Translation approaches for Cross-Language Information Retrieval (CLIR)

Mirco Kocher
Master Student at 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* a CLIR system should translate and *how* it should do it. In case document translation is possible the performance is at least as good as if instead the query would be translated. For a direct translation the machine translation services offer a good result. If a pivot language has to be used it is better to do multiple translations in parallel with different intermediate languages to filter out erroneous additional words and reduce ambiguity. A comparable corpus can be used to create a dictionary and by combining two similarity strategies the overall fitness improves.**

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 different 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.

Mirco Kocher has the student number 09-113-739

II. Classical Information Retrieval

In a within-language retrieval the implementation is essentially separated into two phases, namely, an indexing and a matching phase. Such a information retrieval (IR) 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 documents and queries are vectors in a high-dimensional space. The coefficients represent features with a weighting (like the *tf-idf*[1], Okapi[2], statistical language[3] or the cosine[4] model) and binary vector operations (like the inner product[5], dice[6] or cosine[7] 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. More weighting formulas and feature representations with further details on the topic can be found in the book by Peters et alii [3].

III. Cross-Language Information Retrieval

The classical information retrieval systems are not applicable for the case where the documents are written in languages

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

[2] a more complex probabilistic model in which a term's weight also depends on its discrimination power and on document length

[3] tries to estimate the occurrence probability of terms

[4] the tf-idf normalized by its length

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

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

[7] the inner product is divided by the product of the root of the length of each vector

different from the language used for query formulation. To retrieve documents across languages the classic IR mechanisms have to be extended by Cross-Language Information 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 Information Retrieval (MLIR) system covers all the above cases plus the basic within-language retrieval. 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 periodically 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 misspelled form of another [3].

A. Problems introduced by translation

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 segmented 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 segmented mean ”middle”, ”kingdom” and ”people” which should be compounded and translated to ”Chinese” when translating to English [4].

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”[8]).

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 referring 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 translations, 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

[8]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”.

query now contains five different terms for the word "Material" which will influence the document retrieval.

IV. Related Work

Many previous works in this domain focused on how to translate queries to improve CLIR. Yu and Tsujii extracted a bilingual dictionary from Wikipedia and was able to collect robust and large-scale comparable corpora [5]. He also combined context heterogeneity similarity (terms around a domain specific word are similar to that of its translation in another language) and dependency heterogeneity similarity (a word and its translation share similar modifiers and head) which outperforms both the individual approaches. Gollins and Sanderson translated in parallel across multiple intermediate languages and fused the end results which raised the effectiveness of the system [6]. Depending on the size of the language resourced the lexical triangulation approach to transitive translation may even outperform the direct translation but there is no single best merge strategy for all environments. Savoy and Dolamic used the machine translation tools provided by Google[9], BabelFish[10] or Promt[11] and addressed the question to what extent they can produce adequate results [7]. Independent from the service used the retrieval performance is clearly lower than in a monolingual search especially for queries containing concepts expressed in an ambiguous way or vocabulary that leads to incorrect identification of relevant and non-relevant items.

The following works focused on what to translate to improve CLIR. Oard and Hackett showed that document translation can be approximately as effective as approaches based on query translation and may ultimately be more effective for some application [8]. With the moderately large corpus he noted that the performance gain with document translation depends on the topic but in general appears to perform at least as well as query translation. McCarley [9] ran comparable experiments and came to the same results as Oard and Hackett. He extended the testing with a hybrid system that uses both query translation and document translation which finally produces superior performance than either direction alone.

Fujita presented the limits of CLIR effectiveness and conditions to further improve the results [10]. While the query translation quality should be ideally perfect techniques like the pre-translation query expansion may improve the effectiveness further but also helps to compensate for lost information in the translation. In this paper we compare the results achieved by those previous works according to the relative improvement over monolingual baseline.

V. Methods for comparison

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.

[9]http://translate.google.com/
[10]http://babelfish.com/
[11]http://translation2.paralink.com/

The monolingual retrieval results provide a useful baseline for evaluating cross-language retrieval performance. In this case the untranslated document collection and the query in the same language is used and considered as the practical limit. In the cases where CLIR effectiveness is higher than the monolingual one some test set specific knowledge are available that gives the system with more information about relevance than monolingual topic description [11].

The 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 can therefore define the

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

and the

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

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 as our main metric the mean average precision (MAP) which is the average of many queries' average precision values.

We can also look for a value that closely reflects the expectation of those queriers who are searching for a single good response to their request. The mean reciprocal rank (MRR) average value of the inverse of the rank for the first relevant document. The MRR varies between 1 (the first retrieved item is relevant) and 0 (no correct response up to a given threshold). This value serves as a measure for the ability to extract one correct answer and list it among the top ranked items.

VI. Evaluation

In this evaluation we will interpret the two main studied issues about *what* a CLIR system should translate and *how* is should do it. First we focus on the *what* to translate by answering the questions about whether document or query translation is better according to the results achieved by Oard and Hackett. Then we analyze the *how* to translate by comparing the approaches of the bilingual dictionary extraction by Yu and Tsujii, the lexical triangulation of Gollins and Sanderson and finally the performance of machine translation services examined by Savoy and Dolamic.

