ABSTRACT
In this paper the system WikiTranslate is introduced that performs query translation for cross-lingual information retrieval (CLIR) that only uses Wikipedia. Queries will be 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 performances between 70 and 75% of the monolingual baseline.

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 to enter a query in different languages.

There are two approaches to handle CLIR: Translating the query into the target language or translating the documents into the source language and search 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. In this paper we will treat Wikipedia articles as representations of concepts (units of knowledge). An advantage is that such articles provide more context than sources like online dictionaries. This can be used to perform word sense disambiguation [1], also a well known problem of CLIR. Wikipedia also contains 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 [2]. Therefore redirects may be used for query expansion.

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. [3]) and then extracting translations. However, the first method is chosen since it allows us to make use of the structure of Wikipedia (e.g. using the cross-lingual links, text and internal links in the articles) and thereby enabling us to investigate new possibilities to perform query translation.

The main research question of this paper is: How can Wikipedia be used for query translation 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 very 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 very useful source for CLIR.

Though this is an approach that potentially could be applied to any language pair that is sufficiently covered by Wikipedia, the system will be evaluated by searching with Dutch, French and Spanish queries in an English data collection.

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.

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 [1].

Two important observations are made by [4]. 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, doesn’t 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 [4].

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 [7] 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 [8, 9]. The first step of [9] 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. [8] follows the same approach except that it doesn’t select new query terms but makes a relevance model. The approach of [8] also automatically achieves word sense disambiguation since the method makes use of co-occurrence statistics.

2.2 Wikipedia

Wikipedia is an online, multilingual, free-content encyclopedia where every user can make a contribution to. Soon after its start it began to grow exponentially [10]. 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 [11]. The characteristics of Wikipedia makes it suitable as a semantic lexical resource [2].

Wikipedia has been used for automatic word sense disambiguation [12] and for translation. [13] used it to translate out of vocabulary words and [14] used it to translate queries. Both made use of the cross-lingual links available in Wikipedia to translate terms. One of the steps [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 [3]. The notion that Wikipedia can be treated as a comparable corpus is new and hasn’t 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. PROPOSED APPROACH

3.1 Outline

The approach used by WikiTranslate exists of two important steps: mapping the query to Wikipedia concepts and creating the final query using these found concepts. Below is an overview of the proposed approach:

Step 1: Mapping the Query to Wikipedia concepts

First the most relevant concepts 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 add 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 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).

Using the features of Wikipedia allows 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 is 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.

The use of Wikipedia also has some disadvantages. The coverage is less than dictionaries on basic 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.

3.2 Example

To illustrate the steps of the proposed approach in this paper topic C230 from the Ad hoc task of CLEF (Cross Language Evaluation Forum) of 2004 is chosen. 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 (which is used for evaluation):

<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>

4. EXPERIMENTAL SETUP

4.1 Lucene

Lucene will be used as the underlying retrieval system to retrieve Wikipedia articles at the 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.

4.2 Preprocessing

4.2.1 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 makes use of 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 summarization of the article or most important information about the article (and thus containing the most related words). If the article in fact redirects to another article, the title of the referred page is also stored.

To enhance comparability, the same preprocessing method is used for all languages. We have chosen 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]. Stop words are removed with the lists from the Snowball algorithm [17].

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.

4.2.2 Compound Words

Compound words are words that are formed from multiple words. For example 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. Decomposing 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 is found, and the second part appears in the Wikipedia corpus as well, then it 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.

4.3 Step 1: Mapping the Query to Wikipedia Concepts

4.3.1 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 [8] and [9] 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 inputted in Lucene, retrieving the most relevant Wikipedia concepts. The concepts can be retrieved by searching on the title, text, description or 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 [9].

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 final, documents etc. are not included). The final stemmed query looks as follows


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

4.3.2 Step 1b: Search on every term of the query

However, there are cases where some terms in the query completely disappear in the top results, since they are not very related to the other terms and the other terms appear more often. For example, the results of the 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 quit 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 queries 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:

\[
(+\text{title:tussen}^1.6) \equaltext{descr:atlantis} \equaltext{descr:mir} \equaltext{descr:koppel} \equaltext{descr:eerst} \equaltext{descr:ruimte} \\
\equaltext{descr:shuttle} \equaltext{descr:aankoppel} \equaltext{descr:tuss} \\
\equaltext{descr:amerikan} \equaltext{descr:shuttle} \equaltext{descr:atlantis} \\
\equaltext{descr:mir} \equaltext{descr:ruimte} \equaltext{descr:station}
\]

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 wasn’t able to find a translation of the terms aankoppeling and ruimte at all. The search with the term aankoppeling had no results. The search with the term ruimte mapped to the disambiguation page of ruimte, which didn’t 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.

4.4 Step 2: Creating the Translated Query

4.4.1 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".

4.4.2 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.

