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

University of Twente,
The Netherlands
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 formulated in Dutch, French and Spanish in an English data collection. The system achieved a performance of 67% compared to the monolingual baseline.

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

1 Introduction

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 followed to obtain translations. The aim of this research is to explore the possibilities of Wikipedia for query translation in CLIR.

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

We treat Wikipedia articles as representations of concepts (i.e. units of knowledge). WikiTranslate maps the query to Wikipedia concepts. Through the cross-lingual links translations of the concepts in another language are retrieved. This raises the following sub questions: *How can queries be mapped to Wikipedia concepts?* and *How to create a query given the Wikipedia concepts?*

Our method uses the unique structure of Wikipedia, enabling us to investigate new possibilities to perform query translation. 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 [1], which might make it suitable to handle translations of proper names.

- Continuous contributions of a large community keep the information up-to-date
- Wikipedia articles provide more context in comparison with sources like online dictionaries. This can be used to perform word sense disambiguation [2].
- Presence of redirect pages; pages that represent alternative names of concepts (e.g. synonyms, abbreviations and spelling variants [1]) and that consist of a link that directs to the main article it represents. They may be used for query expansion.

However, the coverage of common words in Wikipedia is smaller than translation dictionaries and some terms have many senses, some very specific and uncommon, making word sense disambiguation more difficult. For example in Wikipedia the term *house* has senses like a novel, song, operating system or a game.

The overview of this paper is as follows. 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

Kraaij et al. [3] make an important observation about CLIR. The final query delivered to the system does not have to be a single translation. Including synonyms and related words can in fact improve performance. One approach to accomplish this is with query expansion or using parallel corpora (e.g. [4,5]). In the first step of Sheridan et al. [5], the best matching documents in the source language are retrieved. Next, frequently occurring words in comparable documents in the target language are selected to compose the final query. Lavrenko et al. [4] follows the same approach except that their method creates a relevance model in the target language.

Wikipedia is an online, multilingual encyclopedia to which everyone can contribute. Its characteristics make it suitable as a semantic lexical resource [1]. Wikipedia has been used for automatic word sense disambiguation [6] and for translation. Su et al. [7] use it to translate out of vocabulary words and Schönhofen et al. [8] use it to translate queries. The notion that it can be treated as a comparable corpus is new and has not been researched much yet except by Potthast et al[9]. Wikipedia can be seen as a comparable corpus since articles are represented in different languages and connected through cross-lingual links.

3 Proposed approach

The approach used by WikiTranslate consists of two important steps: mapping the query in source language to Wikipedia concepts and creating the final query in the target language using these found concepts.

The first step maps the query to Wikipedia concepts. First, the most relevant concepts to the query are extracted after a search with the whole query (step 1a). Next, a search on every term of the query is performed (step 1b) using the internal links from the concepts retrieved with step 1a (called LINKS) or using the text and title of the Wikipedia articles (called CONTENTS).

The second step creates the translation. First, we add articles that redirect to the found Wikipedia concepts to include synonyms and spelling variants (step 2a). Furthermore articles retrieved with step 1a are given more weight (step 2b). Finally, the final query is created using the found concepts (step 2c).

This approach differs from traditional approaches, since we make use of the text and internal links, which are not available for approaches based on dictionaries and parallel corpora. This approach also differs from other approaches using Wikipedia. Su et al. [7] and Schönhofen et al. [8] have only used Wikipedia to enhance their translations. An advantage of our approach is that it allows extraction of phrases from the topics, since the titles of Wikipedia articles are often phrases. Furthermore by adding the top documents from step 1a, the most relevant concepts to the whole query are added. Also related concepts can be added, creating a kind of query expansion effect.

4 Experimental setup

Lucene is used as the underlying retrieval system to retrieve Wikipedia articles. From each article the title, text and cross-lingual links are extracted. The first paragraph of an article is extracted as well, which is called *description*. Because long articles tend to score lower, 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. If the article is a redirect page, the title of the referred page is also stored. Wikipedia articles that represent images, help pages, templates, portal pages and pages about the use of Wikipedia are excluded. 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 [10]. Stemming is best for Dutch and Spanish, but 4-gramming is more suitable for English and French [10]. We use Snowball stemmers to perform stemming [11]. Words are removed with the lists from the Snowball algorithm [11].

To illustrate the steps of the proposed approach we translate the following topic (C230 from the Ad hoc task of CLEF 2004) from Dutch to English:

```
<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>
```

(English: *Atlantis-Mir Docking, Find documents reporting the first space shuttle docking between the US shuttle Atlantis and the Mir space station*).

