Query Translation using Wikipedia-based resources for analysis and disambiguation

Benoît Gaillard
Orange Labs
&
CLLE-ERSS, 5, allées A. Machado, F-31058 Toulouse Cedex 9, France
benoit.gaillard@univ-tlse2.fr

Malek Boualem
Orange Labs
2, Av. Pierre Marzin 22300 Lannion, France
malek.boualem@orange-ftgroup.com

Olivier Collin
Orange Labs
2, Av. Pierre Marzin 22300 Lannion, France
olivier.collin@orange-ftgroup.com

Abstract

This work investigates query translation using only Wikipedia-based resources in a two step approach: analysis and disambiguation. After arguing that data mined from Wikipedia is particularly relevant to query translation, both from a lexical and a semantic perspective, we detail the implementation of the approach. In the analysis phase, lexical units are extracted from queries and associated to several possible translations using a Wikipedia-based bilingual dictionary. During the second phase, one translation is chosen amongst the many candidates, based on topic homogeneity, asserted with the help of semantic information carried by categories of Wikipedia articles. We report promising results regarding translation accuracy.

1 Introduction

Retrieving relevant information from the constantly increasing amounts of available multilingual content on the web is becoming as significant an issue as providing access to content originally was. To address this issue, Cross Language Information Retrieval (CLIR) techniques are used to enable users to retrieve relevant documents in a language different from the language of queries. To compare a query in a language to documents in another language, CLIR systems often apply Machine Translation (MT) techniques either to queries or to all indexed documents. Comparative evaluations (Clough, 2005) suggest that translating documents before indexing them is a slightly better approach to CLIR than translating queries because translations of indexed data tend to be more accurate than translations of queries. This approach has been occasionally studied, for example in (Gaillard et al, 2010), but major CLIR systems seem to favor query translation approaches because translating large indexes into many languages would be too costly. The most direct approach to translate queries is probably to use multilingual dictionaries, as for example in the prototype[1] detailed in (Etzioni et al, 2007). Two major difficulties face the lexical approach. First, the coverage of lexicons is a limiting factor that is difficult and expensive to optimize especially because queries can refer to a great number of named entities and multi-word terms the list of which is constantly and rapidly growing. Secondly, most words have multiple meanings. Selecting the most appropriate translation between several alternatives is a crucial yet challenging task, as queries often provide very little information that can be used to disambiguate.

Wikipedia has features that can provide solutions to these two issues. Thanks to the voluntary contributions of millions of users, it gathers a very significant amount of continuously updated, freely accessible organized knowledge. From it, one can easily extract up to date multilingual dictionaries that have an optimal lexical coverage. Furthermore, Wikipedia content is classified in a hierarchical network of semantic categories associated to articles by contributors, which can help choosing the most appropriate translation between alternatives i.e. disambiguating lexical translations.

This paper shows how organized information extracted from this online encyclopedia can be

[1] Turing Center, Univ. Washington, www.panimages.org/ .

used to solve the two classical issues of limited lexical coverage and of ambiguity, in order to accurately translate queries. We first show that Wikipedia features are well suited to query processing (Section 3), and detail the extraction of data from Wikipedia into a bilingual dictionary and usable semantic information. Section 4 explains how we use this data to analyze multi-term queries in order to extract from them the best combination of lexical units. In addition, we propose a strategy for choosing the most appropriate translation among several alternatives, with semantic techniques based on Wikipedia categories. Section 5 is devoted to the evaluation of the translation accuracy of this method, compared to other state of the art MT services.

2 Related work

2.1 Lexical approaches to Query Analysis and Translation

From Salton (1972) to Nguyen et al (2008) many methods for translating queries with the help of bilingual dictionaries and thesauri have been developed. As Ballesteros and Croft (1997) point out, three main issues need to be tackled when translating queries: dealing with Out Of Vocabulary (OOV) words or getting hold of exhaustive enough dictionaries, resolving ambiguities, and recognizing phrases, multi word locutions or named entities. As a solution to these issues, the authors propose to use local context around query terms in order to add expansion to them, before and after translation. The expansions, on one hand, clarify the meaning of queries and even, on the other hand help to minimize errors if irrelevant words have been added to the query by translation. The authors used phrases extracted from a manually translated parallel corpus, according to grammatical rules. However no explanation is given regarding the detection of multi-term phrases within queries.