What to translate

As presented in section III there are four choices to cross the language gap between query and documents. Obviously the approach of translating nothing is generally not suitable or applicable. By using multiple translation stages the errors that emerge by translation as shown in subsection III-A will stack up resulting in suboptimal performance Therefore there should ideally be only a single translation of either the query or the document collection as presented in the following subsection.

A. Query or document translation

In the work of Oard and Hackett [8] queries of variable length and from different topics are used. They are evaluated in three different scenarios. First as a baseline the monolingual retrieval of German queries in the German SDA[12]/NZZ[13] document collection. Then the retrieval is performed with query translation, meaning the English queries are translated to the German language and used in the same untranslated document collection. Finally the retrieval is tested with document translation, which means that the same English queries are used but the documents are translated to the English language.

When only using short queries the difference between the document translation and query translation is not significant and they preform about equally well. If we look for instance at the precision at 30% recall we get 1/3 precision for the monolingual retrieval and about 9/40 precision for both translation strategies. This also means that CLIR is in this case 5/8 as effective as classical IR. When increasing the recall level the gap diminishes until at 70% recall they both result in the same performance. A notable difference in translation strategy can be detected when using longer queries. The monolingual retrieval still achieves 1/3 precision at 30% recall. At the same level the document translation approaches a precision of 1/4 but the query translation only gets 1/5 precision. The difference in precision between document translation and monolingual already vanishes at 40% recall while the difference between all three tested systems disappears not until the 60% recall level. Some advantage for document translation is apparent for long queries. Looking at the gain in average precision that emerges from using document translation rather than query translation on a query-by-query basis shows a similar result. Even though for some topics the average precision decreases in most cases there is a positive gain. It does appear that document translation is performing at least as well as query translation and both approaches are performing creditably according to tests by Oard and Hackett [8] and also McCarley [9] who ran similar experiments. The results for short and long queries range between 67% and 90% of monolingual average precision on the SDA/NZZ collection.

[12]Swiss news press agency - Schweizerische Depeschenagentur AG

[13]New Journal of Zurich - Neue Zürcher Zeitung

How to translate

Despite the traditional approach of translating term wise using a dictionary the following subsections examine other methods. This is useful for instance if the resources needed for the translation is not available. A dictionary can in this case be collected from comparable corpora and using a triangulation with multiple languages helps in the elimination of wrong translations. The machine translation services provide another way to cross the language gap.

A. Bilingual dictionary extraction

Approaches from the natural language processing domain can be used to extract bilingual dictionaries from comparable corpora. With the inter-language links in Wikipedia Yu and Tsujii [5] builds a comparable corpora and then uses context and dependency heterogeneity similarity to extract such a lexicon. The large size ensures the quantity of the comparable corpora and the manually created links by the article authors guarantees the quality of the collected corpora. The context based strategy observes that a term and its translation appear in similar lexical contexts. The dependency based strategy not only uses the words around translation candidates but utilizes the syntactic analysis of comparable corpora to recognize the meaning of translation candidates. The lexical information comes from the entire sentence instead of just a small frame. Finally the proposed solution is to combine the context heterogeneity similarity and dependency heterogeneity similarity appropriately. The experiment first uses 500 Chinese-English pairs that are randomly selected from the collected pages and equally divided into a testing and development group. In a second experiment 15 Chinese-English pairs are randomly selected from a different independent lexicon to test the effectiveness of the proposed approach in real bilingual dictionary extraction.

Comparing both heterogeneity similarity strategies on the first translation pairs the dependency based method reached a MRR of 0.112 while the context based approach was with 0.053 over 50% lower. The proposed solution that combines the two approaches out performs the individual scores with a MRR of 0.125. This is a 10% improvement to the dependency approach and aver 130% better than when only using the context. In the second experiment with real world data the performance is generally lower. The individual approaches reached a mean reciprocal rank of 0.079 and 0.078, again with the advantage for the context approach. With a MRR of 0.097 the combined solution is better by 1/4 compared to the other two approaches. Yu and Tsujii did not test this extracted dictionary with a retrieval and a comparison to a monolingual

baseline. Nevertheless the improvement over current practices in bilingual dictionary extraction can be seen.

B. Lexical triangulation

While most approaches to CLIR assume that a direct translation between the query and the document exist Gollins and Sanderson [6] researches the situation where this assumption does not hold. Using a pivot language introduced errors due to the additional step of transitive translations. By using lexical triangulation, which combines translations from two different transitive routes, the errors are reduced and the performance improves. As a test Gollins and Sanderson took German queries and a document collection in the English language and first used the two pivot languages Spanish and Dutch and later included Italian as well. If the query was the single German word "fisch", a Spanish suggestion is the two terms "pez, pescado" from where the English translation "pitch, fish, tar, food fish" follows. On the Dutch route we first get "vis" and finally "pisces the fishes, pisces, fish". For the strict merge process only the intersection of the two transitive translation is taken, which is "fish" in this example and representing a good unambiguous translation of the original German word. In the liberal merge again the common translations are preferred but in the absence of a common translation the terms from both routes were used. If no translation is available then the original German term passed unchanged.

