TALN 2010, Montréal, 19-23 juillet 2010

Traduction de requêtes à l'aide de Wikipédia pour la recherche d'information multilingue

Benoît Gaillard, Olivier Collin, Malek Boualem

Orange Labs – 2, Avenue Pierre Marzin, 22300 Lannion, France,
benoit.gaillard, olivier.collin, malek.boualem@orange-ftgroup.com

Résumé Cet article s'inscrit dans le domaine de la recherche d'information multilingue. Il propose une méthode de traduction automatique de requêtes basée sur Wikipédia. Une phase d'analyse permet de segmenter la requête en syntagmes ou unités lexicales à traduire en s'appuyant sur les liens multilingues entre les articles de Wikipédia. Une deuxième phase permet de choisir, parmi les traductions possibles, celle qui est la plus cohérente en s'appuyant sur les informations d'ordre sémantique fournies par les catégories associées à chacun des articles de Wikipédia. Cet article justifie que les données issues de Wikipédia sont particulièrement pertinentes pour la traduction de requêtes, détaille l'approche proposée et son implémentation, et en démontre le potentiel par la comparaison du taux d'erreur du prototype de traduction avec celui d'autres services de traduction automatique.

Abstract This work investigates query translation using only Wikipedia-based resources in a two steps approach: analysis and disambiguation. After arguing that data mined from Wikipedia is particularly relevant to query translation, we detail the implementation of the approach. In the analysis phase, queries are segmented into lexical units that are associated to several possible translations using a bilingual dictionary extracted from Wikipedia. During the second phase, one translation is chosen amongst the various candidates, based on consistency, asserted with the help of semantic information carried by categories associated to Wikipedia articles. These two steps take advantage of data mined from Wikipedia, which is very rich and detailed, constantly updated but also easy and free to access. We report promising results regarding translation accuracy.

Mots-clés : recherche d'information multilingue, traduction de requêtes, Wikipédia.
Keywords:, cross language information retrieval, query translation, Wikipedia.

Introduction

La quantité d'information accessible sur le web ou sur d'autres supports électroniques croit de manière significative et elle est disponible dans des langues de plus en plus variées. Le CLIR (Cross Language Information Retrieval) permet de trouver des contenus dont la langue est différente de celle dans laquelle la requête est formulée. Souvent, pour y parvenir, un système de CLIR se base sur la traduction de la requête ou des documents. Bien que des études (Clough, 2005) montrent que l'approche par traduction des contenus est légèrement meilleure que l'approche par traduction des requêtes, car les logiciels de traduction automatique peuvent s'appuyer sur le texte, alors que les requêtes ne sont généralement constituées que de quelques mots (Jansen et al., 2000), l'approche privilégiée par certains acteurs reste la traduction des requêtes, du fait de la croissance soutenue de la quantité de contenus à indexer et de la diversité linguistique du web. Les approches lexicales de traduction de requêtes rencontrent deux problèmes. D'une part, la couverture lexicale est un facteur limitant et coûteux à optimiser. D'autre part un même mot peut être traduit de différentes façons qui n'ont pas systématiquement le même sens.

L'encyclopédie en ligne Wikipédia permet de proposer des solutions à ces deux problèmes. Grâce à des millions de contributeurs, elle contient une quantité de connaissances conséquente, constamment mise à jour et publiquement accessible, ce qui permet d'en extraire aisément des dictionnaires multilingues dont la couverture lexicale est optimale. De plus, ces connaissances sont organisées par le biais de catégories fournies par les utilisateurs (Guégan, 2006). Le présent article décrit d'abord les données issues de Wikipédia et montre qu'elles ont des propriétés bien adaptées au traitement des requêtes. Il expose ensuite une méthode d'analyse qui permet de segmenter les requêtes en unités lexicales, et une stratégie de choix de traductions parmi plusieurs alternatives, par des techniques sémantiques s'appuyant sur les catégories des articles de Wikipédia. La méthode est validée par l'évaluation de la qualité de la traduction d'un corpus de requêtes.

