Selecting Automatically the Best Query Translations

Pierre-Yves Berger & Jacques Savoy
Computer Science Dept.
University of Neuchatel
2009 Neuchatel, Switzerland
{Pierre-Yves.Berger, Jacques.Savoy}@unine.ch

Abstract

In order to search corpora written in two or more languages, the simplest and most efficient approach is to translate the query submitted into the required language(s). To achieve this goal, we developed an IR model based on translation tools freely available on the Web (bilingual machine-readable dictionaries, machine translation systems). When comparing the retrieval effectiveness of manually and automatically translated queries, we found that manual translation outperformed machine-based approaches, yet performance differences varied from one language to the text. Moreover, when analyzing query-by-query performances, we found that query performances based on machine-based translations varied a great deal. We then wondered whether or not we could predict the retrieval performance of a translated query and apply this knowledge to select the best translation(s). To do so we designed and evaluated a predictive system based on logistic regression and then used it to select the top most appropriate machine-based translations. Using a set of 99 queries and a documents collection available in the German and Spanish languages (extracted from the CLEF-2001 and 2002 test suites), we show that the retrieval performance of the suggested query translation selection procedure is statistically better than the single best MT system, but still inferior to the retrieval performances resulting from manual translations.

1. Introduction

Given the increasing information on the Internet, written in different languages, effective access to these information sources is becoming more important. Accompanying this is a growing need for search systems able to cross the various barriers encountered when accessing different language sources. In a multilingual country such as Switzerland or in Europe in general, people search the multilingual Web in a variety of topic areas (law, economy, culture, education, leisure, and travel). Large international organizations or companies (e.g., Novartis, WTO, and European Parliament) also have similar search requirements. While some users would be perfectly bilingual, others may only be able read documents written in another language but would not be able to formulate a query in that language or, at the very least, provide reliable search terms in a form comparable to those found in the documents being searched (Oard & Resnik 1999; Petrelli *et al*. 2004). In other cases, monolingual users may want to retrieve documents in another language and then automatically or manually translate the retrieved texts into their own language. Finally, documents in other languages may contain information in non-textual formats such as pictures, images, graphics or even statistics. Reinforcing these examples is the fact that in a global context, information produced in a given location is immediately available worldwide.

This paper presents a query-by-query translation selection strategy that is intended to promote effective bilingual searches. We do not propose a new translation tool *per se* but we do suggest the use of existing and freely available translation resources. In order to evaluation our proposed approach, we use a relatively large set of documents written in the German or Spanish languages. The rest of this paper is organized as follows. Section 2 presents the indexing and search strategy used in our experiments while Section 3 describes the main features of the test collections and evaluations carried out in a monolingual context. Section 4 lists and evaluates various translation tools freely available on the Web for both the German and Spanish languages. Section 5 presents and evaluates our query translation selection strategy based on the logistic regression technique developed to select the best translation(s).

2. Indexing and Searching Strategies

In order to retrieve pertinent items that might respond to a user search request, we must first represent the corresponding documents and queries. This is achieved by first extracting a set of indexing terms in order to build a document (or query) surrogate. Usually this step does not require a lot of processing, except for the Chinese or Japanese languages where word boundaries are not explicitly marked, for which a segmentation procedure must be used to find the corresponding words. Second, terms appearing in a stopword list are ignored. Third, the suffixes required by the flexion rules (e.g., plural form) or by the derivational morphology are usually removed by a stemming procedure. Finally, a weight w_{ij} reflecting the importance of each term t_j in describing the semantic content of the document d_i is computed, and is based mainly on the following components:

- term frequency of term t_j within document d_i, denoted by tf_{ij};
- document frequency (denoted df_j) corresponding to number of documents indexed by term t_j. This factor is usually included in the idf_j value (with $idf_j = \log [n/df_j]$, where n indicates the number of documents in the corpus);
- document length (*ceteris paribus*, it seems appropriate to attribute more weight to smaller documents than to longer ones tending to deal with more than one topic).