4.1 Step 1: Mapping the Query to Wikipedia Concepts

This step is based on [4] and [5] 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]. We set a minimum score and maximum number of documents to be included to determine which top documents will be included.

Our example finds the concepts "*space shuttle atlantis*" and "*mir (ruimtestation)*" with the following stemmed query:

```
   (title:atlantis  text:atlantis)  (title:mir  text:mir)  (title:koppel
text:koppel) (title:eerst text:eerst)..(title:station text:station)
```

We also search for every term of the query separately, because with the previous step some terms may not be found. For example, the query *history of literature* yields mostly articles about literature (missing the term *history*). 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. 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 are 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. An exact match has precedence over this step. For the term *tussen* (English: *between*) from our example, the following query is used:

```
   ((+title:tuss)^1.6) (descr:atlantis) .. (descr:ruimt) (descr:station)
```

The following concepts are recognized for our example topic: *America, Atlantis (disambiguation), Coupling, Mir, Mir (disambiguation), Russian Federal Space Agency, Shuttle, Space Shuttle Atlantis, Space Shuttle program,* and *Station.*

4.2 Step 2: Creating the Translated Query

The translation can be expanded by adding the redirect pages referring to the found concepts (adding synonyms etc.). For the concept "*space shuttle atlantis*" the following translations are added: "*atlantis (space shuttle), ov-104, ss atlantis* etc".

The expectation is that the concepts retrieved by step 1a returns the most relevant concepts. Therefore these concepts are given a higher weight than the other concepts For every found concept the translation can be obtained through the cross-lingual links. From every translation, terms like *disambiguation*, *category*, etc. and non-word characters are removed. Translations like *w#y* and *w(y)* are split into *w* and *y*.