1 Quelques aspects de l'état de l'art sur la traduction des requêtes

1.1 Approches lexicales de la traduction et de l'analyse de requêtes

Comme le montrent (Ballesteros et al. 1997), les approches de traduction de requêtes à l'aide de lexiques bilingues telles que (Etzioni et al., 2007) soulèvent trois types de difficultés: traiter les mots qui ne font pas partie du lexique ou disposer d'un lexique suffisamment exhaustif, résoudre les ambigüités et enfin reconnaître les syntagmes et les entités nommées. Ils proposent d'utiliser le contexte local des termes de la requête pour y ajouter des expansions avant et après traduction. Ces expansions permettent de préciser le sens avant de traduire et de minimiser les erreurs si malgré tout des sens non-pertinents ont été ajoutés par la traduction. Dans le même esprit, afin d'optimiser la traduction à l'aide de l'expansion de requêtes, nous avons récemment proposé une étude qui démontre l'intérêt de rapprocher le vocabulaire issu de la traduction des contenus avec celui des requêtes (Gaillard et al., 2010). D'autre part, pour traiter les syntagmes, les auteurs utilisent des syntagmes extraits, à l'aide de règles grammaticales simples, d'un corpus parallèle traduit manuellement. Grâce à la croissance forte du nombre de contributeurs, Wikipédia permet d'accéder à de grandes quantités d'informations d'ordre sémantique et de constituer des lexiques bilingues relativement riches (Zesh et al., 2007). Ainsi, (Jones et al., 2008) parviennent à compléter un moteur classique de traduction

automatique par un lexique de syntagmes issus de Wikipédia. Pour extraire les "phrases", ou "syntagmes" de la requête, les auteurs utilisent la méthode dite de *"Maximum forward matching"*, associée à des règles lexicales simples (Ballesteros et al., 1997). Cette méthode consiste à détecter le plus long syntagme possible dans la requête, en partant du début, puis à répéter récursivement l'opération sur le reste de la requête.

1.2 Approches sémantiques pour la désambiguïsation des requêtes

Une approche, proposée par (Mihalcea, 2007), consiste à désambiguïser les sens des mots à l'aide d'un outil statistique de classification entrainé sur un corpus extrait automatiquement de Wikipédia. Des mesures telles que la *"similarité sémantique"* et la *"relation sémantique"* (*"Semantic Relatedness"*), initialement développées par (Resnik, 1995) et par (Banerjee, Petersen, 2003) pour des wordnets (Fellbaum, 1998), peuvent être appliquées à des requêtes en s'appuyant sur Wikipédia (Strube, Ponzetto, 2007). (Banerjee, Petersen, 2003) lèvent les ambigüités de sens à l'aide de leur contexte par une extension de l'algorithme de (Lesk, 1986). La technique consiste à évaluer le recouvrement entre les mots du contexte du terme ambigu et ceux des définitions de chaque sens dans Wordnet. (Strube, Ponzetto, 2007) proposent et évaluent plusieurs méthodes pour calculer la proximité sémantique de deux mots à l'aide de Wikipédia: une mesure est basée sur la distance à parcourir dans la *folksonomie* de Wikipédia (nombre d'arêtes entre deux nœuds du graphe), une autre mesure se base sur l'information (probabilité d'occurrences) et la dernière adapte à Wikipédia l'approche de (Banerjee, Petersen 2003).

(Bunescu, Pasca, 2006) proposent une méthode de reconnaissance d'entités nommées par un dictionnaire extrait de Wikipédia en prenant en compte les redirections et les pages d'homonymie. La désambiguïsation du sens d'une entité nommée est effectuée à l'aide de la similarité cosinus entre les mots du contexte autour de l'entité nommée (dans la requête) et les mots de l'article correspondant à l'EN candidate. Cette méthode est optimisée en utilisant aussi les catégories des articles de manière pondérée. (Schönhofen et al., 2008) s'appuient sur les articles, considérés comme des concepts, de Wikipédia en langue cible pour désambigüiser puis reformuler les traductions des requêtes. Pour chaque alternative de traduction de chaque mot de la requête, les concepts Wikipédia associés sont sélectionnés. Chaque concept est ensuite noté en fonction de ses liens (définis par les liens hypertextes entre articles de Wikipédia) avec les autres concepts sélectionnés. Une requête est alors générée à partir des concepts les mieux connectés. Cette méthode de désambigüisation s'appuie sur l'*homogénéité thématique* (*topic homogeneity*) (Gledson, Keane, 2008). L'homogénéité thématique peut s'appliquer à des requêtes ou à des textes, en s'appuyant sur des mesures de proximité de mots deux à deux telles que la *"semantic relatedness"*. (Nguyen et al., 2008) s'affranchissent de la traduction lexicale en projetant les requêtes sur les concepts de Wikipédia.

