Cross Language Indexing and Retrieval of the Cypriot Digital Antiquities Repository

Dayu Yuan
Dept. of Computer Science and Engineering
The Pennsylvania State University
University Park, Pennsylvania, USA
duy113@psu.edu

Prasenjit Mitra
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania, USA
pmitra@ist.psu.edu

ABSTRACT

We design and implement a cross-language retrieval system for the Cypriot Digital Antiquities Repository (cyDAR). Users can query either by English and Ancient Greek to search for documents written in Ancient Greek. Because of the lack of dictionary and parallel corpus, we use translation machine to translate the documents. We index both the original Ancient Greek text and translated English text to facilitated multi-language search.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval; I.2.7 [**Natural Language Processing**]: Machine translation

Keywords

Cross-language information retrieval; Digital Library

1. INTRODUCTION

cyDAR contains works describing scientific, philosophical, and social commentaries from antiquity to early Christian era. These works are written in Ancient Greek. It is important to design a cross-language retrieval system to support the search by both English and Ancient Greek. Users, represented by historians or archaeologists, may input a query in Ancient Greek describing what are found in a newly discovered document. They are interested in finding cyDAR documents that are similar to the query. Another type of users with information needs may formulate a query in English and search for related documents. The major goal of the cyDAR project is to help both types of users. In this paper, we design a cross-language retrieval platform to provide multi-lingual accesses to these priceless cyDAR documents.

A cross-language retrieval system can be designed with either *document translation* or *query translation*. For *document translation*, the cyDAR documents are first translated to English. And then, both the original corpus and the translated corpus are indexed separately for search. When an Ancient Greek query comes in, the system looks up the Ancient Greek index. When an English query comes in, the system looks up the English index. For a mixed query containing both the Ancient Greek and English terms, the system searches for both indexes and find the answer by combining results found on both indexes. The *query translation methods* only index the original documents (in Ancient Greek). When an English query or a mixed query comes, the system first translates the query into Ancient Greek and then searches for the documents related to the translated query. Although saves the off-line processing time, the query translation methods take more time on on-line query processing.

We adopt the document translation strategy based on users' requirements. Traditional users search on cross-language systems with no-English queries because they are incapable of formulating a query in English. Users of our system formulate a query in Ancient Greek because they want to find results related to the query in its original form (in Ancient Greek). To help users understand the search results, an English translation of the results are returned as well. Hence, translating the original corpus to English is inevitable and the document translation methods fit better than query translation methods. We further discuss the translation algorithms used for translation. Three candidates, i.e., machine-readable dictionary, statistical matching and machine translation are discussed in details. We choose to use machine translation for our system because of its good performance as shown in CLEF 09 'Ad-hoc track' [2].

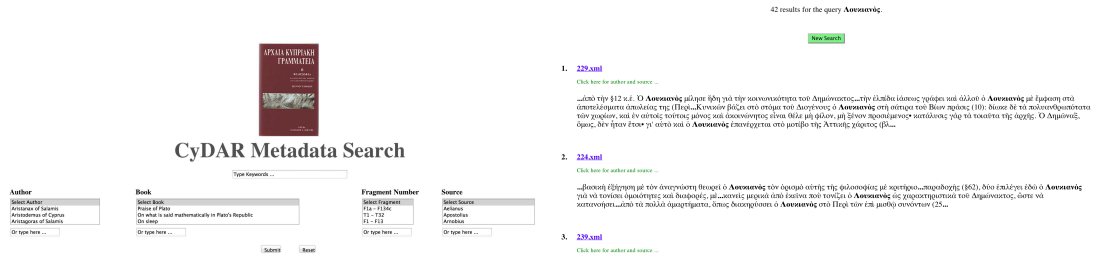
2. DESIGN AND IMPLEMENTATION

Data Description: The cyDAR documents are written in Ancient Greek. They also have Modern Greek as translation. Ancient Greek is written in italics and Modern Greek is in normal font. This observation helps us distinguish the Modern Greek from Ancient Greek. The raw data is a collection of books, each of which contains multiple essays. Each essay is a work describing scientific, philosophical, or social commentary. We define each essay as a retrieval unit. It contains (may be partially) the original description in Ancient Greek, translated version in Modern Greek and commentary in Modern Greek. Other information, such as authors and sources, is also included.