This suggests that a promising approach to tackle the issue of lack of coverage of multilingual dictionaries, and of constituting translation lexicons of phrases or named entities, seem to rely on automatic extraction from parallel or comparable corpora. The recent significant increase in the number of users and contributors to Wikipedia makes it a good source for the construction of rich bilingual lexicons, as shown in (Zesh, 2007), because it provides easy access to large amounts of lexical and semantic information. For instance, Jones et al (2008) add to a regular MT solution a Wikipedia-based phrase dictionary. To detect phrases in queries the authors use a method called "*Maximum forward matching*" combined to simple grammatical rules (Ballesteros and Croft, 1997). This method consists in finding the longest possible phrase in the query, starting from the first word, and then recursively repeating the operation on the remaining part of the query. Detection is therefore performed by comparing character strings of increasing size to entries of the Wikipedia-based bilingual phrase dictionary.

2.2 Semantic approaches to disambiguation

Regarding disambiguation, since its participative category structure is semantically rich, Wikipedia is again a very precious resource. It has been used, for example, in an approach of (Mihalcea, 2007) where disambiguation is achieved by a statistical classification method trained on a Wikipedia-based corpus in which words are tagged with their meanings in context.

Measures like "*Semantic Similarity*" and "*Semantic Relatedness*" (Resnik, 1995) have been used for Wordnet-based applications by (Banerjee and Pedersen, 2003). However, as shown in (Strube and Ponzetto, 2006, 2007) they can also be applied to queries and be based on Wikipedia data. Banerjee and Pedersen (2003) show how to measure semantic relatedness between words, by extending the Lesk algorithm (Lesk, 1996): it consists in measuring the degree of overlap between words of the local context of the ambiguous term, and words of the definition of each sense of the term in the Wordnet thesaurus (Fellbaum, 1998). Strube and Ponzetto (2006) propose and compare various methods to measure the semantic relatedness of two words based on Wikipedia. The first measure is based on the path length between two concepts in the Wikipedia "*folksonomy*" (Guégan, 2006) that emerges from the categories. The second is based on probabilities of word occurrences and the last one adjusts to Wikipedia the (Banerjee and Pedersen, 2003) approach, measuring degrees of overlap between Wikipedia article contents.

Bunescu and Pasca (2006) also use Wikipedia-based semantic proximity to disambiguate the meaning of named entity recognized with a dictionary mined from Wikipedia. Taking into account redirecting and disambiguation pages, the disambiguation is performed using the cosine similarity measure between words of the context around the named entity (in the query) and words of the Wikipedia article for the candidate meaning. They enrich their approach with compari-

sons with vectors of categories associated to the considered articles.

Schönhofen et al (2008) use Wikipedia "concepts" (a subset of Wikipedia articles) in the target language in order to reformulate and disambiguate queries that have already been translated by lexical methods and in which concurrent alternative translations are kept. For each translation alternative of each query word, related Wikipedia concepts are selected. Target language queries are then generated from the most connected concepts thus selecting the most internally consistent alternative.

Wikipedia-based query translation does not need to rely on a lexical approach. For instance, Nguyen et al (2008) translate queries by projecting them onto a Wikipedia-based semantic space and then generating them in the target language, or, inspired by the Explicit Semantic Analysis (ESA) approach (Potthast et al, 2008), one can compare them to the projections of documents on the same conceptual space.

2.3 Originality of our work

Our approach combines several aspects of the techniques that we just summarized, in order to provide an original solution to Wikipedia-based query translation. For instance, like (Bunescu and Pasca, 2006), we use cosine similarity and Wikipedia categories to disambiguate translations. However, when they use the categories in conjunction with textual context in the articles and around the terms of the query, our disambiguation solely relies on the Wikipedia category structure. Moreover their approach is not applied to query translation. In (Schönhofen et al., 2008), we find an approach based on topic homogeneity: only the concepts that are the most similar to each other are used to generate the query in the target language. Our approach to choosing amongst alternatives is also based on topic homogeneity, but we measure it with cosine similarity based on Wikipedia categories whereas they use hyperlinks between articles. Furthermore, they reformulate the queries based on concepts, whereas our translation is more directly lexical.

