Translation approaches for Cross-Language Information Retrieval (CLIR)

Mirco Kocher, Student Number: 09-113-739
Student in Master of Science in Computer Science of the Universities of Bern, Neuchâtel and Fribourg

***Abstract*—This research paper presents and evaluates the issue of providing systems for CLIR. Different translation approaches are first elaborated and then compared. The two main studied issues elaborate what part a CLIR system should translate and how it should do it.**

I. Introduction

PEOPLE may write a query in one language and understand answers given in another. This is for instance when regarding very short text in Question and Answer format or just factual information for travel. Moreover, many documents contain non-textual information such as images, videos and statistics that can be understood regardless of the language involved and do not need translation.

Next to the two most common working languages in the European Union English and French there are 22 other official languages. While the EU encourages all its citizens to be able to speak two languages in addition to their mother tongue many are not bilingual [1]. Some can read documents written in another language but cannot formulate a query in that language. They cannot provide reliable search terms comparable to those found in the documents being searched. The challenge is "given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in differen media or languages appropriately identified." [2]

In this paper we are going to present the main factors that have to be considered when building information retrieval system that handles multiple languages. The rest of this paper is organized as follows. After a general introduction to classical information retrieval and the extension across languages the following part shows related works that we contemplated. Section five presents the methods and notations used. Chapter six provides the results that were achieved and in the conclusion we present our main findings.

II. Classical Information Retrieval

In a within-language retrieveal the implementation is essentially separated into two phases, namely, an indexing and a matching phase. Such a system first indexes the documents offline in advance and in the second step reacts online to the users query. The created index allows to look up features from the query and calculates a score for each matching document. This is faster than searching a large dataset at query execution time with a linear scan and results in an efficient system with effective retrieval. The system builds an inverted index (like a hash table) that allows efficient look-up of a feature and returns a list of all documents containing the given feature. The index is said to be inverted because each feature is associated with a pair containing a document number and the corresponding frequency that denotes how often this term occurs in this document. In the matching phase the system performs n look-ups for a query with n different features. All documents that are not included in the union of these n lists receive a score of 0 and won't be considered further. For all other documents, the similarity scores are calculated according to the number of features they contain. There are different algorithms to calculate a similarity score that are based on the idea that docuemts and queries are vectors in a high-dimensional space. The coefficients represent features with a weighting (like the *tf-idf*[1] or the cosine[2] model) and binary vector operations (like the inner product[3], dice[4] or cosine[5] formula) are used for the calculation of the similarity score. In the end the system returns a ranked list of documents with descending similarity scores.

III. Cross-Language Infromation Retrieval

To retrieve documents across languages, that is written in languages different from the language used for query formulation, the classic information retrieval mechanisms have to be extended by Cross-Lanugage Infromation Retrieval (CLIR) systems. This system manages a language mismatch between query and parts of the document collection where either:

- the document collection is monolingual, but the users can formulate queries in a different language.
- the document collection contains documents in multiple languages and users can query the entire collection in any language.
- the document collection contains documents with mixed-language content and users can query the entire collection in any language.

A Multilingual Infromation Retrieval (MLIR) system covers all the above cases plus the basic within-language retrieval.

[1] term frequency in the document multiplied with the inverse document frequency, that is the number of documnets containing this term

[2] the tf-idf normalized by its length

[3] the sum of the products of the corresponding entries of the two vectors

[4] two times the inner product divided by the sum of the length of each vector

[5] the inner product divided by the product of the root of the length of each vectors

The question is what to translate (such as the query or document only or a combination of both) and how to translate (either using machine-readable dictionaries, with machine translation or applying a statistical approach). There are four choices for crossing the language gap between query and documents:

1) translate the query into the language of the documents
2) translate the documents into the language of the query
3) translate both the query and the documents into an intermediary language
4) translate nothing

There are direct advantages and disadvantages to all options. With the second choice the whole corpus has to be translated which uses more storage space with each covered language and is a time-consuming process. With improving translation systems the whole document collection has to be preiodically re-translated to take advantage of these improvements. However the whole translation process can be shifted to the offline portion and avoids any speed penalty at retrieval time. Also the context of terms is available and helps the disambiguation of the words with multiple meanings. On the contrary in the first choice only the words in the query (which is usually short) are translated and avoids the storage problem. However, since user queries tend to be short and thus offer little context to handle ambiguous terms. The third choice can be used if there is no direct translation available or the quality is poor and the intermediate translation results in a better retrieval. For similar languages such as the Nordic languages (Danish, Swedish and Norwegian) the query might not need to be translated, based on the similar vocabulary and with a spelling correction algorithm one language can be considered as a mis-spelled form of another.

A. Problems

Before applying any translation method the text in question has to be preprocessed. In general the text is transformed to lowercase to improve matching regardless of the capitalization (for instance when the word is at the beginning of a sentence). Compound words that do not exist in the target language have to be segmentated and on the other hand tokens have to be compounded to represent a meaningful word. For example the German word ”Bundesbankpräsident” should be decoumpounded to ”Bund” + es + ”Bank” + ”Präsident” which is then translated to ”federal bank CEO”. Conversely in the Chinese word 中国人, the three logograms when segmeted mean ”middle”, ”kingdom” and ”people” which should be compounded and translated to ”Chinese” when translating to English [3].

Additionally the text is modified using a stemmer which conflates different token of the same word type. For instance the singular and plural form (like ”horse” and ”horses”) or different grammatical cases (such as the English noun ”Prague” in the Czech language where the dative form is ”Praze” and the genitive form is ”Prahy” are merged with the nominative form ”Praha”[6]).