1.3 Spécificités de l'approche proposée par rapport à l'état de l'art

Notre méthode combine plusieurs aspects des techniques que nous venons de mentionner de manière à offrir une solution originale à la traduction de requêtes. Comme (Bunescu, Pasca, 2006), nous utilisons la similarité cosinus pour la désambigüisation, mais, alors qu'ils n'utilisent les catégories que pour pondérer une évaluation de la proximité sémantique de deux mots basée principalement sur les textes des articles, nous fondons notre mesure de proximité uniquement sur les catégories. De plus leur approche n'est pas appliquée à la traduction de requêtes. L'approche de (Schönhofen et al., 2008) consiste à sélectionner les alternatives de traduction par homogénéité thématique. Notre approche s'en inspire, mais nous ne

reformulons pas la requête à l'aide de concepts et notre mesure de proximité est différente. D'autre part la détection de syntagme évoquée par (Jones et al. 2008), (Ballesteros, Croft et al. 1997) n'est pas aussi développée que dans la méthode que nous proposons. La plupart des approches du CLIR par traduction lexicale de requêtes s'appuient accessoirement sur Wikipédia, contrairement à notre approche qui cherche précisément à optimiser l'apport de Wikipédia en terme de connaissances lexicales et sémantiques.

2 Wikipédia: une ressource pour le traitement des requêtes

2.1 Propriétés lexicales des titres d'articles

Les conventions au sujet du nommage des articles de Wikipédia sont définies sur la page explicative de Wikipédia[1]. En particulier, le titre idéal est le titre le plus court qui définit précisément le sujet, ne commence pas par un article grammatical, sauf lorsqu'il fait partie intégrante du sujet, et s'intègre naturellement dans une phrase. Si plusieurs titres sont possibles, le plus commun devrait être utilisé**,** par application du *principe de moindre surprise*[2]. En anglais: "*using names and terms that readers are most likely to look for in order to find the article*". Ces conventions ont pour conséquence qu'une forte proportion de titres d'articles sont des entités nommées, des noms communs et des groupes nominaux. Les requêtes sont en grande majorité constituées d'entités nommées et de groupes nominaux. Cee constat est confirmé par (Jones et al., 2008), selon qui 90% des requêtes comportent au moins un syntagme ou une entité nommée. De plus (Jansen et al., 2000) ont constaté que les requêtes formées de 1 à 4 mots sont très largement majoritaires. Un utilisateur aura tendance à formuler la requête la plus courte possible, sans déterminant sauf si par exemple la requête est un titre de film. Bien entendu, la dénomination d'un sujet la plus commune, choisie pour le titre d'un article, est aussi la plus commune dans un corpus de requêtes. Nous voyons donc que les requêtes présentent généralement des propriétés linguistiques qui correspondent bien à celles des titres d'articles de Wikipédia.

2.2 Propriétés sémantiques des catégories des articles de Wikipédia

(Voss, 2006) décrit la structure des catégories créées et hiérarchisées par les contributeurs de Wikipédia. Cette structure ressemble à une taxonomie, mais elle est plus flexible; ce qui lui permet de capturer la complexité de la sémantique de la langue naturelle. (Strube et al., 2006) appellent *folksonomie* la structure résultant de catégorisation des articles de Wikipédia. Par ailleurs (Zesh et al., 2007) montrent que le graphe constitué par les catégories de Wikipédia partage de nombreuses propriétés avec des réseaux sémantiques lexicaux tels que WordNet (FellBaum, 1998), classiquement utilisés pour des applications de TALN. Cela permet de penser que le graphe des catégories de Wikipédia est une ressource sémantique valide pour des applications de TALN, tout en étant beaucoup plus riche que des thésaurus coûteux à créer et à maintenir manuellement.

