TALN 2010, Montréal, 19-23 juillet 2010

Traduction de requêtes basée sur Wikipédia

Benoît Gaillard, Olivier Collin, Malek Boualem

Orange Labs – 2, Avenue Pierre Marzin, 22300 Lannion, France,
benoit.gaillard@univ-tlse2.fr, 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.

1 Introduction de notre approche comparée à l'état de l'art

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

Les approches lexicales de traduction de requêtes pour la recherche d'information multilingue se heurtent à des difficultés de couverture lexicale et d'ambiguïté. L'encyclopédie en ligne Wikipédia permet de proposer des solutions à ces deux problèmes car d'une part elle met à disposition une quantité conséquente de connaissances, constamment mises à jour et accessibles, ce qui permet d'en extraire aisément des lexiques dont la couverture est optimale et d'autre part ces connaissances sont organisées sémantiquement par le biais de catégories fournies par les contributeurs (Zesh et al., 2007). Nous montrons que les données issues de Wikipédia ont des propriétés adaptées au traitement des requêtes et exposons une méthode de segmentation des requêtes en unités lexicales et une stratégie de choix de traduction parmi plusieurs alternatives, par homogénéité thématique s'appuyant sur les catégories. Beaucoup d'approches de traduction lexicale de requêtes s'appuient accessoirement sur Wikipédia, contrairement à notre approche qui en utilise exclusivement l'apport lexical et sémantique. Par exemple, Ballesteros et al. (1997) utilisent des syntagmes extraits d'un corpus parallèle traduit manuellement, alors que (Jones et al., 2008) utilisent un lexique de syntagmes issus de Wikipédia. Pour extraire les syntagmes de la requête, ces auteurs utilisent la méthode dite de *"Maximum forward matching"* qui consiste à détecter le plus long syntagme possible dans la requête, en partant du début, puis à répéter récursivement l'opération. Notre approche pour la détection de syntagmes, en revanche, s'appuie sur le seul lexique issu des titres d'articles de Wikipédia et maximise la taille des unités lexicales extraites sur la requête dans son ensemble.

1.2 Approches sémantiques pour la désambiguïsation des requêtes

Des mesures telles que la *"similarité sémantique"* initialement développées par (Resnik, 1995) peuvent être appliquées à des requêtes en s'appuyant sur Wikipédia (Strube, Ponzetto, 2006). (Bunescu, Pasca, 2006) proposent une méthode de reconnaissance et de désambiguïsation d'entités nommées (EN) à l'aide d'un dictionnaire extrait de Wikipédia et de la similarité cosinus entre les mots du contexte autour de l'EN et les mots de l'article correspondant à l'EN candidate. Nous fondons notre mesure de proximité uniquement sur les catégories, qui offrent une représentation plus concise du thème sémantique d'un article que celle issue du texte car c'est l'objectif même de leur ajout. (Schönhofen et al., 2008) s'appuient sur les articles, considérés comme des concepts, de Wikipédia en langue cible pour désambigüiser les requêtes par *homogénéité thématique* (*topic homogeneity*) (Gledson, Keane, 2008) pour ensuite les reformuler. Notre approche s'appuie aussi sur l'*homogénéité thématique*, mais nos alternatives sont des traductions directes, alors que les leurs sont des concepts qui, une fois sélectionnés, sont utilisés pour générer les requêtes en langue cible, comme à l'aide d'un pivot.

2 Wikipédia: une ressource pour le traitement des requêtes

Propriétés lexicales et sémantiques des titres et catégories d'articles: Les conventions relatives au nommage des articles sont définies sur la page explicative de Wikipédia[1]. En particulier, le titre idéal est le titre le plus concis, ne commençant pas par un article. Si plusieurs titres sont possibles, le plus commun est utilisé**,** par application du *principe de moindre surprise* (http://fr.wikipedia.org/wiki/Principe_de_moindre_surprise). Ces

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

conventions ont pour conséquence qu'une forte proportion de titres sont des EN et des groupes nominaux et ne comportent que quelques mots, comme la grande majorité des requêtes (Jones et al., 2008). Un utilisateur a tendance à formuler la requête la plus courte possible. 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. Les requêtes présentent donc des propriétés linguistiques qui correspondent à celles des titres d'articles de Wikipédia. (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. 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.

