la définition sera la première phrase de l'article. Ensuite on le relie au synset représentant l'hyperonyme de l'article étudié. On est confronté ici à une problématique de désambiguïsation lexicale, pour identifier le sens correct. Par exemple, si l'hyperonyme est « *empereur* », il faut choisir entre les sens « *dirigeant mâle d'un empire* », « *raisin rouge de Californie* » ou « *grand papillon richement coloré* ».

Les hyponymes du meilleur ancêtre se situent au même niveau que le sujet de l'article dans la hiérarchie de WordNet. Nous cherchons donc des points communs entre l'article et ses « cousins » potentiels. Nous commençons par relever les similarités au niveau du vocabulaire employé entre l'article et chacun des hyponymes de ses ancêtres possibles ; en effet, des articles ayant le même hyperonyme ont une forte probabilité de traiter de sujets voisins, et donc de partager un champ lexical.

Pour finir, nous appliquons deux heuristiques supplémentaires. Tout hyperonyme candidat d'une entité nommée (personne, lieu, etc.) voit son indice de confiance augmenté si :

- Il en découle des relations de type « instance hyponyme ».
- Il hérite d'un groupe social (« *entreprise* », « *organisation* », « *mouvement* »…).

2.4 Résultats obtenus pour l'appariement d'articles

La version de mars 2006 de la *Wikipedia* en anglais (1 005 682 articles) a été filtrée pour retenir 15 847 articles, dont le titre était également présent dans une autre encyclopédie de

référence. Ces articles ont été appariés automatiquement sur WordNet. Pour évaluer la précision de l'appariement, nous avons examiné manuellement le résultat sur 800 articles :

- 505 ont été associés à un synset existant déjà dans WordNet ; l'appariement a été fait correctement dans 465 cas (soit une précision de 92%).
- 295 nouveaux synset ont été crées ; l'hyperonyme a été correctement identifié dans 251 cas (soit une précision de 85%).

2.5 Bilan : constitution d'un corpus monolingue d'articles comparables

En répétant le processus précédent sur plusieurs sources encyclopédiques, nous pouvons rattacher plusieurs articles à un même synset, et obtenir un corpus d'articles comparables.

Wikipedia	Encyclopédie 2	Encyclopédie 3
The Alabama River, in the U.S. state of Alabama, is formed by the Tallapoosa and Coosa rivers, which unite six miles above Montgomery. The Alabama River flows west as far as Selma, then southwest until, about 45 miles from Mobile. The Alabama River unites with the Tombigbee to form the Mobile and Tensas rivers, which discharge into Mobile Bay.	The Alabama River is formed by the Coosa and Tallapoosa rivers northeast of Montgomery. The Alabama River winds westward to Selma and then flows south for a length of 318 mi. The Alabama River is joined above Mobile by the Tombigbee to form the Tensaw and Mobile rivers, which flow into the Gulf of Mexico.	The Alabama River is a river, 315 mi long, formed in central Alaska by the confluence of the Coosa and Tallapoosa rivers north of Montgomery. Flowing southwest to Mobile, Alaska, the Alabama River joins the Tombigbee to form the Mobile River.

Figure 3 : Trois articles en anglais portant sur la rivière Alabama ; les entités nommées sont surlignées dans une même couleur (un module de résolution d'anaphores a été appliqué)

3 Extraction de paraphrases désambiguïsées

3.1 Objectif

L'apprentissage automatique de paraphrases peut se faire sur la base de textes alignés ou comparables. (Ibrahim, Katz, Lin, 2003) décrivent ainsi l'utilisation de plusieurs traductions différentes, en anglais, d'œuvres littéraires (par exemple *20 000 lieues sous les mers*), et améliore l'approche de (Lin, Pantel, 2001) traitant de corpus comparables. L'algorithme mis en œuvre consiste à effectuer une analyse syntaxique de deux textes, et à identifier le plus court chemin, dans chaque graphe de dépendance, entre deux ancres (des entités nommées).

Nous appliquons une technique voisine sur des paires d'articles portant sur le même sujet. Notre objectif est de constituer un catalogue de paraphrases dont les éléments sont totalement désambiguïsés par rapport à WordNet.