4.4.3 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 Atlantisc 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).

5. EVALUATION

5.1 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 will be 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 would also be 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. The amount of decrease of the MAP of the CLIR system indicates how well the system performs.

MAP is the mean of the Average Precision per query. The Average Precision is calculated like this:

\[ AP = \frac{\sum_{r} P(r) \times rel(r)}{C} \]

\( P(r) \) is the precision for a given rank. \( C \) is the number of relevant documents. \( Rel() \) is a binary function of a given relevance of a rank.

5.2 Test Collection

The system will be evaluated with data from the ad hoc task of CLEF (Cross Language Evaluation Forum). 50 topics per source language are available. A topic consists of the following data:

- **Title**
- **Description**: a one sentence description
- **Narrative**: a more complex "narrative" specifying the relevance criteria

WikiTranslate will participate in the ad hoc task of CLEF 2008. In that edition the narrative information is not available. Therefore we won’t use it in this evaluation either.

The final goal is to search in an English collection of documents. 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.

WikiTranslate is evaluated with the topics and relevance assessments of CLEF 2006, 2005 and 2004.

5.3 Results

5.3.1 Analysis of One Run.

Before looking at the overall results, it is very interesting to analyze the results of one run in depth. We have picked the Dutch run with topics from 2004, using the title and description of a topic, and performing word sense disambiguation with the text and title of the system. (CONTENTS)

When analyzing the whole run with 50 topics, we see that 10 topics performed better then the original English queries. 10 topics performed exactly the same, most of them having a MAP of 0.0000 or 1.0000 and 30 topics performed worse than the original English query. A comparison of the performance between the translations and original English queries can be found in figure 1.

When analyzing the queries we see that sometimes new, but relevant phrases are added with the new translations. For example the translation of topic 212 contains the phrase “Drugs in sport” which wasn’t included in the original English topic:

```xml
<title> Sportswomen and Doping </title>
<desc> Find documents talking about women who have been accused of using drugs to improve their sporting results. </desc>
```

The original English topic has a MAP of 0.0418. The translation has a MAP of 0.2632 (an increase of 0.2214).

However the translations of some queries are totally wrong. One of the worst performing is topic C242. The Dutch version can be found below:

```xml
<title> Duurrecord in de ruimte </title>
<desc> Vind documenten over het langste verblijf van een mens in de ruimte, inclusief de naam van de astronaut </desc>
```

The corresponding English topic:

```xml
<title> Record Permanence in Space </title>
<desc> Find documents on the longest stay of a human being in space, including the cosmonaut's name. </desc>
```

For this topic the English topic has a MAP of 0.8750 while the translated topic has a MAP of 0.0194 (a decrease of 0.8556). When looking at the translation of this topic, we see that the system failed to translate the term *verblijf* (English: *stay*) and *langste* (English: *longest*). However, these terms are very crucial for this topic. These translations are missed because these terms are not covered by Wikipedia. This might be caused because basic words and adjectives are not well covered.

An other worse performing topic is topic 233 with a MAP decrease of -0.6204.

The Dutch topic:

```xml
<title> Broeikaseffect </title>
<desc> Vind documenten over wereldwijde klimaatsveranderingen en meer specifiek discussies over het bestaan van het "broeikaseffect". </desc>
```
The corresponding English topic:
<title>Greenhouse Effect</title>
<desc>Find documents about global climate changes and, in particular, discussions about the existence of the "greenhouse effect".</desc>

Unfortunately WikiTranslate only recognized the following concepts: Existence, Greenhouse effect and World Wide Web. It thus missed concepts belonging to these important terms global, climate and changes. Furthermore, the concept Word Wide Web is mapped to the term wereldwijde (English: global), another indication that the system performs not well with translating adjectives.

### 5.3.2 Overall results

To evaluate the system the performance of the bilingual system is compared with the performance of the monolingual English system. 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.

Results from the source languages Spanish (S), French (F) and Dutch (D) can be found below. They have been tested with the available data from the years 2004, 2005 and 2006. Also the results of the English (E) monolingual system are mentioned.

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.