Génération de données issues de Wikipédia: Nous avons extrait de la page de ressources de Wikipédia (http://download.wikimedia.org/enwiki/latest/ downloaded Nov. 2009) un lexique bilingue dans lequel les titres français d'articles de Wikipédia sont associés aux titres anglais correspondants par des liens multilingues entre articles traitant du même sujet. Ce lexique contient 540.920 traductions, dont un grand nombre d'ENs et de syntagmes, comme par exemple: *"Avocat du diable" ⇔ Devil's advocate"; "L'Avocat du diable (film)" ⇔ "Guilty as Sin"*. Les contributeurs de Wikipédia associent des catégories à chaque article ainsi qu'à d'autres catégories par des relations hiérarchiques (thématique ou hyperonymique). Les hiérarchies de catégories associées aux articles de Wikipédia ont été extraites du site de ressources de Wikipédia. Elles ne constituent pas une taxonomie rigoureuse car elles sont librement construites par des contributeurs variés, ce qui fait toute la richesse de cette *folksonomie* (Strube, Ponzetto, 2007). C'est pourquoi, pour ne pas en subir les inévitables inexactitudes ou redondances, nous avons sélectionné environ 20 catégories parmi la ou les centaines de catégories associées à un article directement ou par filiation. Cette sélection détaillée dans (Collin et al. 2010) s'appuie sur l'heuristique selon laquelle le chemin le plus court parmi les chemins reliant les articles à des catégories pseudo-terminales serait le plus pertinent. Cette 20aine de catégories est utilisée de manière non pondérée.

3 Mise en œuvre du prototype de traduction de requêtes

Les deux phases consécutives de la traduction des requêtes sont illustrées Figure 1: D'abord la segmentation en unités lexicales, ensuite la désambigüisation à l'aide des catégories.

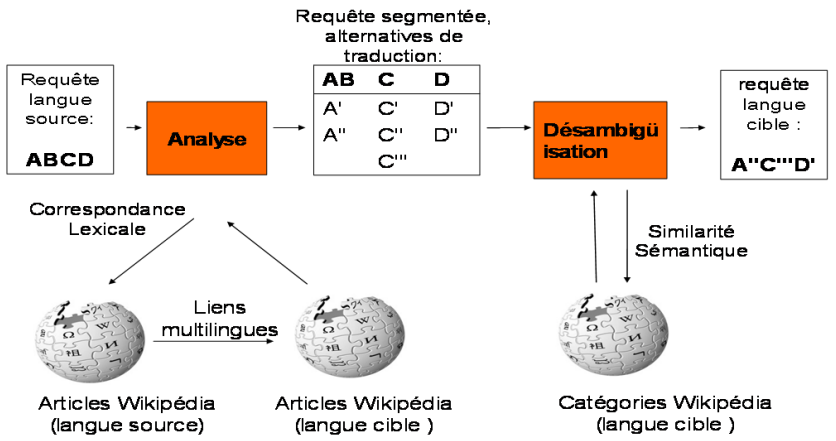

Figure 1: Schéma synoptique de la traduction de requêtes basée sur les titres et les articles de Wikipédia.

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

Une requête est fréquemment constituée de plusieurs mots qui forment une unité lexicale qui se traduit de manière non littérale. Par exemple, *"Amicalement vôtre"* se traduit en anglais par *"The Persuaders"*. Une requête composée de plusieurs mots doit être segmentée 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 la plus correcte. La méthode consiste à vérifier, pour chacune des segmentations candidates, si les unités qui la composent appartiennent au lexique bilingue extrait de Wikipédia, dans l'ordre défini par les trois règles R1 à R3:

• (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 la même taille d'unité, privilégier les unités lexicales en début de requête ("ABC,D" plutôt que "A,BCD").

Une segmentation est acceptée si un pourcentage suffisant (80%, dans ce travail) 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%. Si les règles R1 et R2 correspondent bien à l'intuition, le critère de la règle R3 et celui des 80% sont choisis empiriquement et leur optimalité mériteraient d'être évalué lors de travaux poursuivant le travail présenté ici.

3.2 Désambiguïsation des requêtes traduites par homogénéité thématique

Chaque unité d'une requête peut être traduite par plusieurs alternatives. Nous choisissons la traduction qui maximise l'homogénéité thématique. Chacune des alternatives en langue cible est représentée, dans l'espace vectoriel défini par les catégories, par un vecteur de coordonnée 1 selon chacune des catégories associées à l'article correspondant (environs 20, comme expliqué section 2), 0 selon chacune des autres catégories. La proximité sémantique de deux alternatives est définie par la similarité cosinus de leurs vecteurs de catégories, comme illustré par la Figure 2. Les proximités sémantiques de toutes les paires d'unités traduites sont calculées puis ajoutées, mesurant ainsi l'homogénéité thématique de la requête traduite.