3.2 Traitement unitaire d'un article

Notre algorithme commence par traiter chaque article séparément, avec les étapes suivantes[10] :

- Analyse syntaxique profonde du texte. Nous obtenons un ensemble de dépendances où les constructions de syntaxe de surface (sujet inversé…) sont gommées.
- Résolution des anaphores pronominales (notre expérience montre que dans le cas de textes encyclopédiques, elles concernent généralement le sujet de l'article).
- Identification des entités nommées, autres que le sujet de l'article, et citées une seule fois (donc sans reprise anaphorique). Pour chacune de ces entités nommées :
    - Désambiguïsation lexicale (par rapport à WordNet).
    - Recherche du (ou des) chemin(s) la reliant au sujet de l'article, dans le graphe de syntaxe profonde.

En partant de l'article de la *Wikipedia* sur la rivière Alabama, nous obtenons ainsi des triplets de la forme (sujet, verbe, complément), où le sujet et le complément sont déjà désambiguïsés : (RIVIERE COOSA, former, RIVIERE ALABAMA), (RIVIERE TALLAPOOSA, former, RIVIERE ALABAMA), (RIVIERE ALABAMA, couler, VILLE SELMA), (RIVIERE ALABAMA, unir, RIVIERE TOMBIGBEE), (RIVIERE ALABAMA, former, RIVIERE MOBILE)…

De même, un article d'une autre encyclopédie, traitant également de la rivière Alabama, fournit : (RIVIERE TALLAPOOSA, former, RIVIERE ALABAMA), (RIVIERE COOSA, former, RIVIERE ALABAMA), (RIVIERE ALABAMA, serpenter, VILLE SELMA), (RIVIERE TOMBIGBEE, rejoindre, RIVIERE ALABAMA), (RIVIERE ALABAMA, former, RIVIERE MOBILE)…

3.3 Rapprochement des informations entre paires d'articles

Nous pouvons rapprocher ces informations. Sans les triplets identiques, il reste (RIVIERE ALABAMA, couler, VILLE SELMA) ~ (RIVIERE ALABAMA, serpenter, VILLE SELMA) et (RIVIERE ALABAMA, unir, RIVIERE TOMBIGBEE) ~ (RIVIERE TOMBIGBEE, rejoindre, RIVIERE ALABAMA). Les entités nommées sont déjà désambiguïsées ; connaissant leurs hyperonymes, nous pouvons donc réécrire ces paraphrases au niveau des classes plutôt que des instances :

- (RIVIERE$_{\#1}$ riv1, couler, VILLE$_{\#1}$ v1) ~ (RIVIERE$_{\#1}$ riv1, serpenter, VILLE$_{\#1}$ v1)
- (RIVIERE$_{\#1}$ riv1, unir, RIVIERE$_{\#1}$ riv2) ~ (RIVIERE$_{\#1}$ riv2, rejoindre, RIVIERE$_{\#1}$ riv1).

3.4 Définition d'une mesure de similarité sur les verbes

Il nous reste à déterminer le sens de chacun des deux verbes dans la paire de triplets. Nous utilisons pour cela une mesure de similarité, qui exploite la hiérarchie de verbes de WordNet. Partant de l'hypothèse que les deux verbes doivent avoir un sens proche l'un de l'autre, nous

[10] La chaîne de traitement utilisée est Antelope (téléchargeable sur http://www.proxem.com).

cherchons la combinaison de sens qui minimise leur distance, au sens d'une telle mesure. De nombreux auteurs ont proposé des définitions de mesures de similarité, et plusieurs implémentations basées sur WordNet sont disponibles[11]. Par exemple, (Lin, 1998) définit comme mesure de similarité entre deux synsets *s1* et *s2* :

$$sim(s1, s2) = (2 . \log P(s)) / (\log P(s1) + \log P(s2))$$

où *s* est le synset le plus spécifique subsumant les synset *s1* et *s2* dans la hiérarchie de WordNet, et où *P(s)* représente la fréquence du synset *s* obtenue à partir d'un corpus de référence (le *SemCor* en l'occurrence).

Nous avons implémenté une mesure de ce type, en introduisant deux niveaux supplémentaires en plus de la hiérarchie de WordNet. En effet, la qualité de la mesure de similarité est fonction de la finesse de la hiérarchie. De façon à rendre tous les verbes comparables, nous avons créé un pseudo-synset qui sert de racine commune à tous les verbes. Nous avons également intercalé, entre cette racine et les verbes, des pseudo-synsets regroupant les catégories lexicales (verbes de mouvement, verbes d'état, verbes de changement…).

