WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia

D. Nguyen, A.Overwijk, C.Hauff, R.B. Trieschnigg, D. Hiemstra, F.M.G. de Jong
Twente University
dong.p.ng@gmail.com, arnold.overwijk@gmail.com, c.hauff@ewi.utwente.nl, trieschn@ewi.utwente.nl, hiemstra@cs.utwente.nl, f.m.g.dejong@ewi.utwente.nl

Abstract

This paper presents WikiTranslate, a system which performs query translation for cross-lingual information retrieval (CLIR) using only Wikipedia to obtain translations. Queries are mapped to Wikipedia concepts and the corresponding translations of these concepts in the target language are used to create the final query. WikiTranslate is evaluated by searching with topics in Dutch, French and Spanish in an English data collection. The systems achieved a performance of 67% compared to the monolingual baseline.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

Keywords

Cross-lingual information retrieval, query translation, word sense disambiguation, Wikipedia, comparable corpus

1. Introduction

Cross-lingual information retrieval (CLIR) has become more important in recent years. CLIR enables users to retrieve documents in a different language than their original query. Potential users for CLIR are users who find it difficult to formulate a query in their non-native language and users who are actually multilingual and want to save time by entering a query in one language instead of entering the query in all languages they want to search in.

There are two approaches to handle CLIR: Translating the query into the target language or translating the documents into the source language and searching with the original query. Most of the research in this area uses the first approach, because translating the documents itself is not scalable. The method described in this paper also uses the first approach. Main approaches to query translation are dictionary based translation, the use of parallel corpora and machine translation (MT). A new source to obtain translations is Wikipedia.

This paper introduces WikiTranslate; a system that performs query translation using only Wikipedia as a translation resource. Most Wikipedia articles contain *cross-lingual links*; links to articles about the same concept in a different language. These cross-lingual links can be used to obtain translations. The aim of this research is to explore the possibilities of Wikipedia for query translation in CLIR.

We will treat Wikipedia articles as representations of concepts (i.e. units of knowledge). WikiTranslate will map the query to Wikipedia concepts. Through the cross-lingual links translations of the concepts into language-specific terms are retrieved. Another method would be to generate a parallel corpus with Wikipedia (e.g. [1]) and then to extract translations. However, the first method is chosen since it allows us to make use of the structure of Wikipedia, for example the cross-lingual links, text and internal links (e.g. links in the text that link to other Wikipedia articles), and thereby enabling us to investigate new possibilities to perform query translation.

The main research question of this paper is: *Is Wikipedia a viable alternative to current translation resources in cross-lingual information retrieval?*

The goal is to transform q_s (the query in source language *s*) to q_t (the query in target language *t)*. First we will map query q_s to Wikipedia concepts using the Wikipedia in language *s*. Translations of the mapped concepts to language *t* can be obtained through the available cross-lingual links. With these translations query q_t can be created. This raises the following sub questions: *How can queries be mapped to Wikipedia concepts?* and *How to create a query given the Wikipedia concepts?* Note that mapping a query to concepts is a difficult, but crucial step. One of the difficulties in this step is also a major problem in CLIR itself: word sense disambiguation. A hypothesis is that the structure of Wikipedia (e.g. articles, internal links etc.) makes Wikipedia a useful source for CLIR. The expectation is that

Wikipedia is suitable to handle the challenges mentioned above and will demonstrate that is it a promising approach in the field of CLIR.
The system is first evaluated with the data collections of the CLEF (Cross Language Evaluation Forum) Ad Hoc task of 2004, 2005 and 2006 using different approaches. The best approach is submitted to the ad hoc task of 2008.
First an overview of Wikipedia and related work in the field of CLIR is given. Then WikiTranslate is introduced and the experimental setup is described. Results are then presented and discussed.

2. Related work

2.1. Cross-lingual Information Retrieval

Because queries are usually short, therefore not providing enough context, and composed of keywords instead of sentences, CLIR has to deal with a lot of challenges like out of vocabulary words (OOV), named entity recognition and translation, and word sense disambiguation. The latter one is especially problematic with very short queries and needs to have a very high degree of accuracy to be useful to an IR-system [2] .
Two important observations are made by [3]. First, most IR models are based on bag-of-words models. Therefore, they don't take the syntactic structure and word order of queries into account. Second, queries submitted to IR systems are usually short and therefore not able to describe the user's information need in an unambiguous and precise way. When taking these observations in account, this means that the final query delivered to the system, translated from the source language, does not have to be a single translation. Since MT often only delivers one translation, this might not be the most suitable way to generate new queries. Including synonyms and related words can improve performance [3].
One approach to accomplish this is with query expansion. This can be result-independent or dependent. With result independent expansion word relationships are used. One of the methods for result-independent expansion is the use of semantic relations, but Voorhees [4] shows that this often degrades performance. Result dependent expansion uses documents retrieved by the initial query. In CLIR query expansion can occur before translation, after, or both. Research shows that combining pre- and post-translation has the best performance [5, 6].
Another approach to get multiple translations for a given concept is the use of parallel corpora. Examples of this are [7,8]. The first step of Sheridan et al. [8] was retrieving the best matching documents in the source language. Then words were selected that occurred frequently from the corresponding retrieved documents. After that, a final query was composed with these words. Lavrenko et al. [7] follows the same approach except that it doesn't select new query terms but makes a relevance model. The approach of [7] also automatically achieves word sense disambiguation since the method makes use of co-occurrence statistics.