[6]People sometimes use ”Praha” in English instead of Prague but mostly forget to decline it. It would obviously be ”to be in Praze” and ”go to Prahy”.

Sometimes a stopword list is applied to remove frequent and insignificant terms with the goal to reduce the size of the inverted file. Such a list may contain only one term (”the”) as in the WIN system (Thomson Reuters), nine terms (”an”, ”and”, ”by”, ”for”, ”from”, ”of”, ”the”, ”to”, ”with”) as suggested by the DIALOG system or may like the SMART system include 571 words (e.g. ”a”, ”all”, ”are”, ”is”, ”it”, ”just”, ”while”, ”who”, ”with”, ...). As a consequence, the length of a piece of text in the source language, and the length of its representation in the target language may differ widely.

There are some problems that come with the translation. One of the most prominent problem is the insufficient lexical coverage where some words have no translation such as for abbreviations or proper names. This can be countered by using a specialized thesauri with names of persons (”Gorbachev” in English and ”Gorbatschow” in German), arts (”Mona Lisa” in English is ”La Gioconda” in Italian) and cities (”Lisbon” in English is ”Lisboa” in Portuguese) and a dictionary for codes (”WHO” in English is ”OMS” in French). Nevertheless depending on the previously applied stopword list new problems might occur. The query ”vitamin A” is transformed to ”vitamin” when using the SMART system. In this case any document refering to any vitamin is retrieved even if it is about the vitamin C that is of no use for the querier. The same problems appear for the queries ”IT engineer” and ”WHO goals”. Assuming the systems applies a stopword list that preserves the word ”who” but uses a dictionary of codes. In this case the query ”Who won the Tour de France in 1995” could be translated to ”OMS gagné le tour de France en 1995”. Another prominent problem in all CLIR systems is the translation ambiguity. Each word-by-word translation using a machine-readable dictionary returns more possible expressions for each individual term. By simply using all available traslations, the number of terms in the destination language that is substituted for each term in the source language can vary widely. Assuming we use the Merriam-Webster Spanish Online dictionary to replace each word with all given translation alternatives. The Spanish query ”Contrabando de Material Radiactivo” is transformed (when ignoring the Spanish stopword ”de”) to ”smuggling, contraband; material, physical, real, equipment, gear; radioactive”. The resulting query now contains five different terms for the word ”Material” which will influence the document retrieval.

Something about variety in quality of the available resources...

IV. State of the Art

Most important previous work. Oard with document translation versus query translation. Yu with his comparable corpus from Wikipedia. Gollins and the lexical triangualtion. Savoy using Google machine translation.

Limit of current practices

V. Formalism

To compare different approaches we need clearly defined notations and a common baseline. The evaluation of a system is based on its ability to find and present relevant documents

that are appealing to the users. This effectiveness is measured using the precision and recall. Precision specifies the proportion of a retrieved set that is relevant while the recall indicates the proportion of all relevant documents in the collection that is included in the retrieved set.

TABLE I
A CONTINGENCY TABLE

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

In table I we have A+B+C+D as the collection size with A+C being the relevant documents and A+B representing the retrieved documents. We therefore define

$$Precision = \frac{A}{A+B} \tag{1}$$

and

$$Recall = \frac{A}{A+C}. \tag{2}$$

Since we have a ranked list for each query instead of a set the precision can be calculated at a fixed recall point (like precision at 20% recall) or a fixed rank cutoff (like precision at rank 20). Finally we only want a single number for effectiveness measure and we take the average precision. This is calculated by averaging precision when the recall increases which is the case at each new relevant retrieved document. As an example we have a total of 10 retrieved documents and we know that the documents at rank 2, 4, 5, 7 and 9 are relevant and the rest is irrelevant. For the document at rank 7 we have A = 4 and B = 3 which gives us using (1) a precision of 4 / (4 + 3) = 0.57. Following this calculation at each rank with a relevant document we calculate the average precision to be (0.5 + 0.5 + 0.6 + 0.57 + 0.55) / 5 = 0.544. A single query isn't enough to evaluate the effectiveness of a retrieval system. Therefore we introduce the mean average precision (MAP) which is the average of many queries' average precision values.

VI. Evaluation

Benchmarks. Baseline is monolingual IR.
Interpretation

VII. Conclusion

Recap main idea
Main results found
Improvements/applications
[4] [5] [6] [7] [8]

References

[1] TNS Opinion and European Commission. Europeans and their Languages. *Special Eurobarometer 386*, 2012.

[2] Doug Oard and David Hull. AAAI Symposium on Cross-Language IR. Stanford, Spring 1997.

[3] Jacques Savoy. Information Retrieval and the Internet: Beyond just English - CLIR. University of Neuchâtel, Spring 2013.

[4] Tim Gollins and Mark Sanderson. Improving Cross Language Information Retrieval with Triangulated Translation. In *SIGIR*, pages 90–95, 2001.

[5] Carol Peters, Martin Braschler, and Paul Clough. *Multilingual Information Retrieval - From Research To Practice*. Springer, 2012.

[6] Jacques Savoy and Ljiljana Dolamic. How effective is Google's translation service in search? *Commun. ACM*, 52(10):139–143, 2009.

[7] Kun Yu and Junichi Tsujii. Bilingual Dictionary Extraction from Wikipedia. *Proc. Of Machine translation Summit XII*, 2009.

[8] Douglas W. Oard and Paul G. Hackett. Document translation for cross-language text retrieval at the university of maryland. In *TREC*, pages 687–696, 1997.