TALN 2010, Montréal, 19-23 juillet 2010

Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques

Fatiha Sadat[1] et Alexandre Terrasa[1,2]

(1) Université du Québec à Montréal, 201 av. President Kennedy,
Montréal, QC, H3X 2Y3, Canada
(2) École Supérieure d'Informatique Appliquée EXIA.CESI
11 avenue Neil Armstrong, 33700 Mérignac, Bordeaux, France
sadat.fatiha@uqam.ca, alexandre.terrasa@viacesi.fr

Résumé Cet article présente une approche et des résultats utilisant l'encyclopédie en ligne Wikipédia comme ressource semi-structurée de connaissances linguistiques et en particulier comme un corpus comparable pour l'extraction de terminologie bilingue. Cette approche tend à extraire d'abord des paires de terme et traduction à partir de types des informations, liens et textes de Wikipédia. L'étape suivante consiste à l'utilisation de l'information linguistique afin de ré-ordonner les termes et leurs traductions pertinentes et ainsi éliminer les termes cibles inutiles. Les évaluations préliminaires utilisant les paires de langues français-anglais, japonais-français et japonais-anglais ont montré une bonne qualité des paires de termes extraits. Cette étude est très favorable pour la construction et l'enrichissement des ressources linguistiques tels que les dictionnaires et ontologies multilingues. Aussi, elle est très utile pour un système de recherche d'information translinguistique (RIT).

Abstract Multilingual linguistic resources are usually constructed from parallel corpora, but since these corpora are available only for selected text domains and language pairs, the potential of other resources is being explored as well. This article seeks to explore and exploit the idea of using multilingual web-based encyclopaedias such as Wikipedia as comparable corpora for bilingual terminology extraction. We propose an approach to extract terms and their translations from different types of Wikipedia link information and texts. The next step will be using a linguistic-based information to re-rank and filter the extracted term candidates in the target language. Preliminary evaluations using the combined statistics-based and linguistic-based approaches were applied on Japanese-French, French-English and Japanese-French. These evaluations showed a real open improvement and good quality of the extracted term candidates for building or enriching multilingual ontologies, dictionaries or feeding a cross-language information retrieval system with the related expansion terms of the source query.

Mots-clés : Terminologie bilingue, corpus comparable, Wikipédia, ontologie multilingue.
Keywords: Bilingual terminology, comparable corpora, Wikipedia, mulilingual ontologies.

1 Introduction

Wikipédia, une encyclopédie en ligne multilingue, offre en même temps un réservoir gigantesque de données multilingues qui peut être exploité par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que *Google*[1] ou *Yahoo*[2] ou même le moteur de recherche propre de Wikipedia[3].

L'objectif de notre travail est l'acquisition automatique d'une terminologie bilingue ou multilingue des articles de Wikipédia indépendamment des langues utilisées; mais appliquée pour cette première étude aux paires de langues français-anglais et plus tard au français-japonais. Ce travail sera utilisé pour la recherche d'information multilingue ainsi que pour la construction et l'enrichissement des ressources linguistiques multilingue telles que les dictionnaires et ontologies.

Le contenu de cet article se résume comme suit. La section 2 présente la ressource Wikipédia ainsi que les différentes étapes du processus d'extraction de terminologie bilingue en utilisant une combinaison d'approches statistique et linguistique. Nous montrons dans la section 3 les analyses et évaluations obtenues en implémentant notre approche sur différentes paires de langues incluant le français, le japonais et l'anglais ; et discutons d'autres résultats et extensions auxquels la présente étude est liée. La section 4 conclue cet article et donne des pointeurs et extensions pour le futur.

2 Wikipédia et le Processus d'Extraction de la Terminologie Bilingue

Wikipédia (*prononcé wikipe'dja ou vikipe'dja*) est une encyclopédie collective établie sur Internet, universelle, multilingue et fonctionnant sur le principe du wiki , c-à-d un site Web dont les pages sont modifiables (gratuitement) par tout ou partie des visiteurs du site. Nous exploitons l'aspect multilingue de cette ressource afin d'extraire de la terminologie qui pourra créer ou enrichir les ressources linguistiques existantes, comme les ontologies et dictionnaires multilingues. Dans notre approche, les liens au sein de la même langue seront utilisés afin de chercher l'information dans la langue précisée et ainsi parcourir le corpus monolingue. Aussi, les liens entre différentes langues seront utilisés afin de chercher l'information translinguistique.