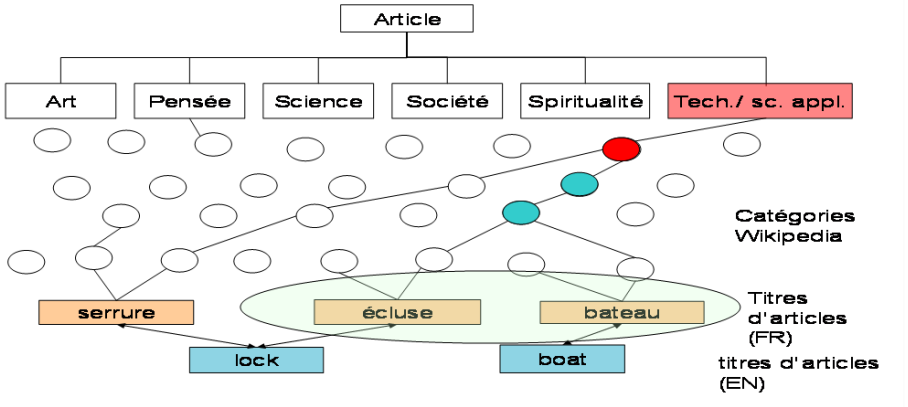

Figure 2:Choix de la traduction de « lock » par "écluse" par homogénéité thématique avec "bateau".

4 Evaluation de la traduction des requêtes et perspectives

Le tableau 1 illustre des résultats satisfaisants. Les lignes 1 et 2 illustrent la bonne couverture lexicale des ENs et termes en montrant que notre prototype est le seul à traduire correctement le titre de film ou le terme. Les lignes 2 et 4 illustrent les bonnes capacités de désambiguïsation par homogénéité thématique: le fruit *avocat* est plus probablement lié à l'*agriculture biologique* que l'homme de loi *avocat*, car ils partagent le thème de l'agriculture.

Source	**Traduction Wikipédia**	**Traduction Systran**	**Traduction Google**
Maman, j'ai raté l'avion	Home Alone	Mom, I missed the plane	Mom, I missed the plane
vélo tout terrain	mountain bike	bicycle any ground	road bike
juge avocat	Judge Lawyer	judge lawyer	Judge Advocate
avocat agriculture biologique	Avocado Organic Farming	lawyer organic farming	Advocate farming

Tableau 1: Illustration de la segmentation, de la traduction des EN et syntagmes et de la désambiguïsation.

La traduction a été évaluée sur un corpus d'environs 7000 requêtes saisies sur un moteur de recherche d'un portail multimédia monolingue publique d'Orange[2], se rapportant au domaine des actualités. Les requêtes, composées en général de quelques mots, comportent de nombreuses ENs et erreurs orthographiques, comme reporté dans (Bouraoui et al. 2010). Nous avons comparé les traductions du prototype avec celles de 3 logiciels de traduction automatique du marché, en libre service[3]: Systran, ProMT et Google. Le taux d'erreur (ER) est évalué par 1 humain qui assigne une note à chaque traduction (0: mauvaise traduction, 0.5: partiellement correcte, 1: bonne traduction). La moyenne M de ces scores est calculée sur la base des 7000 requêtes, même répétées car la performance d'un moteur se mesure par l'usage. Le taux d'erreur est défini par la formule: ER=1-M. Notre prototype ne proposant aucune correction orthographique ou traitement grammatical, nous avons distingué les requêtes comportant des erreurs ou des structures telles que des dates ou des phrases (environ le tiers de toutes les requêtes). Ainsi, le Tableau 2 présente 2 ER différents: pour toutes les requêtes (ER), et pour celles sans erreur ni structure (ER_{-og}).

	Wikipédia	Systran	ProMT	Google
ER	0,131	0,132	0,170	0,077
ER_{-og}	0,100	0,118	0,156	0,064

Tableau 2: Comparaison des Taux d'Erreurs de plusieurs traducteurs appliqués aux requêtes.

La traduction par lexique et catégories de Wikipédia *uniquement* est meilleure que celle de 2 des 3 traducteurs testés. Nos résultats inférieurs à ceux de Google s'expliquent par le fait que nous traitons mal de nombreuses requêtes comportant des opérateurs booléens, des mots

[2]http://www.2424actu.fr

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

simples tels que «de», ou des verbes conjugués. L'évaluation d'un moteur de CLIR tirant parti de cette approche dépasse le cadre de ce travail, mais la qualité relative des traductions obtenues malgré son implémentation minimaliste permet de penser que, lors de travaux futurs, notre approche sera en mesure de contribuer à l'élaboration de moteurs de recherche multilingues performants.

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.

BOURAOUI J.L., GAILLARD B., GUIMIER DE NEEF E., BOUALEM M. (2010). Annotation of linguistic phenomena in query logs. Actes de *Congreso Internacional de Lingüística de Corpus*, May 2010, University of A Coruña.

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.

COLLIN O., GAILLARD B., BOURAOUI J.L., GIRAULT, T. (2010). Constitution d'une resource sémantique issue du treillis des catégories de Wikipédia. Actes de *17è conference sur le Traitement Automatique des Langues Naturelles* (TALN 2010), Montréal, Canada.

GLEDSON A., KEANE J. (2008). Measuring Topic Homogeneity and its Application to Dictionary-Based Word-Sense Disambiguation. COLING 2008, Manchester, UK 273–280.

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.

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.

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.