2.2. Wikipedia

Wikipedia is an online, multilingual encyclopedia where every user can make a contribution to. Soon after its start it began to grow exponentially [9]. Because Wikipedia is driven by its users, its topical coverage depends on their interests, but even its least covered areas are covered well according to [10]. The characteristics of Wikipedia make it suitable as a semantic lexical resource [11].
Wikipedia has been used for automatic word sense disambiguation [12] and for translation. Su et al. [13] used it to translate out of vocabulary words and Schönhofen et al. [14] used it to translate queries. Both used the cross-lingual links available in Wikipedia to translate terms. One of the steps Schönhofen et al. [14] performed was determining the concepts of a query with Wikipedia and constructing a new query.
Wikipedia has been used to find similar sentences across languages to generate parallel corpora [1].The notion that it can be treated as a comparable corpus is new and has not been researched much yet except by [15]. Wikipedia can be seen as a comparable corpus since articles are represented in different languages and connected through cross-lingual links.

3. Motivation for the usage of Wikipedia

Wikipedia has the following advantages compared to the existing resources used to perform query translation (e.g. bilingual dictionaries, parallel corpora etc.):

- Better coverage of named entities and domain specific terms [11].
  It might therefore be suitable to handle one of the problems of CLIR: the translation of proper names.
- The information is very up to date because of the continuous contribution of users.
- Wikipedia articles provide more context in comparison with sources like online dictionaries.
  This can be used to perform word sense disambiguation [2], also a well known problem of CLIR.
- Presence of redirect pages; pages that represent alternative names of concepts and that only consist of a link that directs to the main article the concepts represents (for example *coalition cabinet* redirects to *coalition government*). These redirect pages represent synonyms (words with identical or similar meanings), abbreviations and spelling variants [11]. Therefore redirects may be used for query expansion.

The use of Wikipedia also has some disadvantages. The coverage is less than dictionaries on common words (e.g. *drive*, *stay*, etc.) and some terms have a lot of senses, some of which are very specific and uncommon, thereby making word sense disambiguation more difficult.

4. Proposed approach

4.1. Outline

The approach used by WikiTranslate consists of two important steps: mapping the query to Wikipedia concepts and creating the final query using these found concepts. Some sub steps are optional; these steps have been left out in some tests. Below is an overview of the proposed approach, an elaboration on these steps can be found in section 5.

- **Step 1: Mapping the Query to Wikipedia concepts**
  First the most relevant concepts to the query are extracted after a search with the whole query (step 1a). Then a search on every term of the query is performed (step 1b). This is done in two different ways. Using the internal links from the concepts retrieved with step 1a (which we will call LINKS) and using the text and title of the Wikipedia articles (which we will call CONTENTS).

- **Step 2: Creating the Translated Query**
  First we add articles that redirect to the found Wikipedia concepts (step 2a, optional) to include synonyms and spelling variants. Furthermore articles retrieved with step 1a are given more weight (step 2b, optional)
  Finally, the final query is created using the found concepts (step 2c). There are different ways to combine the concepts and some translations have to be modified.

This approach differs from other approaches in different ways. The most important difference is that the text and internal links of the articles are used. These are not available for approaches based on dictionaries or a parallel corpus. Other research using Wikipedia maps concepts to queries by searching for titles that exactly match with the words in the query ([13], [14]). Furthermore, they only use Wikipedia to enhance their translations (e.g. as a next step after using a bilingual dictionary).
The features of Wikipedia allow us to use different methods to map a term to a Wikipedia article. Only mapping on the titles of the articles (similar to mapping on dictionary entries), using the text of articles (the whole text or only the first paragraph), or both.
An advantage of this approach is that it allows extraction of phrases from the topics, since the titles of Wikipedia articles are often phrases. For example when the concept named "*Ayrton Senna*" is retrieved, quotation marks can be put around the title, causing documents that contain this phrase exactly will score higher. Furthermore by adding the top documents from step 1a, the most relevant concepts to the whole query are added. Also related concepts to the whole query can be added with this step, creating a kind of query expansion effect.