#### Table 2. Results task T+D

<table>
<thead>
<tr>
<th></th>
<th>NO_WSD</th>
<th>LINKS</th>
<th>CONTENT</th>
<th>CONTENT_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+D 2004 E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.2071</td>
<td>0.2275</td>
<td>0.2054</td>
<td>0.2216</td>
</tr>
<tr>
<td>E</td>
<td>(65,03%)</td>
<td>(73,22%)</td>
<td>(66,11%)</td>
<td>(71,32%)</td>
</tr>
<tr>
<td>S</td>
<td>0.2464</td>
<td>0.2343</td>
<td>0.2311</td>
<td>0.2373</td>
</tr>
<tr>
<td>F</td>
<td>(79,30%)</td>
<td>(75,41%)</td>
<td>(80,82%)</td>
<td>(76,37%)</td>
</tr>
<tr>
<td>D</td>
<td>0.2451</td>
<td>0.2447</td>
<td>0.2450</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(78,89%)</td>
<td>(78,76%)</td>
<td>(78,85%)</td>
<td></td>
</tr>
<tr>
<td>T+D 2005 E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.1934</td>
<td>0.2161</td>
<td>0.2320</td>
<td>0.2491</td>
</tr>
<tr>
<td>E</td>
<td>(54,60%)</td>
<td>(61,01%)</td>
<td>(65,50%)</td>
<td>(70,33%)</td>
</tr>
<tr>
<td>S</td>
<td>0.1865</td>
<td>0.2003</td>
<td>0.2341</td>
<td>0.2472</td>
</tr>
<tr>
<td>F</td>
<td>(52,65%)</td>
<td>(56,55%)</td>
<td>(66,09%)</td>
<td>(69,79%)</td>
</tr>
<tr>
<td>D</td>
<td>0.1871</td>
<td>0.2186</td>
<td>0.2195</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(57,19%)</td>
<td>(68,75%)</td>
<td>(69,79%)</td>
<td></td>
</tr>
<tr>
<td>T+D 2006 E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.2343</td>
<td>0.2390</td>
<td>0.2370</td>
<td>0.2395</td>
</tr>
<tr>
<td>E</td>
<td>(72,43%)</td>
<td>(73,88%)</td>
<td>(73,88%)</td>
<td>(74,03%)</td>
</tr>
<tr>
<td>S</td>
<td>0.2396</td>
<td>0.2462</td>
<td>0.2623</td>
<td>0.2723</td>
</tr>
<tr>
<td>F</td>
<td>(73,88%)</td>
<td>(74,99%)</td>
<td>(81,08%)</td>
<td>(84,17%)</td>
</tr>
</tbody>
</table>

Since Dutch is a compound language, the results of using compound splitting are compared with not using compound splitting. The average increase over the Dutch runs using description showed an increase of 5,95% 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 9,90%.

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 11,18%.

Filtering non-related word significantly increases the performance. The average increase over a randomly selected test set was 9,26%.

To compare the results of the different systems, the results are averaged per system and task over every tested language. 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.
It is difficult to make a solid comparison with the performances of 76,78%. However, since Dutch topics were only available for this system, Spanish had an average performance of 71,89% compared to the average performance of French and Spanish using CONTENT_W seems to perform best. Therefore we have right stop words lists (with filtering words like For the task T + D the performance is very dependent on the recall 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 might be due to the expansion of every concept. Therefore, wrongly recognized concepts are also expanded, including a lot of non related translations. Also when manually looking 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 big enough to be used for translations. When manually looking at 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 basic 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.

6. DISCUSSION

It is difficult to make a solid comparison with the performances of other systems. First of all at the ad hoc task of CLEF 2004, 2005 and 2006 French, Spanish and Dutch are not chosen as a source language. Furthermore, 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 only a simple retrieval underlying retrieval system is used (Lucene with standard parameters). Therefore the performance of the English monolingual system is lower than of state of the art systems. However, since the system achieves performances around 70 and 75% of the monolingual baseline, which are manually created queries, the results are very reasonable.

Table 3 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 is very dependent 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 manually looking 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.

Table 3. Summary of runs

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Average (% Monolingual system)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T NO WSD</td>
<td>72.71 %</td>
</tr>
<tr>
<td>T LINKS</td>
<td>71.88 %</td>
</tr>
<tr>
<td>T CONTENT</td>
<td>74.89 %</td>
</tr>
<tr>
<td>T CONTENT_W</td>
<td>72.70 %</td>
</tr>
<tr>
<td>T + D NO WSD</td>
<td>68.98 %</td>
</tr>
<tr>
<td>T + D LINKS</td>
<td>71.44 %</td>
</tr>
<tr>
<td>T + D CONTENT</td>
<td>73.18 %</td>
</tr>
<tr>
<td>T + D CONTENT_W</td>
<td>74.98 %</td>
</tr>
</tbody>
</table>

Table 3. Summary of runs

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. Also when manually looking 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 big enough to be used for translations. When manually looking at 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 basic 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.

7. 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. Adding spelling variants and synonyms through redirects showed to decrease performance. Giving some concepts more weight can improve performance.

We have demonstrated that Wikipedia is a useful source for query translation in CLIR. We believe that the unique structure of Wikipedia (e.g. text and internal links) can be very useful for use 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 easy 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 basic 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. An other 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 [19] or a bilingual dictionary. A possibility is using these other resources if no translation can be found with Wikipedia.

ACKNOWLEDGEMENTS
I would like to thank my supervisors Claudia Hauff, Djoerd Hiemstra, Franciska de Jong and Dolf Trieschnigg for their valuable guidance during the research.

REFERENCES