Phrase detection approaches are mentioned by (Jones et al, 2008) or (Ballesteros and Croft, 1997) but they are not as detailed as the method we describe further on in the article. We describe an algorithm that not only looks for one phrase in the query, but that seeks to find the optimal combination of phrases, multi word locutions and named entities for the query as a whole. For example, let us consider a query consisting of five words represented here by A, B, C, D and E. Let us imagine that [AB], [CDE] and [ABC] are phrases or named entities. Following the maximum forward match approach mentioned in (Jones et al., 2008), the query will be analyzed as [ABC][D][E], whereas according to our algorithm the best analysis will be [AB] [CDE].

Most lexical query translation approaches that use Wikipedia use it as a complement to other bilingual dictionaries whereas our approach solely relies on Wikipedia. The work presented here does not try to propose as accurate a query translation as the state of the art, whilst using only Wikipedia resources. Its goal is to describe an approach that maximizes the benefits of Wikipedia lexical and semantic information for query translation.

3 Wikipedia as a resource to query processing

3.1 Lexical properties of Wikipedia titles

The query translation prototype described in this paper is based on the titles of Wikipedia articles. Naming conventions for Wikipedia articles are defined on the Wikipedia's naming conventions policy page[2]. This page states that titles should be recognizable, easy to find, precise, concise and consistent with other articles and uses. More explicitly, the convention states that "easy to find" means "*using names and terms that readers are most likely to look for in order to find the article*". These conventions imply lexical patterns that are similar to pattern found in logs of queries A significant proportion of titles are named entities, common nouns or noun phrases and very few of them are sentences or conjugated verbs.

Various analysis of themes and linguistic patterns of logs of queries (Jansen, 2000), (Bouraoui et al, 2010) have shown that the majority of queries consist in named entities and noun phrases and contain 1 to 4 words. What's more, users tend to formulate queries as concisely and precisely as possible. The "most common denomination" convention suggests that the title of an article should be, as much as possible, what comes the most naturally to the mind of someone thinking of the subject. We see from this comparison that queries and Wikipedia article titles present very similar lexical features.

[2] http://en.wikipedia.org/wiki/Wikipedia:Naming conventions accessed Feb. 2010

3.2 Semantic properties of the Wikipedia category graph

Voss (2006) describes the structure which arises from the categories associated by contributors to articles. Contributors can also propose hierarchical relations between categories. The categories and their hierarchy make a structure that is similar to taxonomy but is more flexible than a classification or ontology. Strube and Ponzetto (2006) call this structure a *folksonomy*. In addition, Zesh et al (2007) show that the Wikipedia category graph shares many properties with semantic nets such as WordNet (Fellbaum, 1998) that are often used for natural language processing (NLP) applications. They even show that methods traditionally used with semantic nets, such as semantic relatedness perform well on the Wikipedia Category Graph and that Wikipedia has a good coverage for nouns. This suggests that the category graph of Wikipedia is a valid resource for NLP applications, whilst being much richer than thesauri that are expensive to manually build and maintain.

3.3 Mining lexical and semantic multilingual information from Wikipedia

Data used in this paper was mined from Wikipedia through the Wikimedia downloads page[3], as part of a general approach to semantic resource extraction described in (Collin et al, 2010). We extracted a bilingual (English/French) dictionary from the translation table[4] and the table of French articles[5]. Direct relations between French article titles and English article titles were stored in the form of a table that directly associates titles with their various translations: *"Avocat (fruit)"* ⇔ *"Avocado"* or *"Avocat (métier)"* ⇔ *"Lawyer"*, for example. This translation table is comparable to a bilingual dictionary having 540.920 links. Its specificity is that it contains an important quantity of named entities and phrases, such as for instance: *"Avocat du diable"* ⇔ *Devil's advocate"; "L'Avocat du diable (film)"* ⇔ *"Guilty as Sin"*, that can be directly used for lexical translation.

The technique we used to resolve ambiguities consists in refining the semantic and thematic scope of articles with the help of their associated categories. There are not always many of them (especially in French), and they often are not informative enough to perform a satisfactory disambiguation. Moreover, linguistic processing for disambiguation is often based on hyperonymy or themes. For instance, one can use the knowledge that the fruit named (in French) "*avocat*" is part of the agriculture theme, whereas the court-based profession of "*avocat*" is of the law theme. Therefore, we have extended the representation of article's semantics with parent categories. In the Wikipedia category graph, every category has a parent category that generalizes it, following a thematic or hyperonymic direction. The highest category (parent of all other categories) is the category "*article*".