4.2. Example

To illustrate the steps of the proposed approach topic C230 from the Ad hoc task of CLEF 2004 is used. We will take a look at the translation of Dutch to English.
The Dutch topic is formulated like this:

```
<title> Atlantis-Mir Koppeling </title>
<desc> Vind documenten over de eerste space shuttle aankoppeling tussen de Amerikaanse shuttle
Atlantis en het Mir ruimte station. </desc>
```

The corresponding English topic looks like this:

```
<title> Atlantis-Mir Docking </title>
<desc> Find documents reporting the first space shuttle docking between the US shuttle Atlantis
and the Mir space station. </desc>
```

5. Experimental setup

5.1. Lucene

Lucene will be used as the underlying retrieval system to retrieve Wikipedia articles at steps 1a, 1b and 2a. Lucene gives every document a relevance score to the query, based on the Vector Space Model [16] and the Boolean model [17].

5.2. Preprocessing

5.2.1. Indexing Wikipedia

From each article the title, text and cross-lingual links are extracted. The first paragraph of an article is extracted as well, which will be called "description". This is done because the mapping of queries to Wikipedia articles uses the text of an article. However, since some articles contain a lot of text (e.g.: article about Europe), they tend to score lower than short articles while they are actually more related to the searched term. To prevent this, instead of searching on the whole text, the search scope can be limited to the first paragraph, since the first paragraph usually contains a summary of the article or most important information about the article (and thus containing the most related words). If the article is in fact a redirect page, the title of the referred page is also stored. Wikipedia articles that represent images (e.g. have the title "image:xxx"), help pages, templates, portal pages and pages about the use of Wikipedia are not included.
To enhance comparability, the same preprocessing method is used for all languages. We choose stemming, although there is no uniform best way of preprocessing for all languages [18]. Stemming is best for Dutch and Spanish, but 4-gramming is more suitable for English and French [18]. We use Snowball stemmers to perform stemming [19]. Words are removed with the lists from the Snowball algorithm [19].

5.2.2. Compound Words

Compound words are words that are composed of multiple words. For example the Dutch word *zeilboot (*English: *sailing boat),* which is a combination of *zeil* (English: *sail*) and *boot* (English: *boat*). Of the source languages used with evaluation, only Dutch contains compounds words. Decompounding is performed on the Dutch queries, since compound splitting improves the performance of a search with compound languages [18]. The decompounding algorithm used by WikiTranslate is based on the one described in [18]. To check if a term exists in the lexicon, a search is done on the article titles of Wikipedia and checked if there are any results. If the first part of a compound word is found, and the second part appears in the Wikipedia corpus as well, it is also treated as a compound part (however compound parts that have their own article are given precedence over parts that only appear in the corpus).
Note that because only Wikipedia is used, this cannot be done at indexing time. Therefore it only is performed on the query itself. Compound parts are added to the original query.

5.3. Step 1: Mapping the Query to Wikipedia Concepts

5.3.1. Step 1a: Search with whole query

As explained in section 2.2 Wikipedia can be viewed as a comparable corpus. The proposed approach is based on [7] and [8] as we also retrieve the best matching documents in the source language and use them to create a new query.
First the original query is put in Lucene, retrieving the most relevant Wikipedia concepts. The concepts can be retrieved by searching on the title, text, description or a combination of these fields. The top documents will be considered as relevant and will be used for translations. With this method word sense disambiguation is performed automatically [8]. A crucial step is to determine which top documents will be included in the final translation. Different experiments are carried out by limiting the score and/or allowing a maximum number of documents to be included.
Thus for our example (see section 3.2) a search is performed with the query "*atlantis mir koppeling eerste space shuttle aankoppeling tussen amerikaanse shuttle atlantis mir ruimte station*" on the text and title of Wikipedia articles (note that non-relevant words like *find, documents* etc. are not included). The final stemmed query looks as follows

```
(title:atlantis text:atlantis) (title:mir text:mir) (title:koppel text:koppel) (title:eerst
text:eerst) (title:spac text:spac) (title:shuttl text:shuttl) (title:aankoppel text:aankoppel)
(title:tuss text:tuss) (title:amerikan text:amerikan) (title:shuttl text:shuttl) (title:atlantis
text:atlantis) (title:mir text:mir) (title:ruimt text:ruimt) (title:station text:station)
```

After searching with this query the following concepts scored higher than de minimum score used: *"space shuttle atlantis"* and *"mir (ruimtestation)"*.