There are different possibilities to put the translations together. We can include every found translation as a phrase query (e.g. "x y"), as an OR query of its terms (e.g. x y), or both (e.g. “x y” x y). 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  ….
space^3.0 shuttle^3.0 atlantis^3.0.. "mir"^3.0 mir^3.0
```

Note that concepts from step 1a (*space shuttle atlantis* and *mir (ruimtestation)*) are given a higher weight (3.0). Other concepts have a standard weight (1.0).

5 Evaluation

WikiTranslate is evaluated on retrieval of English documents using translated Dutch, French and Spanish queries. The system is first evaluated with the data of CLEF 2006, 2005 and 2004. The best performing system is also evaluated with data of CLEF 2008. Note that a different data collection is used in the evaluation of 2008.

Experiments have been carried out using only the title of the topic (T), or using the title and description (T+D) of a topic. 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. Furthermore, no query expansion is applied and every translation is added as a phrase query and as an OR-query of its terms. A stop list is used to filter particularly query words (e.g. “*find*”, “*documents*”, “*describe*”, “*discuss*” etc.) from the description.
To compare the results of the different systems, the results are averaged per system and task over every tested language. Table 1 shows the results. For each run the MAP of the bilingual system is compared with the MAP of the monolingual system (%Mono).

Table 1. Summary of runs 2004, 2005 and 2006

Task	ID	% Mono.	Task	ID	% Mono.
T	NO_WSD	72.71%	T+D	NO_WSD	68.98%
T	LINKS	71.88%	T+D	LINKS	71.44%
T	CONTENT	74.89%	T+D	CONTENT	73.18%
T	CONTENT_W	72.70%	T+D	CONTENT_W	74.98%

CONTENT_W with T + D performs best. Averaging the runs with these settings over the years 2004, 2005 and 2006 shows us that Spanish had an average performance of 71.89% and French had an average of 76.78%.

When creating the final query, different options with using phrase queries and OR queries are possible. Including translations only as a phrase query results in a average MAP decrease of 0.0990. Random tests are used to determine the effect of different steps. Including redirects showed an average MAP decrease of 0.118. Filtering non-related words lead to an MAP increase of 0.0926.

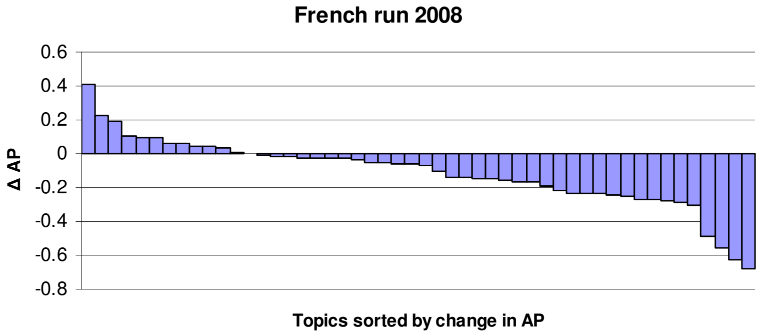
The system CONTENT_W (using T+D) has been submitted to the CLEF ad hoc task 2008. The results can be found in table 2.

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%)

The French run (which contains 50 topics), which had the best performance, is analyzed in more detail. 12 translations performed better than the original topics and 38 performed worse. An overview can be found in figure 1.

Figure 1. A comparison of original (French) and translated (English) topics.



When analyzing the queries we see that sometimes new, but relevant terms are added with the new translations. For example the translation for topic 477 contains the term "*investment*" which wasn't included in the original English topic about web advertising. Furthermore the translations *internet* and *advertising* were given a higher weight. The translation had an increase of AP from 0.1954 (from 0.0345 to 0.2299).

However the translations of some queries are totally wrong. One of the worst performing is topic 457. The translation showed an AP decrease of 0.2340 (from 0.2626 to 0.0286). 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.

6 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. We also used a standard information retrieval system (Lucene) and have not paid further attention to this. 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 due to use of a different data collection [12].

Table 1 shows that word sense disambiguation doesn't improve when we only use the title, but it improves if we also use the description. For the task T + D the performance depends on the right stop words lists. Without filtering these words the performance decreases. This can be explained because WikiTranslate retrieves concepts related to these terms, but not related to the query.

Including every found translation only as a phrase query significantly decreases the performance of the system. Query expansion using spelling variants and synonyms also decreases the performance of the system. Because every concept is expanded, 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.

WikiTranslate performs particularly well with translating proper nouns. Translations that are missed are most of the times adjectives and common words. However, these terms are sometimes crucial (e.g. *longest*). Sometimes translations were missed, because the system wasn't able to find the corresponding concepts due to shortcomings of the used stemmers.

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. The best approach uses the text and titles of the articles.

We have demonstrated that it is possible to achieve reasonable results using only Wikipedia. We believe that it can be valuable alternative to current translation resources and that the unique structure of Wikipedia can be very useful in CLIR. The use of Wikipedia might also be suitable for Interactive CLIR, where user feedback is used to translate, since Wikipedia concepts are very understandable for people.

Wikipedia allows translating phrases and proper nouns especially well. In addition it is very scalable since the most up to date version of Wikipedia can be used. The coverage of Wikipedia for the languages Dutch, French and Spanish seems to be enough to get reasonable results. The major drawback of Wikipedia is the bad coverage of common words. To cope with missed translations other resources like EuroWordNet [13] or a bilingual dictionary might be incorporated.

We believe that with further research a higher performance can be achieved. The method to map concepts can be refined by also using pages like disambiguation pages, and by filtering concepts which are not very related to the other retrieved concepts (used by [8]). Also the query weighting method can be refined.

It would be also interesting to explore other methods of query expansion using Wikipedia. Internal links that occur often at the retrieved concepts or internal links in the first paragraph of the retrieved concepts could be added. However since query expansion can cause query drift, it might be better to give the added concepts a lower weight. Furthermore we should only expand very relevant and related concepts.

Acknowledgments: 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. Zesch, T., Gurevych, I., Mühlhäuser, M.: Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In: Data Structures for Linguistic Resources and Applications, pp. 197-205, 2007.
2. Sanderson, M.: 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. Kraaij, W., Nie, J.-Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. In: Comput. Linguist., vol. 29, pp. 381-419, 2003.
4. Lavrenko, V., Choquette, M., Croft, W. B.: 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.
5. Sheridan, P., Ballerini, J. P.: 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
6. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. In: the North American Chapter of the Association for Computational Linguistics (NAACL 2007), Rochester, 2007.
7. Su, C.-Y., Lin, T.-C., Shih-Hung, W.: Using Wikipedia to Translate OOV Term on MLIR. In: The 6th NTCIR Workshop Tokyo, 2007
8. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Performing Cross-Language Retrieval with Wikipedia. In CLEF 2007 Budapest, 2007.
9. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based Multilingual Retrieval Model. In: 30th European conference on information retrieval Glasgow, Scotland, 2008.
10. Hollink, V., Kamps, J., Monz, C., Rijke, M. de: Monolingual Document Retrieval for European Languages. In: Inf. Retr., vol. 7, pp. 33-52, 2004.
11. Stemming algorithms for use in information retrieval, http://www.snowball.tartarus.org
12. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad Hoc Track Overview. In: Working Notes for the CLEF 2008 Workshop
13. Vossen, P.: EuroWordNet: a multilingual database for information retrieval. In: Proceedings of the DELOS workshop on Cross-language Information Zurich, Switzerland , 1997.