Constitution d'une ressource sémantique issue du treillis des catégories de Wikipedia

Olivier Collin (1)
Benoît Gaillard (2)
Jean-Léon Bouraoui (3)

(1) Orange Labs – Av Pierre Marzin, 22300 Lannion
olivier.collin@orange-ftgroup.com
benoit.gaillard@orange-ftgroup.com
jeanleon.bouraoui@orange-ftgroup.com

Résumé Le travail présenté dans cet article s'inscrit dans le thème de l'acquisition automatique de ressources sémantiques s'appuyant sur les données de Wikipedia. Nous exploitons le graphe des catégories associées aux pages de Wikipedia à partir duquel nous extrayons une hiérarchie de catégories parentes, sémantiquement et thématiquement liées. Cette extraction est le résultat d'une stratégie de plus court chemin appliquée au treillis global des catégories. Chaque page peut ainsi être représentée dans l'espace de ses catégories propres, ainsi que des catégories parentes. Nous montrons la possibilité d'utiliser cette ressource pour deux applications. La première concerne l'indexation et la classification des pages de Wikipedia. La seconde concerne la désambiguïsation dans le cadre d'un traducteur de requêtes français/anglais. Ce dernier travail a été réalisé en exploitant les catégories des pages anglaises

Abstract This work is closely related to the domain of automatic acquisition of semantic resources exploiting Wikipedia data. More precisely, we exploit the graph of parent categories linked to each Wikipedia page to extract the semantically and thematically related parent categories. This extraction is the result of a shortest path length calculus applied to the global lattice of Wikipedia categories. So, each page can be projected within its first level categories, and in addition their parent categories. This resource has been used for two kinds of applications. The first one concerns the indexation and classification of Wikipedia pages. The second one concerns a disambiguation task applied to a query translator for cross lingual search engine. This last work has been performed by using English categories lattice.

Mots-clés : Wikipedia, plus court chemin, désambiguïsation, classification, traduction de requête.

Keywords : Wikipedia, shortest path, disambiguation, classification, query translation.

Introduction

Nous inscrivons ce travail dans le cadre très vaste de la constitution de ressources sémantiques lexicales. Il s'agit d'affecter à chaque entrée d'un lexique pré-identifié une ou plusieurs étiquettes permettant de la caractériser au sein d'une taxonomie ou un treillis organisé hiérarchiquement. Cette représentation doit notamment permettre d'abstraire et généraliser l'espace des entrées lexicales, différencier les entrées de même forme (homonymes) et regrouper les entrées de sens équivalent (synonymes). Cette approche est généralement issue de la communauté linguistique qui vise une modélisation exhaustive et précise des entrées lexicales. Une approche alternative consiste à plonger les entrées lexicales dans un espace vectoriel. La plupart du temps chaque entrée est modélisée par un vecteur issu du comptage de ses voisins qui co-occurrent dans une collection de documents. Des techniques standards de data-mining ou clustering permettent ensuite de projeter chaque entrée dans un sous-espace qui peut être hiérarchisé ou non. Une mesure est généralement associée à ces représentations de manière à exprimer une proximité des entrées lexicales dans l'espace correspondant. Cette mesure permet ensuite de traiter des problèmes tels que l'expansion sémantique (proximité sémantique) ou la désambiguïsation (différenciation sémantique). Nous proposons donc ici une alternative à cet espace de représentation : un sous ensemble du treillis des catégories de Wikipedia. Depuis sa création, de nombreux auteurs (Suchanek et al, 2008), (Mihalcea, 2007), (Ponzetto et al., 2007), (Zesch et al., 2007), (Strube et al. , 2006), cherchent à exploiter les données associées aux pages de Wikipedia pour constituer une ressource sémantique exploitable. Notre travail se situe aussi dans cette ligne : nous extrayons automatiquement, sans aucune intervention humaine, un sous-treillis dont la structure produite, bien que linguistiquement imparfaite, permet dans de nombreux cas d'observer des relations d'hyperonymie et de catégorisation thématique pertinentes.