5.3.2. Step 1b: Search on every term of the query

However, there are cases where some terms in the query completely disappear from the top results, since they are not closely related to the other terms and the other terms appear more often. For example, the results of the Dutch query "*Geschiedenis van de literatuur"* (English: *history of literature*) will contain mostly articles about literature at the top. This would yield translations with words related to *literature*, but the term *history* would not appear in the final translation. Since this term is important for the relevance of the documents, leaving this term out will have a much lower precision. Thus, when only using the current approach, important terms may be left out of the final query, affecting the performance of the system.
To avoid this problem, every term in the query is searched separately to find Wikipedia concepts. This step is quite similar to the mapping of a query to dictionary entries, but Wikipedia offers new ways of mapping them. However, this introduces the problem we wanted to avoid: word sense ambiguity. Two different methods are used to map concepts to an individual term.
The first method (which we will call LINKS) uses the internal links of relevant concepts found in step 1. The expectation is that these terms are related to the top relevant documents of the first search. Therefore the internal links from the top documents of the first search are extracted. The search on every term is first only performed on these links. If no concepts can be found, or the found concepts are hardly relevant (i.e. have a low score), then the search is performed on the whole Wikipedia corpus. It is also possible to go deeper: including the internal links of the internal links from the top documents etc.
The second method (called CONTENTS) searches with the whole query, but gives the searched term more weight. The scope is limited to the title and the first paragraph of the text of a Wikipedia article. If an exact match with a Wikipedia title is found, the exact match is used.
For our example topic, a search is performed on every term of the query "*atlantis mir koppeling eerste space shuttle aankoppeling tussen amerikaanse shuttle atlantis mir ruimte station*". For the term *tussen (*English: *between)*, the following query is used:

```
((+title:tuss)^1.6) (descr:atlantis) (descr:mir) (descr:koppel) (descr:eerst) (descr:spac)
(descr:shuttl) (descr:aankoppel) (descr:tuss) (descr:amerikan) (descr:shuttl) (descr:atlantis)
(descr:mir) (descr:ruimt) (descr:station)
```

However, the term is not very relevant. It doesn't occur as an article in Wikipedia, but since the rest of the query is also included the following concepts are extracted: "*spaceshuttle atlantis*" and "*russische ruimtevaart"*. Because the concept "*spaceshuttle atlantis*" redirects to the concept "*space shuttle atlantis",* the latter one is used.

The system is not able to find a translation of the terms *aankoppeling* and *ruimte* at all. The search with the term *aankoppeling* has no results. The search with the term *ruimte* maps to the disambiguation page of *ruimte* which does not contain a cross lingual link.
The following concepts are recognized with step 1a and 1b for our example topic: *America, Atlantis (disambiguation), Coupling, Mir, Mir (disambiguation), Russian Federal Space Agency, Shuttle, Space Shuttle Atlantis, Space Shuttle program,* and *Station.*

5.4. Step 2: Creating the Translated Query

5.4.1. Step 2a (optional): Retrieve articles that redirect to the found Wikipedia concepts.

The translation can be expanded by adding synonyms and spelling variants of the found concepts to the query. This can be done by retrieving all Wikipedia articles in the English Wikipedia that redirect to the found concepts. For example, for the concept *“"space shuttle atlantis"* the following translations are added: *“atlantis (space shuttle), ov-104, shuttle atlantis, atlantis (space shuttle), atlantis (shuttle), ss atlantis, space shuttle atlantis, atlantis space shuttle”.*

5.4.2. Step 2b (optional): Weighting the query

The expectation is that the concepts retrieved by step 1a returns the most relevant concepts to the whole query. Therefore these concepts are given a higher weight than the other concepts.
Thus in our example the concepts *“space shuttle atlantis”* and *“mir (ruimtestation)”* are given a higher weight.

5.4.3. Step 2c: Creating the final query

For every found concept the translation can be obtained through the cross-lingual links. From every translation, terms like ‘*disambiguation’*, ‘*category*’, etc. are removed. The translation is also modified by removing non-word characters.
Sometimes a cross-lingual link refers to a part inside an article, having the form *w#y,* for example *eurovision_song_contest#winners.* It is clear that this translation has to be modified, since this translation is very unlikely to appear. An option is to take only the first or second part of the translation. The part after the hash sign is very specific. The first part gives the context. Therefore both terms are useful and worth retaining. Because of this, the translation is split. So this translation becomes “*eurovision song contest” ”winners”.*
Some concepts have titles like *‘atlantis (space shuttle)’.* The part between the parentheses gives more explanation about the meaning of the article. An option is to remove the part between the parentheses. But since this part gives more information about the word, it will be related to the query. Therefore the title is split in two parts. So this query becomes “*atlantis*” and “*space shuttle*”.
There are different possibilities to put the translations together. An option is putting quotation marks around every found title concept (e.g. “*Space Shuttle Atlantis”),* no quotation marks (e.g. *Space Shuttle Atlantis)*, or both (e.g. “*Space Shuttle Atlantis” Space Shuttle Atlantis).*
If no translation for a word can be found, the original word is added to the query.
The final translation of our example topic looks as follows (without step 2a):

```
"station"^1.0 station^1.0 "russian federal space agency"^1.0 russian^1.0 federal^1.0 space^1.0
agency^1.0 "mir"^1.0 mir^1.0 "coupling"^1.0 coupling^1.0 "america"^1.0 america^1.0 "shuttle"^1.0
shuttle^1.0 "space shuttle program"^1.0 space^1.0 shuttle^1.0 program^1.0 "space shuttle
atlantis"^3.0 space^3.0 shuttle^3.0 atlantis^3.0 "atlantis"^1.0 atlantis^1.0 "ruimte"^1.0
ruimte^1.0 "aankoppeling"^1.0 aankoppeling^1.0  "mir"^3.0 mir^3.0
```