The necessary data to characterize the semantic scope of Wikipedia articles with the help of their associated categories was extracted from Wikimedia download sql files[6]. With these tables we listed all the paths between articles and the terminal category "*article*". Since the quantity of such paths is too large (hundreds of paths for some articles), we made a relevant selection among all these paths, based on the assumption that the most relevant information is carried by the shortest path that links each of the article's categories to the terminal category. In fact, after some testing we realized that paths linking to the *"Article"* category were less relevant than paths to the set of categories one level or two below the *"Article"* category, pointed to by the category page[7]. This set contains 150 *pseudo terminal* categories. For each article, we selected one path per associated category: the shortest path to one of the "pseudo terminal" categories. If several paths were of equal length, they were all selected. Table 1 illustrates the results for the word "*avocat*".

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

Table 1. Shortest paths to a pseudo terminal category, for two distinct meanings of the word *"avocat"*.

Each Wikipedia article was associated with the few selected paths, and this category-based semantic representation was used to disambiguate query translations, as we will detail in the subsequent sections.

[3] http://download.wikimedia.org/enwiki/latest/ downloaded Nov. 2009
[4] frwiki-latest-page.sql; frwiki-latest-langlinks.sql
[5] frwiki-latest-langlinks.sql
[6] frwiki-categorylinks.sql (fr) and enwiki-categorylinks.sql (en)
[7] http://fr.wikipedia.org/wiki/Wikipédia:Catégories

4 Functional aspects of the prototype

4.1 Two successive steps for translation

Queries are translated in two successive phases, as illustrated in Figure 1.

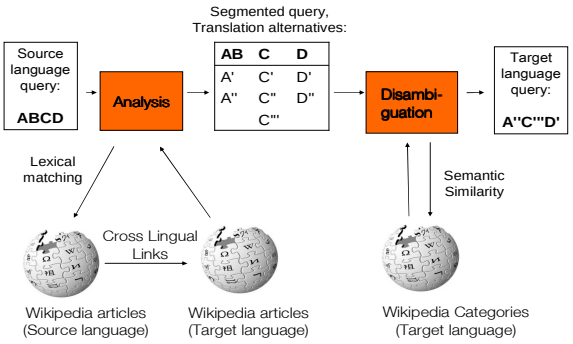

Figure 1. Wikipedia-based query translation

First of all, there is an analysis phase during which queries are segmented in lexical units that the Wikipedia-mined bilingual dictionary can translate. This phase associates one or several candidate translations to each lexical unit of the query, based on the multilingual Wikipedia links. We provide more details on this phase in section 4.2. The second phase is the disambiguation. Since there often are several alternatives for each lexical unit, many combinations can be candidates to the final translation. In the case when queries are segmented into several lexical units, we choose the best combination, according to topic homogeneity with a specific method based on Wikipedia categories (Section 4.3).

4.2 Segmentation of the query

Word for word translation of queries is often inaccurate because queries tend to include phrases, named entities or multi-word terms. For instance the title of the series "*The persuaders*" would never be literally translated to "*amicalement vôtre*", its French title that literally means "*friendly yours*". Many Wikipedia titles are made of several words and their equivalent titles in a different language are non literal translations of that lexical unit. In order to translate a query that has several words, it is therefore necessary to segment it into lexical units.

In order to detail our segmentation algorithm, let us consider the example of a query composed of the 4 words A, B, C and D. Provided that only consecutive words can form a lexical unit, this query can be segmented in 8 different ways: "ABCD"; "ABC,D"; "AB,CD"; "A,BCD"; "A,BC,D"; "AB,C,D"; "A,B,CD" or "A,B,C,D". The choice of the best segmentation is based on the assumption that, if a succession of words can be translated as a whole, translating subunits of it would harm the accuracy of the translation. Our method consists in verifying, for each candidate segmentation, that its lexical units belong to the Wikipedia-mined bilingual dictionary (section 4.2) and therefore have one or several possible translations. This verification is made in decreasing *order of units' sizes*, until an acceptable segmentation is found. More precisely, the order of units' sizes is defined by three rules:

- The fewer lexical units in a segmentation, the better: the "A,B,CD" segmentation is preferred to the "A,B,C,D" one (R1).
- For the same number of units, the segmentation with the longest unit is favored: "ABC,D" is preferred to "AB,CD" (R2).
- For the same number of units and maximum size, the segmentation whose longest unit is the earliest is preferred: "ABC,D" is preferred to "A,BCD" (R3).