---

[1] http://en.wikipedia.org/wiki/Wikipedia:Naming_conventions accessed Feb. 2010

[2] http://fr.wikipedia.org/wiki/Principe_de_moindre_surprise

2.3 Génération de données issues de Wikipédia

A partir des tables des articles et des tables de traduction de la page de ressources de Wikipédia[3], nous avons extrait un dictionnaire bilingue anglais/français dans lequel les titres français d'articles de Wikipédia sont directement associés aux divers titres anglais correspondants: *"Avocat (fruit)"* ⇔*"Avocado"* et *"Avocat (métier)"*⇔*"Lawyer"*, par exemple. Cette table constitue un dictionnaire bilingue contenant 540920 traductions, dont un grand nombre d'entités nommées et de syntagmes, comme par exemple: *"Avocat du diable"* ⇔ *Devil's advocate"; "L'Avocat du diable (film)"* ⇔ *"Guilty as Sin"*.

La technique présentée ici pour résoudre les ambigüités s'appuie sur les catégories que les contributeurs associent à chaque article de Wikipédia. De plus, chaque catégorie est associée à d'autres catégories par des relations hiérarchiques. La catégorie de plus haut niveau est la catégorie *"Article"*. Ces relations hiérarchiques peuvent suivre un axe thématique ou un axe d'hyperonymie. Les hiérarchies de catégories associées aux articles de Wikipédia ont été extraites du site de ressources de Wikipédia. Ces tables fournissent la liste de tous les liens entre articles et catégories et des liens entre catégories. Les liens entre la catégorie terminale *"Article"* et ses sous-catégories[4] étant moins informatifs, nous n'avons conservé que les chemins des articles aux sous-catégories de *"Article"* : environs 150 catégories *"pseudo-terminales"*. Il faut rappeler ici que la hiérarchie de catégories n'est pas une taxonomie rigoureusement construite, que les catégories et leurs liens hiérarchiques sont ajoutés par des contributeurs variés. Ces contributions font toute la richesse spécifique de cette *folksonomie* (Strube, Ponzetto, 2007) mais nécessitent des approches adaptées pour en tirer le meilleur parti sans en subir les inévitables inexactitudes ou redondances. C'est pourquoi nous avons sélectionné le chemin le plus court parmi les chemins reliant l'article aux catégories pseudo-terminales. Ce processus d'extraction sémantique aboutit à associer chaque titre d'article à une ou plusieurs listes hiérarchisées d'environ 20 catégories. Le Tableau 1 fournit un exemple de listes de catégories associées au deux sens du mot *"avocat"*.

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

Tableau 1: Plus court chemin de l'article avocat vers une catégorie pseudo terminale, pour deux sens distincts..

3 Mise en œuvre du prototype de traduction de requêtes

3.1 Vue d'ensemble de la traduction des requêtes

La traduction des requêtes est effectuée en deux phases consécutives illustrées par la Figure 1: D'abord la segmentation en unités lexicales traduisibles par les titres de Wikipédia, ensuite la désambigüisation à l'aide des catégories, en langue cible.

[3] http://download.wikimedia.org/enwiki/latest/ downloaded Nov. 2009

[4] http://fr.wikipedia.org/wiki/Wikipédia:Catégories

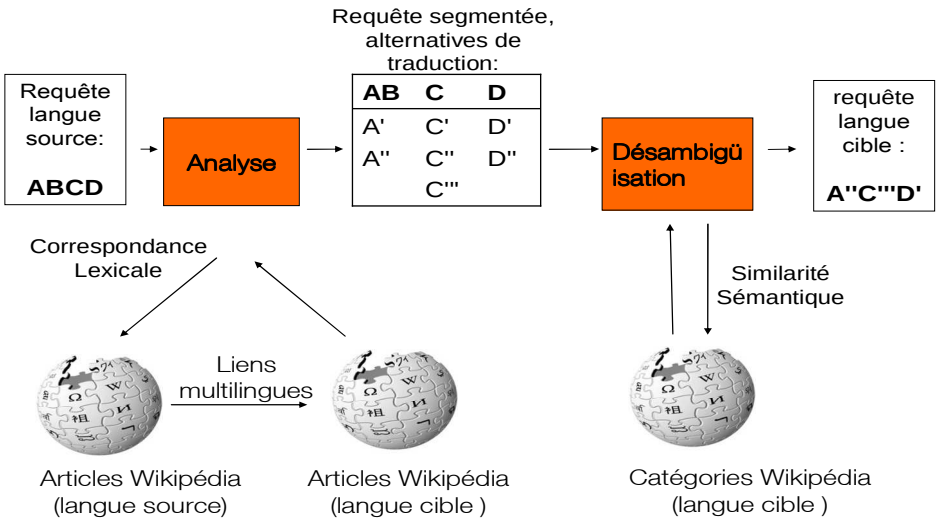

Figure 1: Schéma synoptique de la traduction de requêtes basée sur les titres et les articles de Wikipédia.