Note that concepts from step 1a are given a higher weight (3.0). The rest of the translations have a standard weight (1.0).

6. Evaluation

6.1. Evaluation Method

The Dutch-English, French-English and the Spanish-English language pairs are tested. The test topics are in Dutch, French and Spanish, and the goal is to search in an English data collection.
Comparing the results of different language pairs is very interesting, since the system in theory should be language independent. Except from the preprocessing step (stemming), the system is the same for every language. It is also

interesting to see how much the size of the Wikipedia affects the results. The French Wikipedia is the largest one of the source languages that is used and contains more than 654 000 articles. Dutch follows with more than 435 000 and Spanish has more than 357 000 articles. WikiTranslate will be evaluated by comparing the mean average precision (MAP) of the cross-lingual system with the MAP of the monolingual system.
WikiTranslate is first evaluated with the data of CLEF 2006, 2005 and 2004. In this evaluation we use the data collections of the Los Angeles Times 1994 (113,005 documents) and the Glasgow Herald 1995 (56,472 documents), which contain English newspaper documents. The use of Wikipedia should fit these data collections since both contain a lot of named entities. Note however that the Wikipedia data used is from 2008, and the test collections are from 1994 and 1995. This might especially affect queries about persons and news items.
The best performing system is also evaluated with data of CLEF 2008. However note that a different data collection is used in the evaluation of 2008. As data collection the catalog of The European Library (1,000,100 documents) is employed, which is very sparse (many records only contain a title, author and subject heading information) and multilingual.

6.2. Results

6.2.1. Overall results of experiments with data of CLEF 2004, 2005 and 2006

Experiments have been carried out with 2 tasks: using only the title of the topic (T), or using the title and description (T+D). Tests are performed with the following systems: No word sense disambiguation (NO_WSD), word sense disambiguation using links (LINKS), word sense disambiguation through text (CONTENT) and word sense disambiguation through text and weighted query terms (CONTENT_W).
The basic underlying system uses parameters that are determined experimentally by varying the parameters over different ranges. Furthermore no query expansion is applied, every translation is added with and without quotation marks and decompounding is used with Dutch. From the description non-relevant words (*e.g. "find", "documents", "describe", "discuss" etc.*) are filtered out with a stop list, since these terms are not needed to translate. It will even affect the translation itself if these words will be translated incorrectly. The results from the experiments with Spanish (S), French (F) and Dutch (D) can be found below in the Appendix.
To compare the results of the different systems, the results are averaged per system and task over every tested language. The results can be seen in table 1. It is reasonable to average over the different years, since all topics made use of the same data collection and the formulation of the topics is similar.

Table 1. Summary of runs

Task	ID	Average (% Monolingual system)
T	NO_WSD	72,71 %
T	LINKS	71,88 %
T	CONTENT	74,89 %
T	CONTENT_W	72,70 %
T + D	NO_WSD	68,98 %
T + D	LINKS	71,44%
T + D	CONTENT	73,18%
T + D	CONTENT_W	74,98%

CONTENT_W seems to perform best. Therefore we have compared the average performance of French and Spanish using this system. Spanish had an average performance of 71,89% compared to the monolingual baseline. French had an average of 76,78%. However since Dutch topics were only available for the year 2004, it is not possible to compute an average for Dutch.
Since Dutch is a compound language, the results of using compound splitting are compared with not using compound splitting. The MAP of the Dutch runs using long queries (title and description) was on average 0.0595 higher when compound splitting was used.

When creating the final query different options with using quotation marks are possible. A random test set is used to experiment with using quotation marks, not using them or using both. Using both or no quotation marks showed differences in results per run, but averaging over the test set shows no significant difference. Only using quotation marks results in a decrease of 0.0990.
An option was including spelling variants, synonyms etc. through the redirects of the found concepts. However, results show that this degrades the performance significantly. The average decrease of a randomly selected test set was 0.1118.
Filtering non-related word significantly increases the performance. The average increase over a randomly selected test set was 0.0926.