1 Génération des données

1.1 Description du treillis des catégories de Wikipedia

Chaque page de Wikipedia est indexée par un ensemble de catégories mères visibles et cliquables par l'utilisateur en bas de chaque page. La page "*Tom Cruise*" est ainsi indexée par les catégories suivantes : *Acteur américain, Producteur américain, Naissance dans l'État de New York, Naissance en 1962, Personnalité américaine d'origine allemande, Personnalité américaine d'origine britannique, Personnalité américaine d'origine irlandaise, Scientologie.* Ces catégories de bas de page expriment majoritairement un ou plusieurs rôles sémantiques. Chacune de ces catégories constitue à son tour une page de Wikipedia qui possède elle-même des "catégories mères". Cette hiérarchie de catégories n'est pas une taxonomie rigoureusement construite, les catégories et leurs liens hiérarchiques sont ajoutés librement par les contributeurs les plus variés, ce qui fait toute la richesse spécifique de cette structure mais obscurcit les relations sémantiques exploitables sous-jacentes (Guégan 2006), (Strube 2006). Cet espace de catégories constitue un treillis, c'est-à-dire un graphe orienté, partant des pages et aboutissant à une catégorie mère supérieure unique, la catégorie "*Article*". Ce treillis est linguistiquement hiérarchisé, chaque catégorie associée à une page possède des "catégories mères" qui généralisent chaque "catégorie fille" suivant un axe thématique ou un axe d'hyperonymie. A ce stade de nos travaux, nous ne pouvons dissocier automatiquement ces deux axes. Pour chaque catégorie, plusieurs généralisations s'effectuent en parallèle. Par exemple, dans le cas de la page "*Tom Cruise*", un axe de généralisation de la catégorie *Acteur_américain* est : *Artiste_américain>Art_aux_Etats-Unis>Art_par_pays>Art>Article*.

1.2 Constitution du treillis

Les données brutes permettant de constituer le treillis sont issues de deux tables[1] provenant du site de téléchargement des données françaises de Wikipedia[2]. Les jointures adéquates entre ces tables permettent de mettre en relation chaque page ou catégorie de Wikipedia avec ses catégories mères, constituant une représentation "à plat" du treillis. Le lien page/catégories mères est univoque, et la combinatoire du treillis est liée aux liens catégories/catégories, nous avons donc choisi de dissocier les liens page/catégorie des liens catégorie/catégorie. Nous obtenons au final, d'une part 873 468 pages pointant sur 3 770 343 catégories mères directes (en moyenne 4.31 catégories par page), d'autre part 119 492 catégories et 244 817 liens entre catégories (en moyenne 2.04 catégories mères directes par catégorie). Notre but étant d'étendre l'espace des catégories de bas de page, seul le treillis des catégories a ensuite été traité. Les catégories de portails ont aussi été utilisées. Le nombre de chemins possibles atteignant la catégorie "*Article*" peut être très important, souvent plusieurs dizaines ou centaines de chemins. Cette information trop dense et finalement peu informative est par conséquent inexploitable telle quelle. Techniquement, la structure "à plat" du treillis ne permet pas une navigation efficace. Nous avons utilisé le package NetworkX[3] pour charger le treillis des catégories en mémoire (119 492 nœuds et 244 817 arcs) et utiliser ses fonctionnalités de parcours de graphes pour réaliser notre filtrage. Nous avons ensuite tenté d'effectuer un choix pertinent dans cet espace de manière à constituer un sous-treillis du treillis global.

1.3 Extraction d'un sous-treillis

L'extraction du sous-treillis a été réalisée en utilisant une hypothèse forte : l'information pertinente est portée par les chemins les plus courts, partant de chaque catégorie mère associée à la page courante, atteignant la catégorie "*Article*". Toutefois, pour les pages françaises de Wikipedia, nous avons constaté empiriquement que les chemins qui atteignent la catégorie "*Article*" étaient moins pertinents que les chemins atteignant l'ensemble des catégories pointées par la page *Wikipedia:Catégories*[4]. Nous avons donc utilisé les 150 "catégories terminales" CAT_TERM de cette page : *Mouvement culturel, Art contemporain, Artisanat, Design, Art par pays, Rayonnement culturel* ….. L'heuristique est la suivante :

Pour chaque page
    Pour chaque catégorie de page
        Prendre les plus courts chemins atteignant CAT_TERM