3.2 Segmentation des requêtes à l'aide des titres d'articles traduits

La traduction mot à mot d'une requête est souvent erronée car une requête est fréquemment constituée de plusieurs mots qui forment une unité lexicale. En particulier les titres de film se traduisent la plupart du temps de manière non littérale. Par exemple, *"Amicalement vôtre"* (FR) se traduit par *"The Persuaders"* (EN). De nombreux titres d'articles de Wikipédia sont constitués de plusieurs mots, et le titre de l'article correspondant dans une autre langue est la traduction non littérale de cette unité. Pour traduire une requête composée de plusieurs mots ou termes, il est donc nécessaire de la segmenter préalablement en unités lexicales. Par exemple, la requête ABCD (composée des Quatre mots A,B, C et D) peut se décomposer en: "ABCD"; "ABC,D"; "AB,CD"; "A,BCD"; "A,BC,D";"AB,C,D";"A,B,CD" ou "A,B,C,D". Le choix de la meilleure segmentation se base sur l'hypothèse que lorsque plusieurs mots successifs peuvent se traduire comme une unité, cette traduction est plus correcte que la traduction qui serait obtenue en traduisant des sous-unités. La méthode consiste à vérifier, pour chacune des segmentations candidates, si les unités qui la composent appartiennent au lexique bilingue extrait de Wikipédia et ont donc une ou plusieurs traductions possibles. Cette vérification s'effectue dans l'ordre défini par les trois règles R1 à R3, jusqu'à l'obtention d'une segmentation dont un pourcentage suffisant d'unités sont traduites. Cet ordre est défini par trois règles:

• (R1) Minimiser le nombre d'unités lexicales ("A,B,CD " plutôt que "A,B,C,D ").
• (R2) Pour le même nombre d'unités, maximiser la taille de la plus grande unité lexicale ("ABC,D " plutôt que " AB,CD ").
• (R3) Pour le même nombre d'unités, privilégier les unités lexicale en début de requête ("ABC,D " plutôt que "A,BCD ").

Une segmentation est acceptée si un pourcentage suffisant de mots (de la requête en langue source) se trouvent dans des unités lexicales qui donnent lieu à une traduction non vide. Par exemple, si le découpage [AB][C][DE] se traduit par [A'B'][][D'E'], alors ce pourcentage est de 80%. Les résultats présentés ici sont obtenus avec un pourcentage d'acceptabilité de 80%.

3.3 Désambiguïsation des requêtes traduites par homogénéité thématique

Chaque unité d'une requête peut être traduite par plusieurs titres d'articles de sens différents. Nous proposons de choisir la traduction qui maximise l'homogénéité thématique. Le champ sémantique de chacune des alternatives de traduction en langue cible est représenté par un vecteur d'une vingtaine de catégories associées par les utilisateurs aux titres d'article, comme expliqué dans la section 2.3. La proximité sémantique de deux alternatives de traduction est définie par la similarité cosinus de leurs vecteurs de catégories associés. La Figure 2 illustre ce calcul de proximité sémantique. Les proximités sémantiques de toutes les paires d'unités traduites sont calculées puis ajoutées. La somme obtenue est une mesure de l'homogénéité thématique de la requête traduite.