6.3.2. Analysis of CLEF 2008

The system CONTENT_W using the title and description has been submitted in the CLEF ad hoc task 2008. The results can be found below:

Table 2. Results run 2008

Language	MAP
English (monolingual)	0.3407
French	0.2278 (66,86%)
Spanish	0.2181 (64,02%)
Dutch	0.2038 (59,82%)

To illustrate the weak and strong characteristics of WikiTranslate it is interesting to analyze the results of one run in depth. We will take the French run, which had the best performance. When analyzing the whole run with 50 topics, we see that 12 topics performed better than the original English topics and 38 topics performed worse. A comparison of the performance between the translations and original English queries can be found in figure 1.

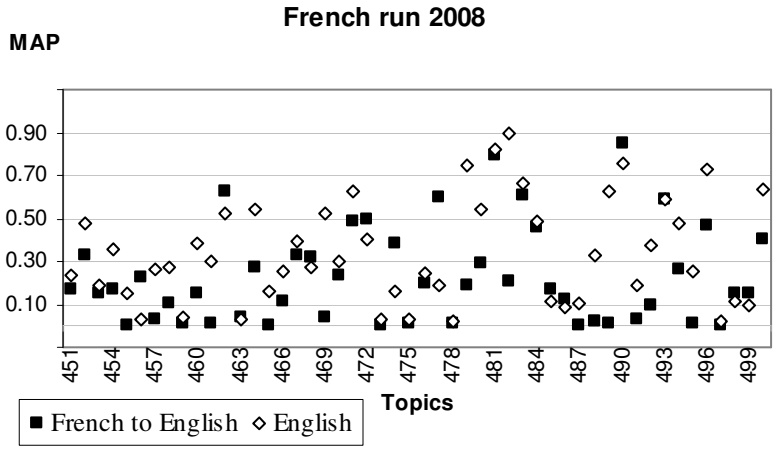

Figure 1. A comparison of French and English topics.

When analyzing the queries we see that sometimes new, but relevant terms are added with the new translations. For example the new query for topic 477 contains the term "*investment"* which wasn't included in the original English topic:

```
<top>
<num> 10.2452/477-AH </num>
<title> Web Advertising </title>
<desc> Find books about the distribution and impact of marketing campaigns on the World Wide Web
/ Internet. </desc>
</top>
```

Furthermore the translations *internet* and *advertising* were given a higher weight, which also will have increased the performance. The original English topic has a MAP of 0,0345. The translation has a MAP of 0,2299 (an increase of 0,1954).
However the translations of some queries are totally wrong. One of the worst performing is topic 457. The French version can be found below:

```
<top>
<num> 10.2452/457-AH </num>
<title> La chasse au grand gibier en Afrique </title>
<desc> Descriptions non fictives d'expéditions de chasse au grand gibier en Afrique. </desc>
</top>
```

The corresponding English topic:

```
<top>
<num> 10.2452/457-AH </num>
<title> Big Game Hunting in Africa </title>
<desc> Real-life descsriptions of big game hunting expeditions in Africa. </desc>
</top>
```

For this topic the English topic has a MAP of 0,2626 while the translated topic has a MAP of 0,0286 (a decrease of 0,2340). When looking at the translation of this topic, we see that the system had difficulties translating the term *fictives* (English: *fictional*). It mapped the concepts "*Planets in science fiction*" and "*Fictional brands*" to this term.
The bad coverage of adjectives in Wikipedia might have caused this.

7. Discussion

It is difficult to make a solid comparison with the performances of other systems. First of all since the approach of WikiTranslate is different than other approaches, it is reasonable to have a lower performance than state of the art systems that use well researched methods. It is also important to keep in mind that we used a standard information retrieval system (Lucene) and have not paid further attention to this. Therefore the performance of the English monolingual system is lower than that of state of the art systems. At the ad hoc task of CLEF 2004, 2005 and 2006 French, Spanish and Dutch are not chosen as a source language, which makes it even harder to compare. However, since the system achieves performances around 70 and 75% of the monolingual baseline, which are manually created queries, these results are very reasonable. The performance of the system with the dataset of 2008 is significantly lower. This might be caused because it is evaluated with a different data collection. Organizers of the Ad Hoc task 2008 have indicated that this task is quite different from previous Ad Hoc tasks. The data is very sparse and they estimate that only 66% of the English collection has an English subject header [20].
Table 1 shows that the results lie quite close to each other. Word sense disambiguation doesn't seem to improve the performance if only the title is used. However when also the description is used, word sense disambiguation does improve the performance.
For the task T + D the performance depends on the right stop words lists (with filtering words like *document, information* etc.). As the results shows, without filtering these words the performance decreases. This can be explained because WikiTranslate will retrieve concepts related to these terms, but not related to the query.
Putting titles of every found concept between quotation marks significantly decreases the performance of the system. This can be explained when we look at the translations. Some found concepts are too specific. Also some concepts have names like "*Soft and hard drugs*". When we search with quotation marks around it, the retrieval system will only retrieve documents that contain exactly this phrase. However, documents that contain these terms, but not exactly in this form are also relevant. Therefore fewer documents are retrieved with this method.
Query expansion using spelling variants and synonyms significantly decreases the performance of the system. This might be due to the expansion of every concept. Therefore wrongly recognized concepts are also expanded, including a lot of non related translations. Furthermore when we manually look at the redirects, some redirects are very global or not very related to the concept. The performance might be better if the expansion method is more refined, thus not expanding every concept. Also performance might improve if the original term is given more weight than its redirects.
The coverage of Wikipedia seems to be large enough to be used for translations. When inspecting the translations, it seems that some translations weren't missed because they were not covered, but because the system wasn't able to