As a first IR model, we may use the classical $tf \cdot idf$ vector-space model within which the indexing strategy is based on both term frequency and document frequency. Moreover, each indexing weight may vary only within the range of 0 to 1 through using cosine normalization as defined by the following equation.

$$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}} \qquad (1)$$

In addition to this vector-space approach, we also considered the Okapi probabilistic model (Robertson *et al.*, 2000), which in several tracks during the last CLEF evaluation campaign usually resulted in one of the highest performance levels (Peters *et al.*, 2006). In this IR model the indexing weight is defined according to the following formula:

$$w_{ij} = \left((k_1+1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right) \quad \text{with } K = k_1 \cdot \left[(1-b) + b \cdot {ld_i}/{avdl}\right] \qquad (2)$$

where K represents the ratio between the length of d_i measured by ld_i (or the sum of its tf_{ij}) and the collection mean is represented by *avdl*. The constants b and k_1 were fixed respectively at 0.5 and 1.5 for all corpora, while *avdl* was set according to the values shown in Table 1.

When indexing documents (or queries) written in the German or Spanish languages, we replaced diacritic characters by their corresponding non-accentuated letter. For these languages, we used a freely available stopword list and most common suffixes were removed using a simple stemmer (see www.unine.ch/info/clef/). Finally for the German language, we applied an automatic decompounding procedure (Savoy, 2004b). In the experiments reported in this paper, both compounds (e.g., “Atomtests” or “Bundesbankpräsident”) and their composite parts (e.g., “Atom,” and “Test” or “Bund,” “Bank,” and “Präsident”) were left in documents and queries. Thus both forms were used during the indexing of documents and queries. As reported by Braschler & Ripplinger (2004), decompounding German words may significantly improve retrieval performance.

3. Test Collections and Evaluation Methodology

To evaluate our new query translation selection strategy, we merged the CLEF 2001 (Peters *et al*., 2002) and CLEF 2002 (Peters *et al*., 2003) test suites in order to obtain a larger number of queries. From these text collections, we selected only the German and Spanish corpora, made up of newspaper collections such as the *Der Spiegel* (1994-1995, German) and *Frankfurter Rundschau* (1994, German), together with various articles edited by news agencies such as *EFE* (1994, Spanish) and the Swiss news agency (1994, German). The documents included in these corpora cover news events that occurred mainly in 1994.

	German	Spanish
Size (in MB)	527 MB	509 MB
Number of documents	225,371	215,738
Number of distinct indexing terms / document		
Mean	119.1	111.8
Standard deviation	109.7	55.4
Median	89	99
Number of queries	99	99
Number rel. items	4,068	5,548
Mean rel./ query	41.1	56.0
Maximum	212 (Q# 42)	321 (Q# 95)
Minimum	1 (Q# 64)	1 (Q# 64)

Table 1: Various statistics on the German and Spanish corpora

As shown in Table 1, both corpora are of comparable size. On the other hand, the mean number of distinct indexing terms per document is also relatively similar across the corpora (around 120 for the German corpus and 112 for the Spanish). Table 1 shows that when examining the number of relevant documents per query, the mean number varies across the languages, from 41.1 for the German corpus to 56.0 for the Spanish, with the latter also containing the larger number of relevant items (5,548).

An inspection of the available topics reveals they reflect a diversity of information needs (such as “European single currency,” “Civil War in Afghanistan” or “The Clementine space probe”) rather than being limited to a narrow subject range. Following the TREC model, each topic was structured into three logical sections, namely a brief title (denoted briefly by the letter T), a one-sentence description (denoted by the letter D) and a narrative part specifying the relevance assessment criteria (denoted by N). The queries used in our experiments are based only on the TD topic formulation, the official query formulation employed during the CLEF 2001 and 2002 evaluation campaigns.