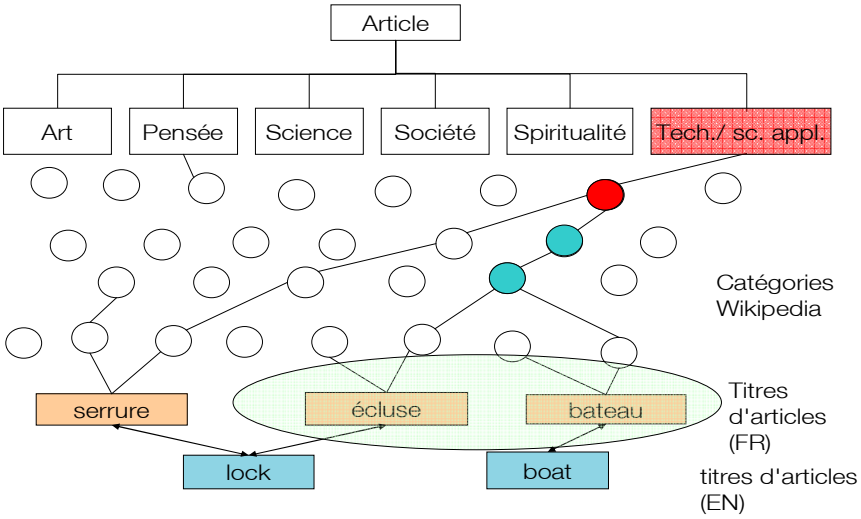

Figure 2:Choix de la traduction de lock par "écluse" car, "écluse" et "bateau" sont sémantiquement plus proches que "serrure" et "bateau".

4 Evaluation de la traduction des requêtes

4.1 Exemples de traductions comparées

Cette section propose quelques comparaisons de traductions d'exemples de requêtes qui illustrent des cas d'usage dans lesquels notre prototype donne des résultats satisfaisants.

Source	Traduction Wikipédia	Traduction Systran	Traduction Google
Michel blanc	Michel Blanc	White Michel	Michel Blanc
Maman, j'ai raté l'avion	Home Alone	Mom, I missed the plane	Mom, I missed the plane
gérard depardieu velo tout terrain	Gerard Depardieu Mountain Bike	Gerard depardieu bicycle any ground	Gérard Depardieu road bike

Tableau 2: Illustration de la segmentation et de la traduction des entités nommées et des syntagmes.

Source	Traduction Wikipédia	Traduction Systran	Traduction Google
juge avocat	Judge Lawyer	judge lawyer	Judge Advocate
avocat agriculture biologique	Avocado Organic Farming	lawyer organic farming	Advocate farming
lock boat	Ecluse Bateau	fermez à clef le bateau	lock bateau
lock door	Serrure Porte	porte de serrure	serrure

Tableau 3: Exemples de désambiguïsation de requêtes traduites.

4.2 Comparaison des taux d'erreurs de traduction

La traduction est évaluée sur un corpus de 750 requêtes "différentes" parmi 6900 requêtes saisies sur un moteur de recherche multimédia et monolingue. Nous avons comparé les traductions du prototype avec celles de trois logiciels de traduction automatique du marché, en libre service[5]: Systran, ProMT et Google.

Le taux d'erreur (ER) est évalué manuellement: chaque traduction est notée (0 pour une mauvaise traduction, 0.5 pour une traduction partiellement correcte et 1 pour une bonne traduction). La moyenne M (resp. M_w) de ces scores est calculée sur la base des 750 requêtes différentes les unes des autres (resp. des 7000 requêtes, même répétées). Le taux d'erreur est défini par la formule: ER=1-M, (resp. ER_w=1-M_w). Notre prototype ne propose aucun traitement grammatical ou de correction orthographique, donc nous avons distingué les requêtes comportant des erreurs orthographiques ou des structures grammaticales. Nous obtenons ainsi 6 taux d'erreur différents: pour toutes les requêtes (ER, resp ER_w), pour les requêtes sans erreur ni structure grammaticale (ER_{-og}, resp. ER_{w-og}) et pour les requêtes avec erreur orthographique ou structure grammaticale ($ER_{|og}$, resp. $ER_{w|og}$). Les résultats sont présentés dans le Tableau 4.

	Wikipédia	Systran	ProMT	Google
ER_w	**0,131**	**0,132**	**0,170**	**0,077**
ER	**0,331**	0,245	0,298	0,177
ER_{w-og}	**0,100**	**0,118**	**0,156**	**0,064**
ER_{-og}	**0,175**	0,155	0,225	0,111
$ER_{w\|og}$	0,713	0,373	0,410	0,286
$ER_{\|og}$	0,711	0,461	0,477	0,340

Tableau 4: Comparaison des Taux d'Erreurs de plusieurs traducteurs appliqués aux requêtes.