find the corresponding concepts. These problems were sometimes caused by the shortcomings of the use of stemmers. For example, the term *boicots* (Spanish) wasn't stemmed properly, and therefore not mapped to the term (*boicot*). Translations that are missed are most of the times adjectives and common words. However, these terms are sometimes crucial (e.g. *longest*). WikiTranslate performs particularly well with translating proper nouns.
The analysis of one single run showed that some topics performed even better than the original ones. This indicates that this method is very promising.

8. Conclusion & Future Work

In this paper the system WikiTranslate is introduced that performs query translation using only Wikipedia as translation source. WikiTranslate maps queries to Wikipedia concepts and creates the final query through the obtained cross-lingual links. We have experimented with different methods to map a query to Wikipedia concepts. The best approach made use of the text and titles of the articles. Furthermore we have experimented with different methods to create a query given the Wikipedia concepts. Adding spelling variants and synonyms through redirects showed to decrease performance. Giving some concepts more weight can improve performance.
We have demonstrated that it is possible to achieve reasonable results using only Wikipedia. We believe therefore that it can be valuable alternative to current translation resources and that the unique structure of Wikipedia (e.g. text and internal links) can be very useful in CLIR. The use of Wikipedia might also be suitable for Interactive CLIR, where user feedback is used to translate the query, since Wikipedia concepts are very understandable for people.
An advantage of using Wikipedia is that it allows translating phrases and proper nouns especially well. In addition it is very scalable since it is easy to use the most up to date version of Wikipedia which makes it able to handle actual terms.
The coverage of Wikipedia for well-represented languages like Dutch, French and Spanish seems to be enough to get reasonable results. However, the major drawback of Wikipedia is that sometimes concepts are not covered (mainly common words).
We believe that with further research a higher performance can be achieved. In particular the method to map concepts can be refined. It is possible to make more use of the structure of Wikipedia, e.g. also using the category pages, disambiguation pages and making more use of the internal links. Furthermore a method to filter concepts that are not very related to the other retrieved concepts (already used by [14]) might improve performance.
Also experiments with different methods of preprocessing (e.g. using n-grams instead of stemming) can be interesting, since they might be more suitable.
The query weighting method used by the system is very basic, but already showed to improve performance. Therefore the expectation is that a more refined method can even improve performance more. Also adjusting the weights of concepts retrieved at step 2b can improve performance. Furthermore the added weights might also be more adjusted, for example to the relevance score given by Lucene.
Furthermore it would be interesting to explore other methods of query expansion using Wikipedia. A method would be to add the internal links that occur often at the retrieved concepts. Another possible method is adding concepts that appear as internal links in the first paragraph of the retrieved concepts. However since query expansion can sometimes cause query drift, it might be better to give the added concepts a lower weight.
To cope with translations that are not covered by Wikipedia (usually basic words and adjectives), it is possible to incorporate other resources like EuroWordNet [21] or a bilingual dictionary. A possibility is using these others resources if no translation can be found with Wikipedia.

Acknowledgements

This paper is based on research partly funded by IST project MESH (http://www.mesh-ip.eu) and by bsik program MultimediaN (http://www.multimedian.nl).

References

[1] S. F. Adafre and M. de Rijke, "Finding Similar Sentences across Multiple Languages in Wikipedia," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, pp. 62--69.

[2] M. Sanderson, "Word sense disambiguation and information retrieval," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* Dublin, Ireland: Springer-Verlag New York, Inc., 1994, pp. 142 – 151

[3] W. Kraaij, J.-Y. Nie, and M. Simard, "Embedding web-based statistical translation models in cross-language information retrieval," *Comput. Linguist.,* vol. 29, pp. 381-419, 2003.