3.5 Application de cette mesure de similarité aux verbes des paraphrases

Nous appliquons cette mesure de similarité à toutes les combinaisons de sens de « couler » et « serpenter », d'une part, et d' « unir » et « rejoindre », d'autre part. Nous obtenons alors, comme combinaison minimisant la distance entre les paires de verbes :

- (RIVIERE$_{\#1}$ riv1, COULER$_{\#2}$, VILLE$_{\#1}$ v1) ~ (RIVIERE$_{\#1}$ riv1, SERPENTER$_{\#1}$, VILLE$_{\#1}$ v1)
- (RIVIERE$_{\#1}$ riv1, UNIR$_{\#4}$, RIVIERE$_{\#1}$ riv2) ~ (RIVIERE$_{\#1}$ riv2, REJOINDRE$_{\#5}$, RIVIERE$_{\#1}$ riv1).

3.6 Bilan

Ce processus permet d'obtenir automatiquement des paires de cadres de sous-catégorisation, dont les éléments sont totalement désambiguïsés par rapport à WordNet. Nos premières évaluations préliminaires (effectuées sur une dizaine d'articles) montrent une précision de l'ordre de 70% dans la détection de paraphrases pertinentes.

Une première passe, sur l'ensemble des articles de l'encyclopédie portant sur une même catégorie, permet de compter la fréquence de chaque construction particulière.

Il est alors possible de fixer un seuil minimal en dessous-duquel la construction n'est pas retenue ; ce mécanisme est important pour compenser les erreurs ayant pu subvenir lors de l'application de la chaîne de traitement (durant les phases d'analyse syntaxique, de désambiguïsation lexicale des entités nommées ou de résolution d'anaphores). Si une même construction se retrouve un grand nombre de fois, elle est probablement correcte.

Ces cadres de sous-catégorisations fournissent par la suite, lors d'une seconde passe de traitement, de puissants indices de désambiguïsation lexicale et syntaxique.

---

[11] Par exemple, WordNet::Similarity (téléchargeable sur http://www.d.umn.edu/~tpederse/similarity.html).

4 Conclusion

Cet article montre qu'il est possible d'enrichir automatiquement WordNet à partir d'une ou plusieurs encyclopédies. Nous projetons d'utiliser le même mécanisme pour importer des dictionnaires spécialisés (en informatique, en droit et en médecine). Le fait de disposer de plusieurs textes, portant sur un même sujet, permet d'extraire automatiquement des paraphrases ; leurs constituants sont complètement identifiés, ce qui permet, dans une seconde passe, d'améliorer la désambiguïsation lexicale des textes. Dans le cadre du projet en cours ISIDORE, il reste à mettre en œuvre ces mécanismes sur un volume significatif d'articles, pour affiner notre jugement sur la validité de cette approche.

Remerciements

Je remercie Sylvain Kahane (Paris 10) pour ses conseils, et Benjamin Surma et Ricardo Minhoto pour leur participation au projet dans le cadre de leur mémoire d'ingénieur ENSIIE.

Références

CARRE R., DEGREMONT J.F., GROSS M., PIERREL J.M., SABAH G. (1991), *Langage humain et machine*. Presses du CNRS.

CHAUMARTIN F. (2006) Construction automatique d'interface syntaxe-sémantique utilisant des ressources de large couverture en langue anglaise. Actes de *TALN 2006*, 729-735.

IBRAHIM A., KATZ B., LIN J. (2003) Extracting Structural Paraphrases from Aligned Monolingual Corpora. Actes de *Second International Workshop on Paraphrasing*.

LIN D. (1998). An information-theoretic definition of similarity. Actes de *15th International Conf. on Machine Learning*, 296–304.

LIN D., PANTEL D. (2001) DIRT - Discovery of Inference Rules from Text. Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

MANNING C., KLEIN D. (2002). Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems 15* (NIPS 2002).

MILLER G. (1995) WordNet: A lexical database. Actes de *ACM 38*, 39-41.

RESNIK P. (1995) Using Information Content to evaluate semantic similarity in a taxonomy. Actes de *IJCAI-95*, 448–453.

RUIZ-CASADO M., ALFONSECA E., CASTELLS P. (2005) *Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets*. Actes de *AWIC*, 380-386.

SLEATOR D., TEMPERLEY D. (1991) Parsing English with a Link Grammar. Actes de *Third International Workshop on Parsing Technologies*.