
Dong Nguyen, Arnold Overwijk, Claudia Hauff, Dolf R.B. Trieschnigg, Djoerd Hiemstra, and Franciska de Jong

University of Twente,
The Netherlands
{dong.p.ng,arnold.overwijk}@gmail.com,
{c.hauff,trieschn,f.m.g.dejong}@ewi.utwente.nl
hiemstra@cs.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).