Nous conservons tous les plus courts chemins de même longueur. Pour une page donnée, afin de limiter le nombre total de chemins par page, nous avons expérimentalement choisi de ne conserver que les 15 chemins les plus courts ainsi que d'éliminer les chemins de longueur supérieure à 8. Voici le résultat obtenu pour les pages associées à "*Avocat*" :

Avocat_(fruit)	*Fruit_alimentaire>Plante_alimentaire>Plante_utile>Agriculture*
Avocat_(métier)	*Métier_du_droit>Droit*
	Personnalité_du_droit>Droit

[1] *frwiki-latest-page.sql* et *frwiki-categorylinks.sql*
[2] http://download.wikimedia.org/frwiki/latest/.
[3] http://networkx.lanl.gov/
[4] http://fr.wikipedia.org/wiki/Wikipédia:Catégories

Nous obtenons un ou plusieurs arbustes, pointant vers une ou plusieurs catégories générales, qui décrivent à la fois des relations hyperonymiques et thématiques. Dans le cas d' "*Avocat*", ces données caractérisent bien le contexte sémantique et thématique des deux hypothèses. Dans d'autres cas, les données obtenues sont plus nombreuses mais restent pertinentes. Les données suivantes, plus fournies, caractérisent la page de "*Tom Cruise*" :

Acteur_américain>Artiste_américain>Art_aux_Etats-Unis>Art_par_pays
Acteur_américain>Acteur_par_nationalité>Acteur>Personnalité_de_la_tévision
Acteur_américain>Artiste_américain>Artiste_par_pays>Artiste
Personnalité_américaine_d'origine_allemande>Diaspora_allemande>Diaspora>Géographie_politique>Politique
Personnalité_américaine_d'origine_allemande>Diaspora_allemande>Diaspora>Migration>Société
Personnalité_américaine_d'origine_britannique>Personnalité_américaine_par_origine_ethnique_ou_nationale>Personnalité_par_origine_ethnique_ou_nationale>Migration>Société
Personnalité_américaine_d'origine_irlandaise>Diaspora_irlandaise>Diaspora>Géographie_politique>Politique
Personnalité_américaine_d'origine_irlandaise>Diaspora_irlandaise>Diaspora>Migration>Société
Portail:Cinéma>PortailArt>Portail:Culture
Portail:EtatsUnis>Portail:Amérique>Portail:Géographie>Portail_du_domaine_géographique
Producteur_américain>Cinéma_américain>Cinéma_aux_Etats-Unis>Cinéma_par_pays>Art_par_pays
Producteur_américain>Producteur_de_cinéma_par_nationalité>Producteur_de_cinéma>Producteur>Artiste
Scientologie>Groupement_spirituel>Spiritualité_autres>Spiritualité
Scientologie>Groupement_spirituel>Petit_mouvement_religieux>Religion

Ces données constituent, pour chaque page, un nouvel espace de représentation que nous pouvons maintenant utiliser pour traiter différents types de problèmes. Différentes mesures (cosinus, Jaccard…) peuvent être utilisées pour différencier ou regrouper les pages.

2 Cas d'usage

2.1 Indexation et classification

Nous avons indexé les pages de Wikipedia, non seulement à partir de leurs catégories parentes mais aussi à partir des catégories participant au sous-treillis associé. Nous pouvons ainsi récupérer directement toutes les pages indexées par "*Acteur_américain"*, "*acteur"* ou "*personnalités*". Ceci constitue la toute première utilisation de notre travail, Wikipedia ne proposant pas ce niveau d'indexation pour l'ensemble des pages. Nous avons ainsi récupéré directement le sous ensemble des pages indexées par le thème *"informatique*" (25 140 pages). Les chemins de toutes les catégories associées à une page ont ensuite été transformés en un seul "sac de catégories". Au final, chaque page est représentée par un vecteur de catégories contenant l'ensemble de toutes ses catégories parentes. Nous pouvons ainsi utiliser les techniques standards de classification ou de "data-mining" s'appuyant sur des vecteurs de traits. Bien que la hiérarchie soit ainsi perdue, le mélange de catégories très spécifiques telles que "*Matériel_informatique"* et de catégories très générales telles que "*Informatique"* reste une information très structurante. Par exemple, une partie du vecteur associé à "*Disque dur multimédia"* contient : *Matériel_audio-vidéo, Audiovisuel, Médias, électronique, Multimédia, Informatique, Stockage_informatique, Matériel_informatique, Techniques_et_sciences_appliquées, Stockage_informatique, Industrie, économie…*