Acceptability of candidate segmentations: The chosen segmentation is the first, in the order defined by the three rules R1 to R3, for which "*most of*" the units are translated. "*Most of*" is defined by a percentage of words of the source language query that belong to units that have translations. If the segmentation [AB][C][DE] is translated by [A'][][B'], where the single word unit [C] has no translation, the percentage of translated word is 80%. However, if a query is segmented as [ABC][DE] and is translated by [A'][], where [DE] has no translation, then its percentage is 60%. All the results presented here are based on an 80% threshold of acceptability.

4.3 Maximizing topic homogeneity of the translation

Each unit can be translated in several ways. Since choosing the most likely translation in the case of single unit queries is not in the scope of this article, when (during evaluation) we met a query segmented into only one lexical unit, we randomly chose one of the alternatives. We focus here on the case in which the query has several units that can be translated independently, each of them by one or several alternatives. In that case, our approach consists in choosing the combination for which the units are the most semantically close to each other, the combination that maximizes the " *topic homogeneity*" (Gledson and Keane, 2008) of the query. For example, let us compare the query Q1 "*avocat juge*" (*"lawyer/avocado judge"*) and the query Q2 "*avocat*

agriculture biologique" (*"lawyer/avocado organic farming"*). In Q1, the meaning "*lawyer*" is semantically close to *"judge"*, they both belong to the semantic field of the law. In the query Q2, conversely, the meaning "*avocado*" is semantically closer to "*organic farming*". Therefore Q1 should be translated by "*lawyer judge*" whereas Q2 should be translated by "*avocado organic farming*". To describe the semantic field of each translation alternative of a unit, we use the category paths described in section 3.3. Each translation alternative is thus associated to approximately 20 categories. We then represent a candidate unit translation with a vector whose dimensions are the Wikipedia categories. The semantic proximity of two lexical units is then measured by the cosine similarity of their category vectors. The Figure 2 illustrates this semantic proximity measure.

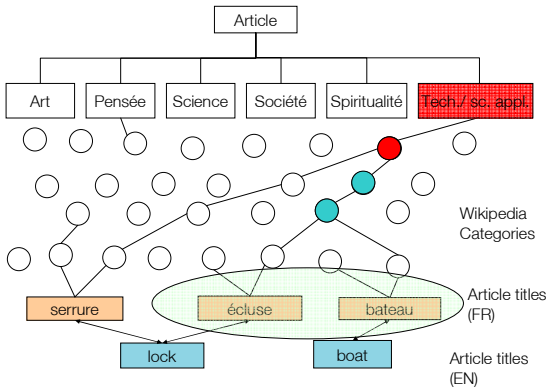

Figure 2. Translation of the query "*lock boat*" (English to French). *"écluse"* and *"bateau"* are semantically closer than *"serrure"* and *"bateau"*.

In the general case, for any number of units, we choose the combination of alternatives for which the sum of the cosine similarities is the greatest. This sum can be considered as a measure of the topic homogeneity of the generated query.

5 Experimental validation

5.1 Query specific optimal lexical coverage

We display in Table 2, use cases in which our approach improves query translation thanks to an optimal lexical coverage of query specific domains such as named entities or terms.

Source	**Wikipedia prototype**	**Systran**	**Google**
Maman, j'ai raté l'avion	Home Alone	Mom, I missed the plane	Mom, I missed the plane
Michel blanc	Michel Blanc	White Michel	Michel Blanc
Amicalement votre	The persuaders	in a friendly way your	friendly your
gérard depardieu velo tout terrain	Gerard Depardieu Mountain Bike	Gerard depardieu bicycle any ground	Gérard Depardieu road bike
Recherche d'information	Information Retrieval	Search for information	Information Retrieval
Prise de la bastille	Storming of the Bastille	Storming of the Bastille	Bastille

Table 2. Query analysis and translation of phrases and named entities.