[4] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* Dublin, Ireland: Springer-Verlag New York, Inc., 1994, pp. 61 - 69

[5] P. McNamee and J. Mayfield, "Comparing cross-language query expansion techniques by degrading translation resources," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* Tampere, Finland: ACM, 2002.

[6] L. Ballesteros and W. B. Croft, "Resolving ambiguity for cross-language retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* Melbourne, Australia: ACM, 1998, pp. 64 - 71

[7] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* Tampere, Finland: ACM, 2002, pp. 175 - 182.

[8] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the SPIDER system," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* Zurich, Switzerland: ACM, 1996, pp. 58 - 65

[9] J. Voss, "Measuring Wikipedia," in *the 10th International Conference of the International Society for Scientometrics and Informatics*, 2005, pp. 221--231.

[10] A. Halavais and D. Lackaff, "An Analysis of Topical Coverage of Wikipedia," *Journal of Computer-Mediated Communication,* pp. 429–440, 2008.

[11] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and Accessing Wikipedia as a Lexical Semantic Resource," *Data Structures for Linguistic Resources and Applications,* pp. 197-205, 2007.

[12] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation," in *the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, Rochester, 2007.

[13] C.-Y. Su, T.-C. Lin, and W. Shih-Hung, "Using Wikipedia to Translate OOV Term on MLIR," in *The 6th NTCIR Workshop* Tokyo, 2007

[14] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány, "Performing Cross-Language Retrieval with Wikipedia," in *CLEF 2007* Budapest, 2007.

[15] M. Potthast, B. Stein, and M. Anderka "A Wikipedia-based Multilingual Retrieval Model," in *30th European conference on information retrieval* Glasgow, Scotland, 2008.

[16] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM,* vol. 18, pp. 613-620, 1975.

[17] Lucene (2008). "Apache Lucene - Scoring" [online]. Available: http://lucene.apache.org/java/2_3_1/scoring.html

[18] V. Hollink, J. Kamps, C. Monz, and M. d. Rijke, "Monolingual Document Retrieval for European Languages," *Inf. Retr.,* vol. 7, pp. 33-52, 2004.

[19] Snowball (2007). "Stemming algorithms for use in information retrieval" [online]. Available: http://www.snowball.tartarus.org

[20] CLEF 2008 Ad-Hoc Track (2008) [online]. Available: http://www.clef-campaign.org/2008/2008Ad-hoc.html

[21] P. Vossen "EuroWordNet: a multilingual database for information retrieval" *In Proceedings of the DELOS workshop on Cross-language Information* Zurich, Switzerland , 1997.

Appendix

Table 1. Results task T

	NO_WSD	LINKS	CONTENT	CONTENT_W
	MAP			
T 2004 E	0.3079			
S	0.1973 (64,08%)	0.1956 (63,53%)	0.2080 (67,55%)	0.2160 (70,15%)
F	0.2082 (67,62%)	0.2192 (71,19%)	0.2186 (71,00%)	0.2083 (67,65%)
D	0.2159 (70,12%)	0.2042 (66,32%)	0.2249 (73,04%)	0.2076 (67,42%)
T 2005 E	0.2698			
S	0.1890 (70,05%)	0.1846 (68,42%)	0.1894 (70,20%)	0.1855 (68,75%)
F	0.1856 (68,79%)	0.1881 (69,72%)	0.2051 (76,02%)	0.2021 (74,91%)
T 2006 E	0.2803			
S	0.2280 (81,34%)	0.2222 (79,27%)	0.2314 (82,55%)	0.2186 (77,99%)
F	0.2437 (86,94%)	0.2375 (84,73%)	0.2351 (83,87%)	0.2299 (82,01%)

Table 2. Results task T+D

	NO_WSD	LINKS	CONTENT	CONTENT_W
	MAP			
T+D 2004 E	0.3107			
S	0.2021 (65,05%)	0.2275 (73,22%)	0.2054 (66,11%)	0.2216 (71,32%)
F	0.2464 (79,30%)	0.2343 (75,41%)	0.2511 (80,82%)	0.2373 (76,37%)
D	0.2451 (78,89%)	0.2642 (85,03%)	0.2447 (78,76%)	0.2450 (78,85%)
T+D 2005 E	0.3542			
S	0.1934 (54,60%)	0.2161 (61,01%)	0.2320 (65,50%)	0.2491 (70,33%)
F	0.1865 (52,65%)	0.2003 (56,55%)	0.2341 (66,09%)	0.2472 (69,79%)
T+D 2006 E	0.3235			
S	0.2343 (72,43%)	0.2390 (73,88%)	0.2390 (73,88%)	0.2395 (74,03%)
F	0.2390 (73,88%)	0.2426 (74,99%)	0.2623 (81,08%)	0.2723 (84,17%)