Le regroupement des pages peut ensuite nous permettre de réaliser une forme d'expansion sémantique de requêtes portant sur les pages de Wikipédia en regroupant des pages partageant des catégories communes. La page *"Segment_de_réseau"* est ainsi voisine de la page *"Ethernet"* en partageant 14 catégories : *'électronique','Informatique','Télécommunications','Portail:Science', 'Protocole_réseau','Portail:Informatique','Normes_et_standards_informatiques','Portail:Technologie','Composant_électronique','Protocole_de_técommunication','Normalisation_des_técommunications','Techniques_et_sciences_appliquées','Matériel_informatique', 'Connectique'.* Cette expansion, appliquée à des pages de produits tels que les jeux vidéo, devient un système de recommandation. Les pages *'Mario_Golf'* et *'CyberTiger'*

correspondent ainsi à des produits similaires qui partagent les catégories suivantes : *'Informatique', 'Projet_jeu_vidéo', 'Golf', 'Jeu_Nintendo_64', 'Jeu_vidéo', "Application_de_l'informatique", 'Jeu_vidéo_sorti_en_1999', 'Jeu_vidéo_de_golf', 'Sport_individuel', 'Techniques_et_sciences_appliquées', 'Sport', 'Audiovisuel', 'Projet:Jeu_vidéo', 'Médias'.*

2.2 Désambiguïsation appliquée à la traduction de requêtes

Une désambiguïsation s'appuyant sur l'espace des catégories anglaises a été appliquée aux hypothèses de traduction des termes de requêtes d'un moteur cross-lingue utilisant nos ressources (Gaillard et al, 2010). L'exemple suivant est issu du sous-espace des catégories généré à partir des pages et catégories anglaises de Wikipedia[5], tel que décrit en section 1. La figure 1 montre la proximité des différentes pages de traduction de "avocat" et "Tom Cruise" dans l'espace des plus courts chemins sélectionnés (seules les catégories terminales sont représentées). Dans le cas présent la proximité entre "Tom Cruise" et "lawyer" est essentiellement due à leur hyperonyme commun "People" ainsi qu'à une catégorie intermédiaire de Wikipedia "People by occupation". Les chemins issus de "avocado" sont complètement disjoints de ceux qui sont issus de "Tom Cruise". Une distance est calculée entre chaque vecteur de catégories associé aux hypothèses de traduction de chaque terme. La distance la plus faible permet de choisir la bonne solution de traduction (ici "lawyer").

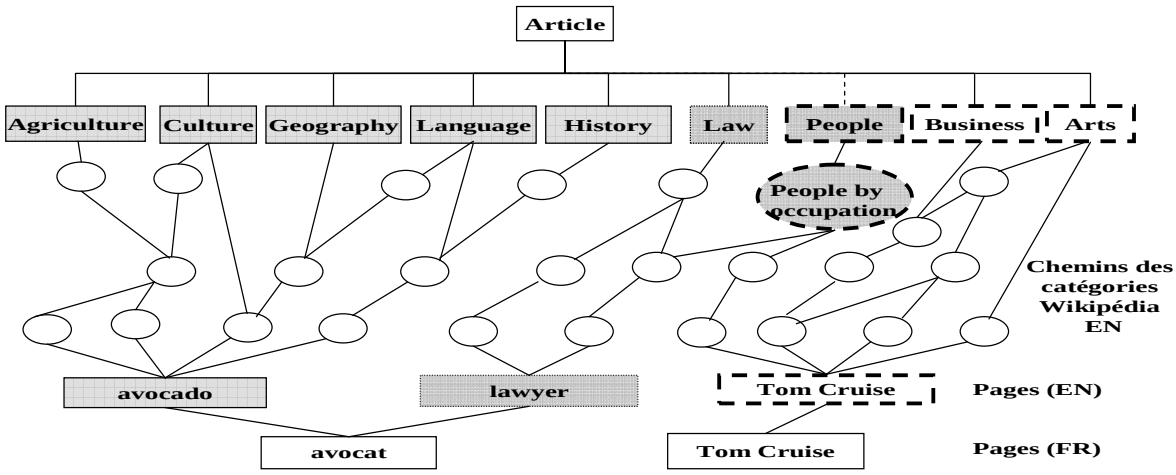

Figure 1 : Désambiguïsation de la requête "Avocat Tom Cruise".