En premier, nous considérons une requête composée de *n* mots en langue source *S* qu'on utilise pour interroger Wikipédia. L'article obtenu est utilisé pour la construction de notre corpus comparable en langue *S*. L'utilisation des liens interlangues pour la même requête permet de construire un corpus comparable en langues source *S* et cible *T*. De cette manière, nous aurons construit un premier article comparable pour la requête de départ.

L'exploitation des liens existants dans les articles obtenus en premier en langue *S* et *T,* servira à agrandir et à approfondir le contenu et la taille de notre corpus comparable. Dans cette étude, nous nous basons sur des profondeurs de la recherche documentaire sur Wikipédia afin d'exploiter les liens dans la même langue ; c-à-d l'article obtenu en premier aura une profondeur égale à 0, l'utilisation des liens de cet article aboutira à une profondeur égale à 1 et plus - l'utilisation de *n fois* les liens des articles aboutiront à une profondeur égale à *n*. Plus nous parcourant les profondeurs dans la même langue, plus la taille de notre corpus devient volumineuse et plus le corpus sera plus consistant et pertinent pour nos recherches. Le parcours en parallèle dans l'autre langue sera inévitable puisqu'il s'agit d'un corpus comparable (c-à-d dans deux langues au moins).

Le processus d'extraction de terminologie bilingue des documents de Wikipédia se base en premier sur une approche statistique présentée par Sadat et al. (2003, 2004) et plus tard sur une combinaison avec l'information linguistique afin de ré-ordonner et raffiner la terminologie résultante.

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://www.wikipedia.org/

La phase statistique reprend et adapte la méthode proposée par Sada et al. (2003), Dejean et al. (2002) et Morin et al. (2004), et a pour objectif de réaliser l'alignement des termes de la langue source avec ceux de(s) la langue(s) cible(s). En considérant les corpus comparables construits à partir des articles de Wikipédia, nous appliquons les étapes du processus d'extraction de terminologie bilingue comme suit:

1. *Repérages des termes sources et cibles des deux corpus comparables ;*
2. *Construction des vecteurs contextes dans les deux langues ;*
3. *Transfert du contenu des vecteurs contextes de la langue source vers la langue cible en utilisant les traducteur interlangue de Wikipédia*
4. *Construction des vecteurs de similarité en utilisant des mesures de similarité tel que le cosinus, la distance de Jaccard et le coefficient Dice ;*
5. *Les informations linguistiques des mots sources et cibles sont utilisées afin de ré-ordonner les termes et leurs traductions candidates et ainsi éliminer les termes cibles inutiles.*

3 Evaluations

Nous évaluons dans cette section l'approche décrite précédemment en utilisant des paires de langues incluant le français, l'anglais et le japonais ainsi que le site Wikipédia comme réservoir de corpus comparables. Différentes profondeurs, c-à-d tailles des corpus comparables serons considérées. Toutefois, faute de temps et des limitations de ressources matériels informatiques disponibles dans notre laboratoire, nous nous sommes arrêtés à la profondeur 3 avec des résultats prometteurs.

Les premiers résultats obtenus avec la paire de langues français-anglais et d'après l'exploitation des liens dans un même article (représenté par les profondeurs) sont décrits dans le tableau 1. D'autres exemples pour les paires de langues incluant le japonais vers le français ou l'anglais sont présentés dans les tableaux 2 et 3, respectivement. Nous avons exploré les différents termes extraits en langue cible et nous avons constaté qu'une ontologie bilingue (ou multilingue) pourrait être construite en suivant ce processus. Les termes extraits ont une certaine relation sémantique avec les termes sources et les articles qui leurs sont liés pourront être exploités dans le but d'extraire les relations sémantiques et ainsi construire une ontologie multilingue.

4 Conclusion

Nous avons proposé une première étude d'une encyclopédie en ligne multilingue afin d'extraire de la terminologie bilingue et possiblement multilingue en utilisant les liens inter-langues et différentes profondeurs des liens monolingues. Nos résultats sont très encourageants et prometteurs pour la construction d'une ontologie multilingue. Les prochaines évaluations porteront sur plus de profondeurs afin d'agrandir la taille des corpus comparables et ainsi obtenir une meilleure terminologie. La définition des relations sémantiques entre les différents termes en langues source et cible est à l'étude. Aussi, pour le moment, nous avons utilisé une seule requête pour interroger le site de Wikipédia et ainsi obtenir des corpus comparables. Une éventuelle extension sera l'utilisation de plusieurs requêtes et des évaluations pour déterminer la qualité des traductions. D'autres paires de langues sont aussi à l'étude.