5.2 Improved query analysis and disambiguation

We display in Table 3 use cases in which our method of query segmentation and disambiguation improves query translation.

Source	**Wikipedia prototype**	**Systran**	**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
house grey's anatomy	Dr House Grey's Anatomy	l'anatomie du gris de maison	grey's anatomy House

Table 3. Segmentation and disambiguation, from french to english and from english to french.

5.3 Comparison with baseline MT systems

The translation accuracy of our approach was measured on a corpus of 750 queries issued from the log of a monolingual, publicly available on the Orange portal, multimedia search engine[8] over three days. Many of these 750 queries were typed in on several occasions, the most frequent query was typed in 2021 times. So the total number of queries in the corpus is about 7000, ranging from 1 to a dozen words.

We compared the translations of these queries by our approach with the translations by three well known MT services freely available online: The online Systran solution[9], the ProMT online application[10], and the Google CLIR service[11]. We manually evaluated the Error Rate (ER) of each translator on the corpus, using the following method: each translation was given by the anno-

[8] http://www.2424actu.fr
[9] http://www.systran.fr/
[10] http://tr.voila.fr/
[11] http://www.google.fr/language_tools?hl=fr

tator an accuracy score (0 for a wrongly translated query or not translated at all, 0.5 for a partially correct translation and 1 for a good translation). Translations were randomized so that the annotator could not ascribe a translation to a translator. The mean score M was computed over all these scores and the ER was defined by the formula: ER=1-M.

M can be computed based on the 750 queries or based on each occurrence of each query (over the 7000 occurrences). We call the latter a weighted mean and the resulting ER is called the weighted ER (ER_w). In the introduction, we pointed out that our prototype has no spelling mistake or grammatical processing module. Therefore, in order to compare its score with the three other translators, we also measured the ER over the subset of queries that have no spelling mistake and no grammatical feature, selected by hand. Since query are usually very short, this set is still of significant size. Each MT service or prototype was therefore given 6 different scores: ER over all the queries, ER over all the queries that have no spelling mistake or grammatical feature (ER_{-sg}) and ER over the queries that do have spelling mistakes or grammatical features ($ER_{|sg}$), these three rates weighted (ER_w) or "flat". The results are presented in Table 4:

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

Table 4. Comparison of ER of various MT solutions

On the subset of queries that have no spelling mistake or grammatical feature, our ER is equal or slightly lower than the ER of other MT solutions, except Google. Since our translation is solely based on Wikipedia, results comparable to state of the art MT are impressive as they suggest that the Wikipedia lexical coverage of queries is almost optimal. Furthermore, we perform better with weighted queries, meaning that Wikipedia's lexical coverage also fits lexical statistics of queries.

6 Comments and further research

The evaluation shows that translating queries based on titles of articles and on categories of Wikipedia is accurate in comparison to other established MT solutions, especially for a large proportion of queries consisting in phrases and named entities. Nevertheless, in order to have a complete prototype of query processing for CLIR it would be necessary to combine our Wikipedia techniques with various other techniques and data. First of all Wikipedia is a rich resource but only in specific areas of language, as pointed out in (Schönhofen et al, 2008). Therefore common words bilingual dictionaries need to be combined with our Wikipedia-mined resources. In order to improve the robustness of the prototype, lemmatization techniques, spelling and grammar processing should be applied to queries. Furthermore it is possible to enrich the Wikipedia-mined data by taking into account redirection and disambiguation pages and links, as in (Bunescu et al, 2006). Finally, we noticed that some queries have a structure that carries meaning in itself, and elements that require specific processing outside of translation.

7 Conclusion

This paper describes a query translation approach for CLIR, based on multilingual lexical and semantic information mined Wikipedia. The proposed approach combines a generalization of query segmentation techniques such as "*maximum forward matching*" with a disambiguation technique based on *topic homogeneity*, which is measured on the basis of the similarity of categories associated to the various alternatives of each lexical unit of the query. The approach has been experimentally validated because it shows satisfying translation accuracy in comparison to established MT solutions, without using any state of the art linguistic MT processing which would obviously greatly improve it. The implementation of a MT prototype integrating this processing with our approach is the natural next step of this work and is likely to show much better translation accuracy than baseline MT systems. The approach is therefore a promising first step towards a solution to the issue of building and updating multilingual dictionaries for phrases and named entities, and to the issue of disambiguation of lexical translation of short queries.