[5] http://www.systran.fr/; http://tr.voila.fr/; http://www.google.fr/language_tools?hl=fr

5 Interprétation des résultats, conclusions et perspectives

La traduction des requêtes basée sur les titres d'articles et les catégories de Wikipédia est généralement de bonne qualité, surtout pour les requêtes contenant des entités nommées. Pour capitaliser sur ces résultats, il serait intéressant d'intégrer le prototype présenté ici à un module plus général de traitement des requêtes utilisant des données et des techniques complémentaires. Comme signalé dans (Schönhofen et al., 2008), les mots simples ne font pas l'objet d'articles dans Wikipédia. L'usage de lexiques bilingues classiques serait donc complémentaire à celui des données issues de Wikipédia. En second lieu, les unités lexicales qui composent les requêtes ne sont traduites par notre méthode que si leur orthographe correspond exactement à celle d'un titre d'article de Wikipédia. Pour améliorer la robustesse du prototype il est nécessaire d'appliquer à la requête des techniques de correction automatique, de lemmatisation ou d'expansion (Ballesteros, Croft, 1997). D'autre part il est possible d'enrichir les données extraites de Wikipédia en prenant en compte les pages d'homonymies et les redirections (Bunescu, Pasca, 2006). Enfin, certaines structures de requêtes sont porteuses de sens, ainsi que certains éléments qui nécessitent des traitements spécifiques, comme les opérateurs booléens 'OR' et 'AND'. L'approche de segmentation et de désambiguïsation de requêtes basée uniquement sur Wikipédia permet d'obtenir des résultats assez satisfaisants bien qu'il ne soit pas optimisé par des données ou des techniques adaptées au traitement des requêtes. En particulier, il offre une solution au problème de mises à jour régulières de lexiques d'entités nommées et aux difficultés liées à la désambigüisation en contexte réduit (requêtes de quelques mots). Ainsi, nous pensons que cette approche, une fois intégrée à un mécanisme plus général de traitement des requêtes pour le CLIR, sera en mesure d'améliorer l'état de l'art de la Recherche d'Information Multilingue (Gaillard, 2009).

Références

BALLESTEROS L., CROFT W. B. (1997). Phrasal translation and Query Expansion Techniques for Cross Language Information Retrieval. Actes de *20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR 1997, 84-91.

BANERJEE S., PEDERSEN T. (2003). Extended gloss overlaps as a measure of semantic relatedness. Actes de *18th International Conference on Artificial Intelligence* IJCAI-03, Acapulco, Mexico.

BUNESCU R. C., PASCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. Actes de *11th conference of the European Chapter of the Association for Computational Linguistics* (EACL-2006), Trento, Italy, 9-16.

CLOUGH P. (2005). Caption and Query translation for Cross-Language Image Retrieval. *Lecture notes in Computer Science*, 3491, 614-625, Springer-Verlag.

ETZIONI O. REITER K., SODERLAND S., SAMMER M. (2007). Lexical translation with application to image search on the Web. Actes de *Machine Translation Summit XI*, Bente Maegaard (Eds.).

FELLBAUM C. (Ed) 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

GAILLARD B., BOUALEM M. (2009). Recherche d'information Multilingue: Etat de l'art, besoins, usages et évaluation. *Rapport Interne, Orange Labs*.

GAILLARD B., BOURAOUI, J. L., GUIMIER DE NEEF E., BOUALEM M. (2010). Expansion de requêtes pour la recherche d'information multilingue. CORIA 2010, *7è édition de la COnférence en Recherche d'Information et Applications*, Sousse, Tunisie. (*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 Computational Linguistics*, Manchester, UK 273–280.

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

JANSEN B. J., GOODRUM A., SPINK A. (2000). Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web Journal*, 3(4), 249-254.

JONES, G.J.F., FANTINO F., NEWMAN E., ZHANG Y. (2008). Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. *2nd International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, Hyderabad, India, 34-41.

LESK M. E. (1996). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from and ice cream cone. *5th Annual Conference on Systems Documentation*, Toronto, Ontario, Canada, 24-26.

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

NGUYEN D., OVERWIJK A., HAUFF C., TRIESCHNIGG D., HIEMSTRA D., DE JONG F. M. G. (2008). WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia, *Lecture notes in computer science*, 5706, (CLEF 2008), 58-65.

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.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *International Joint Conference for Artificial Intelligence* (IJCAI-95), 1, 448-453.

SCHÖNHOFEN P., BENCZUR A., BIRO I., AND CSALOGANY K. (2008). Cross-Language Retrieval with Wikipedia. *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval*,5152, (CLEF 2007) 72-79.

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

VOSS J. (2006). Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints, cs/0604036.*

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.