The mean average precision for all runs in case of the monolingual baseline was 0.289. Then the direct translation scored 0.0549 precision which makes it 81% below the monolingual retrieval. This difference in performance can be explained by the poor vocabulary coverage and the inability tho choose the most common sense of a word. All three transitive translation via the Spanish, Dutch and Italian pivot languages scored with a precision of 0.0106, 0.0044 and 0.0026 respectively considerably worse. The triangulation of Spanish and Dutch improves the performance to a precision of 0.0436 in the strict merge process and 0.0403 in the liberal merging. Each intermediate language adds some noise by adding erroneous additional words and therefore introduces ambiguity to the translation. Comparing different routes serves as a noise cancellation and preserves only the correctly translated terms. When triangulating all three pivot languages the precision raised to 0.0558 when using liberal merging but fell to 0.038 in the strict merge process. The improvement over direct translation of the former is not statistically significant. Even though transitive translation introduces more ambiguity the performance might increase when combining evidence from several transitive translations [12]. The results for the German queries are about 20% of monolingual average precision when using transitive translation to match the English document collection. Compared to the direct translation the triangulations have depending on the pivot languages and merge process an average precision that ranges between 50% and 100%.

C. Machine translation

Without directly evaluating the translation services provided by online machine translation services Savoy and Dolamic [7] tested and analyzed various systems in term of their ability to retrieve items automatically based on a translated queries. The document collection contains French articles published by the newspaper Le Monde and the SDA. As a baseline French queries of almost 300 topics were used for a monolingual retrieval. Then English language topics were translated from English to French for the evaluation. Further exploratory works was to compare short queries containing only the title of the topic and long queries including the title and the description of the corresponding topic. The three models *tf-idf*, Okapi and language model were used and tested with respect to the MRR and MAP. Next to Google the translation services by BabelFish and Promt was evaluated.

In any case the Okapi model performed better than the other two IR models. First when looking at Google's translation service with short queries the mean average precision using the Okapi model in the monolingual retrieval was 40% with a MRR of 2/3. With a MAP of only 1/4 and a mean reciprocal rank of 51% the *tf-idf* model was significantly worse. Compared to the first baselines the retrieval using the translation service scored a precision of 0.341 and a MRR of 0.582 which is a relative difference of -15% and -12% respectively. With the lower baseline the *tf-idf* model also got worse results for when the translation service was used. The performance in MRR decreased by 21% to 0.39 and the MAP fell by 25% to 1/5. Looking at Google, BabelFish and Promt when using the long queries that consist of the additional topic description the performance increased. Assuming the Okapi IR model is used the MRR for the monolingual retrieval increased to 0.736 which is a relative difference of 11%. The MRR of Google, BabelFish and Promt with short queries is with $\pm 1\%$ the same and not significantly different. Similarly the long queries improved the performance of all systems by 13% $\pm 0.5\%$. The results for short and long queries depending on the used IR model range between 75% and 90% of the monolingual reciprocal rank on this document collection. Clearly, in mean, a translated query may retrieve the needed information. For some queries the search systems encountered difficulties to find at least one relevant answer. In the monolingual run 30 short queries retrieved no relevant item in the 20 documents with the highest scores. When adding a translation stage the topics may become more ambiguous or use vocabulary that leads to incorrect identification of relevant items. With English topics the translation systems Google, BabelFish and Prompt increased the number of queries without a relevant document to 60, 64 and 56, respectively.

VII. Conclusion

In this paper we studied the concept of CLIR then elaborated and compared different translation approaches. Depending on the languages, the size of the corpus and the available resources the best approach among the existing ones based on the different studies is different.

Experiments were performed to compare query and document translation based CLIR systems. The machine readable dictionaries and semantic rules were used identically for both translation approaches. Document translation is the less

common approach as translation is frequently computationally expensive and the document sets are usually large. Some indications were observed where document translation may be more effective than query translation for some applications. This occurs since a long document offers the translation engine more opportunities to translate key words and phrases appropriately. But in general no clear advantage for either the query translation system or the document translation system was found. The latter approach might be a practical method on moderately large collections. A future work extending the current state might experiment with a hybrid system incorporating both translations. In this case two retrieval systems would be run parallel and the results could be merged to the arithmetic mean of the retrieval scores.

The usage of a comparable corpora like Wikipedia to extract a dictionary is a common approach.

Recap main idea.

Main results found

Improvements/applications

REFERENCES

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

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

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

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

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

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

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

[9] J. Scott McCarley. Should we Translate the Documents or the Queries in Cross-language Information Retrieval? In *ACL*, 1999.

[10] Sumio Fujita. Notes on the Limits of CLIR Effectiveness NTCIR-2 Evaluation Experiments at Justsystem. In *Proceeding of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pages 181–188, 2001.

[11] Jinxi Xu and Ralph M. Weischedel. TREC-9 Cross-lingual Retrieval at BBN. In *TREC*, 2000.

[12] Lisa A. Ballesteros. Cross-Language Retrieval via Transitive Translation. In *Advances in Information Retrieval*, pages 203–234. Springer, 2000.