References

Ballesteros, Lisa and W. Bruce Croft, 1997. Phrasal translation and Query Expansion Techniques for Cross Language Information Retrieval. *20th annual international ACM SIGIR conference on Research and development in information retrieval*, 84-91.

Banerjee, Satanjeev and Ted Pedersen, 2003. Extended gloss overlaps as a measure of semantic relatedness. *18th International Conference on Artificial Intelligence*, Acapulco, Mexico.

Bunescu, Razvan C. and Marius Pasca, 2006. Using encyclopedic knowledge for named entity disambiguation. *11th conference of the European Chapter of the Association for Computational Linguistics* , Trento, Italy, 9-16.

Bouraoui, Jean Léon, Benoît Gaillard, Emilie Guimier De Neef, Malek Boualem, 2010. Annotation of linguistic phenomena in query logs. *Congreso Internacional de Lingüística de Corpus, May, University of A Coruña, (To appear).*

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

Collin, Olivier, Benoît Gaillard, Jean Léon Bouraoui, Thomas Girault, 2010. Semantic resource extraction from the Wikipedia category lattice. LREC 2010. *Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods? (*To appear)

Etzioni, Oren, Kobi Reiter, Stephen Soderland and Marcus Sammer, 2007. Lexical translation with application to image search on the Web. *Proceedings of the Machine Translation Summit XI*, Bente Maegaard (Eds.).

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

Gaillard, Benoît, Jean Léon Bouraoui, Emilie Guimier De Neef, Malek Boualem, 2010. Query Expansion for Cross Language Information Retrieval Improvement. *Fourth international conference on Research Challenges in Information Science (RCIS 2010)*, (To appear).

Gledson, Anne, and John Keane, 2008. Measuring Topic Homogeneity and its Application to Dictionary-Based Word-Sense Disambiguation. *22nd International Conference on Computational Linguistics*, Manchester, UK 273–280.

Guégan, Marie, 2006. Catégorisation par les contributeurs des articles de l'encyclopédie Wikipédia.fr. *Mémoire de master de recherche informatique université Paris XI*, LIMSI CNRS.

Jansen, Bernard J., Abby Goodrum and Amanda Spink, 2000. Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web Journal*, 3(4), 249-254.

Jones, Gareth J.F., Fabio Fantino, Eamonn Newman and Ying Zhang, 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, Michael 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, Rada, 2007. Using Wikipedia for Automatic word Sense Disambiguation. *North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, Rochester.

Nguyen, Dong, Arnold Overwijk, Claudia Hauff, Dolf Trieschnigg, Djoerd Hiemstra and Fransiska M. G. De Jong, 2008. WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia, *Lecture notes in computer science*, 5706, (CLEF 2008), 58-65..

Ponzetto, Simone P. and Michael Strube, 2007. Deriving large scale taxonomy from Wikipedia. *22nd national conference on Artificial intelligence*, AAAI Press, Vancouver, British Columbia, Canada, 1440-1445.

Potthast, Martin, Benno Stein and Maik Anderka, 2008. A Wikipedia-Based Multilingual Retrieval Model. *30th European Conference on IR Research*, Craig Mcdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven and Ryen W White (Eds), 522-530.

Resnik, Philippe, 1995. Using information content to evaluate semantic similarity in a taxonomy. *International Joint Conference for Artificial Intelligence*, 1, 448-453.

Salton, Gerard, 1972. Experiments in multi-lingual information retrieval. *Technical report TR 72-154*, Computer Science Department, Cornell University.

Schönhofen, Peter, Andras Benczur, Istvan Biro and Karoly Csalogany, 2008. Cross-Language Retrieval with Wikipedia. *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval*,5152, (CLEF 2007) 72-79.

Strube, Michael and Simone P. Ponzetto, 2006. WikiRelate!: Computing Semantic Relatedness Using Wikipedia. *21st national conference on artificial intelligence*, 1419-1424.

Voss, Jakob, 2006. Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints, cs/0604036*.

Zesch, Torsten, Iryna Gurevych and Max Mühlhäuser, 2007. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. *Biannual Conference of the Society for Computational Linguistics and Language Technology*, 213-221.