System Architecture: Our system is implemented based on document translation. Figure 1 shows the overall design of our system. The system comprises three components: preprocessing, indexing and searching. In the preprocess-

DocEng'13, September 10–13, 2013, Florence, Italy.

ACM 978-1-4503-1789-4/13/09.
http://dx.doi.org/10.1145/2494266.2494298.



(a) (b)

Figure 2: Search Interface and Result

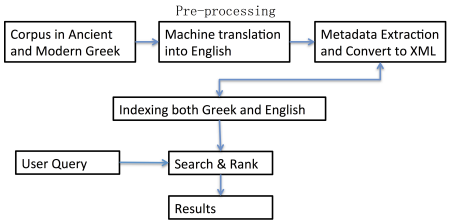


Figure 1: System Framework

ing step, the cyDAR documents are first segmented into retrieval units, and then translated into English by translation machines (will introduce later). Metadata is further extracted and each document is converted to a XML file. In the indexing step, those XML files are indexed by Solr (lucene.apache.org/solr/) with language-aware stemming. The indexing and matching steps are standard procedures as in monolingual retrieval system, we focus on the preprocessing step, especially the translation step.

2.1 Translation

Translation with machine-readable dictionary maps words (terms) from one language to another by looking up a dictionary. The assumption of this method is that words are independent to each other as in the bag-of-word model. Lexilogos(http://www.lexilogos.com/) can be used as a resource of dictionary. Lexilogos contains several dictionaries and it solves the translation problem to some extend, but not sufficient to address all issues. One challenge is the ambiguity and synonymy. That is, a word T may be translated to multiple words that may not be directly related to T given the context [1]. Another challenge is the out-of-word phenomena. The dictionary fails to map a word from one language to another language. In our system, given a word in Ancient Greek, the lexilogos dictionary may find no mapping words in English. There are many reasons for the out-of-word phenomena. One is because of the lack of coverage of the dictionary. Another is because of the morphemes of Ancient Greek. Hence, we dot not choose machine-readable dictionary methods. Statistical translation is an extension of the machine-readable dictionary. Statistical translation maps a word T to a set of words with different probability. The probability can be either context dependent or independent. However, in order to learn the probability, a large amount of parallel corpus are needed. Previous work has studied data collected from Wikipedia [3] or other sources for the learning task. There are very few parallel corpuses in both English and Ancient Greek. Hence, we rule out statistical translation methods. Machine translation is a third option for the system implementation. Recent study on CLEF 09 'Ad-hoc track' [2] have shown that cross-language retrieval platform implemented with Google Translator can perform more than 90% as good as monolingual IR system, given popular languages. Although straightforward, it is hard to translate documents or queries in Ancient Greek directly to English by cutting-edge translation machines because Ancient Greek is not a popular language. However, we can use the translation machines to translate from Modern Greek to English, both of which are popular languages. In addition, in each retrieval unit, there is always a paragraph of Modern Greek as the translation of the original description in Ancient Greek. Thus, by the bridging of Modern Greek, we can use translation machine with a high precision.

2.2 Implementation and Demo

Figure 2 shows the query interface of the system. Four types of metadata can be searched together with the free text queries. These four types, i.e., author, book, fragment number and sources, are important on exploring the documents (retrieval units). Given an English query "Athenian War", the system searches the index and returns lists of results with order. Similarly, given an Ancient Greek query equivalent to Lucian, a list of documents containing this word are returned, as shown in Figure 2(b).

3. CONCLUSION AND FUTURE WORK

We introduce the design of a cross-language retrieval platform to support both English and Ancient Greek search. A document translation method is adopted. Metadata information is extracted to facilitate structural search. Our system has shortages as well. Given the two queries with the same meaning, but one in English and the other in Ancient Greek, the ranking of the results may be different. This is because we normalize the ranking score considering the length of documents. Since words in Ancient Greek and English are not mapped one-to-one, the length of a original document may be different from its translation. We plan to address this challenge in further study.

4. REFERENCES

[1] P. C. Carol Peters, Martin Braschler. *Multilingual Information Retrieval From Research To Practice*. Springer, 2012.

[2] N. Ferro and C. Peters. Clef 09 ad hoc track overview: Tel and persian tasks. In *CLEF (1)*, pages 13–35, 2009.

[3] M.-H. Li, V. Klyuev, and S.-H. Wu. Multilingual sentence alignment from wikipedia as multilingual comparable corpora. In *HC '10*, pages 167–171, 2010.