Références

ADAR E., SKINNER M., WELD D. S. (2009). INFORMATION ARBITRAGE ACROSS MULTI-LINGUAL WIKIPEDIA, PROCEEDINGS OF THE SECOND ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, FEBRUARY 09-12, 2009, BARCELONA, SPAIN.

DEJEAN, H., GAUSSIER, E., SADAT, F. (2002). AN APPROACH BASED ON MULTILINGUAL THESAURI AND MODEL COMBINATION FOR BILINGUAL LEXICON EXTRACTION. IN PROCEEDINGS OF COLING'02, TAIWAN.

ERDMANN M., NAKAYAMA K., HARA T., NISHIO S. (2008B). EXTRACTION OF BILINGUAL TERMINOLOGY FROM A MULTILINGUAL WEB-BASED ENCYCLOPEDIA. J. INFORM. PROCESS.

Profondeur	Mot source (fr.)	Nombre de candidats (ang.)	Traduction idéale (ang.)	Rang
0	Organisation	14	Organization	1
	Maladie	101	disease	14
	Santé	89	Health	1
	Hôpital	19	Hospital	3
1	Admission	90	admission	1
	Algue	88	Algae	1
	Thérapie	52	Therapy	1
3	Abeille	443	Bee	1
	Narcotique	1656	Narcotic	1
	Assurance	289	Insurance	2

Tableau 1: Exemples de la terminologie bilingue (français-anglais) extraite suivant les différentes profondeurs des articles de Wikipédia

Profondeur	Mot source (jap.)	Nombre de candidats (fr.)	Traduction idéale (fr.)	Rang
0	感染	14	Infection	1
1	イングランド	16	Angleterre	2
	けっか	6	Résultat	1
	世界	14	Monde	1
3	アレルギー	236	Allergie	2
	セルロース	233	Cellulose	2
	ワクチン	102	Vaccin	1

Tableau 2: Exemples de la terminologie bilingue (japonais-français) extraite suivant les différentes profondeurs des articles de Wikipédia

Profondeur	Mot source (jap.)	Nombre de candidats (ang.)	Traduction idéale (ang.)	Rang
0	細菌	17	Bacteria	2
1	感染	36	Infection	2
	ドイツ	4	Germany	1
3	ワクチン	88	Vaccine	2
	ヒト	472	Human	1
	微生物	84	Microorganism	1

Tableau 3: Exemples de la terminologie bilingue (japonais-anglais) extraite suivant les différentes profondeurs des articles de Wikipédia

KUN Y., TSUJII J. (2009). BILINGUAL DICTIONARY EXTRACTION FROM WIKIPEDIA. (2009A). IN PROCEEDINGS OF MT SUMMIT XII PROCEEDINGS 2009.

MORIN, E., AND DAILLE, B. (2006). COMPARABILITE DE CORPUS ET FOUILLE TERMINOLOGIQUE MULTILINGUE. TRAITEMENT AUTOMATIQUE DES LANGUES (TAL), 47(1):113-136, 2006.

SADAT, F., YOSHIKAWA, M., UEMURA, S. (2003). LEARNING BILINGUAL TRANSLATIONS FROM COMPARABLE CORPORA TO CROSS-LANGUAGE INFORMATION RETRIEVAL: HYBRID STATISTICS-BASED AND LINGUISTICS-BASED APPROACH. IN PROCEEDINGS OF EACL'203, WORKSHOP ON INFORMATION RETRIEVAL WITH ASIAN LANGUAGES - VOLUME 11, SAPPORO, JAPAN. PAGES: 57–64.

SADAT, F. (2004). KNOWLEDGE ACQUISITION FROM COLLECTIONS OF NEWS ARTICLES TO CROSS-LANGUAGE INFORMATION RETRIEVAL. IN PROCEEDINGS OF RIAO 2004 CONFERENCE, AVIGNON, FRANCE, PP. 504-513.

VÉRONIS, J. (2000). PARALLEL TEXT PROCESSING: ALIGNMENT AND USE OF TRANSLATION CORPORA. DORDRECHT, KLUWER ACADEMIC PUBLISHERS