Conclusion

Nous avons filtré le treillis des catégories de Wikipedia en utilisant une stratégie de plus court chemin. Le résultat est un sous-treillis permettant de prolonger chaque catégorie par quelques chemins qui généralisent chaque catégorie racine suivant une hiérarchie de relations hyperonymiques et thématiques. Bien que ces relations ne soient pas explicitement typées, l'espace de représentation généré est utilisable à des fins d'expansion sémantique et de désambiguïsation de traduction de requêtes portant sur des noms de pages de Wikipedia. Notre ressource est donc déjà opérationnelle. Les résultats obtenus par cette stratégie relativement simple semblent confirmer l'hypothèse sous jacente qui suppose que les chemins

[5] http://download.wikimedia.org/enwiki/latest/

pertinents sont les plus courts. Les perspectives concernent tout d'abord la validation et la recherche d'un fondement théorique à cette hypothèse, la théorie du *Minimum Description Length principle* semble être la voie naturelle. D'autre part, un travail d'analyse et de comparaison de ces données par rapport à d'autres ressources sera poursuivi. Nous pensons notamment mettre en œuvre un typage explicite automatique des différentes relations intervenant au sein du treillis.

Références

GAILLARD B., BOUALEM M., COLLIN O. (2010). Query Translation using Wikipedia-based resources for analysis and disambiguation. Actes de la conférence *14th Annual Conference of the European Association for Machine Translation (EAMT 2010), Saint Raphaël*, à paraître.

GLEDSON A., KEANE.J. (2008). Measuring Topic Homogeneity and its Application to Dictionary-Based Word-Sense Disambiguation. *Coling 2008, 22nd International Conference on Computa-tional Linguistics, Manchester,* UK 273–280.

GUEGAN M.(2006). Catégorisation par les contributeurs des articles de l'encyclopédie Wikipedia.fr. *Mémoire de master de recherche informatique université paris XI, LIMSI CNRS*

MIHALCEA R. (2007). Using Wikipedia for Automatic word Sense Disambiguation. Actes de *NAACL 2007,*196-203

NASTASE V., STRUBE M. (2008). Decoding Wikipedia catégories for knowledge acquisition, Actes de *23rd national conference on Artificial intelligence (AAAI 2008),* 1219-1224

PENG Y., MAO M. (2008), Blind Relevance Feedback with Wikipedia : Enterprise Track, *Proceedings of The Seventeenth Text REtrieval Conference (TREC 2008),* 18-21

PONZETTO S.P., STRUBE M. (2007). Deriving a large scale taxonomy from Wikipedia. AAAI'07. Actes de *22nd national conference on Artificial intelligence,* 1440-1445.

SCHÖNHOFEN P., BENCZUR A, BIRO I., CSALOGANY K. (2008). Cross-Language Retrieval with Wikipedia. *Advances in Multilingual and Multimodal Information Retrieval, Revised selected paper of CLEF 2007, Springer,* pp 72-79

STRUBE M., PONZETTO S. P. (2006). WikiRelate! : Computing Semantic Relatedness Using Wikipedia. Actes de *AAAI 2006,* 1419-1424

SUCHANEK F.M., KASNECI G.,WEIKUM G. (2008). Yago : A large Ontology from Wikipedia and WordNet. *Journal of Web semantics, Elsevier,*203-217

TIEN-CHIEN L., SHIH-HUNG W.(2008), Query Expansion via Link Analysis of Wikipedia for CLIR, *Proceedings of NTCIR-7 Workshop Meeting,* 125-131

ZESCH T, GUREVYCH I (2007) Analysis of the Wikipedia Category Graph for NLP Applications. Actes de *Workshop TextGraphs-2 : Graph-Based Algorithms for Natural Language Processing,* 1-8

ZESCH. T., GUREVYCH I.,MÜHLHÄUSER M. (2007). Analysing and Accessing Wikipedia as a Lexical Semantic Resource. Actes de *Data Structures for Linguistic Resources and Applications,* 197-205