To measure retrieval performance, we adopted the mean average precision (MAP) as computed by `trec_eval`. To determine whether or not a search strategy might be better than another, we applied a statistical test. More precisely, we stated the null hypothesis (denoted H_0) specifying that both retrieval schemes achieved similar performance levels (MAP), and this hypothesis would be rejected at a 95% confidence level (two-tailed test). Such a null hypothesis plays the role of a devil's advocate, and this assumption will be accepted if two retrieval schemes return statistically similar MAPs, and rejected if not. Thus, in the tables found in this paper we underline statistically significant differences based on a two-sided non-parametric bootstrap test (Savoy, 1997). However, a decision to accept H_0 is not equivalent to the opinion that the null hypothesis H_0 is true, but rather it represents the fact that "H_0 is not shown to be false," resulting in insufficient evidence against H_0.

	Mean average precision	
Language IR model	German 99 queries	Spanish 99 queries
Okapi	**0.3975**	**0.5478**
tf · idf	<u>0.2917</u> (-25.1%)	<u>0.3566</u> (-34.9%)

Table 2: Evaluation of the German and Spanish collections (monolingual IR, TD queries)

As a first experiment, we used the topics written in the German and Spanish languages to search into the corresponding documents collection (monolingual IR). The resulting MAP depicted in Table 2 indicates that the Okapi search model proposes the best performance. Moreover, the performance differences with the classical *tf · idf* model are statistically significant.

4. Automatic Query Translation

In the previous section, the topics and the documents were written in the same language. In this section we address the problem of bilingual information retrieval, where the topics are expressed in one language in and the documents retrieved are written in another language. In this particular study, queries are written in English and the retrieved documents are from the German or Spanish collections. Rather than proposing a new translation system, we have based our bilingual search model on existing translation tools freely available on the Web. An overview of various translation strategies used to cross the languages barrier is presented in Section 4.1. An evaluation of several translation resources using the Okapi probabilistic model is reported in Section 4.2.

4.1. Translation Strategies

In order to cross the language barriers, for the most part two general translation strategies have been suggested. First we might translate all documents into the topic language (usually in English) and then the user might submit his/her request in this specified language (see for example Braschler *et al.*, 2002). Second, we might build an index for each language and upon receiving the user's request the system would then translate it into the targeted language, before performing a monolingual search in this language (Chen & Gey, 2004; McNamee & Mayfield, 2004; Savoy, 2004a; 2005).

As an alternative we might propose mixed strategies based on both documents and queries translation (Braschler, 2004; Chen & Gey, 2004), but the implementation of these translation schemes would be more complex and require more computer resources. The required document or query translation could be obtained using several translation resources (Grefenstette, 1998; Peters *et al.*, 2006) such as bilingual or multilingual machine-readable dictionary (MRD) as well as multilingual thesauri (e.g. EuroWordNet), machine translation

(MT) systems or probabilistic translation models trained on large amounts of parallel corpora (Gao & Nie, 2006).

Finally, some authors have suggested “crossing” the language barriers without using a specific translation tool, viewing one text written in one language as misspelled expressions of the other. For example, Buckley *et al.* (1998) suggest that we could view English as misspelled French, and then apply adjunct rules to help the system to correctly spell English documents in French.

In our experiments, we decided that our translation strategy would be based on query translation, thus requiring fewer computer resources than document translation. Moreover, the query translation approach would be clearly much simpler to implement, especially when faced with dynamic document collections (those in which new documents are included, others removed or modified).

4.2. Query Translation Evaluation

In this study, we based our query translation strategy on translation resources freely available on the Web. More precisely, we studied MT systems that would automatically provide a complete translation of a given query into the desired target language, and also bilingual MRD able to provide one or more translation alternatives for each search keyword. Listed below are the five MT systems we consulted, and also a bilingual dictionary (Babylon):

REVERSO (PROMT)	webtranslation.paralink.com/
BABELFISH	babel.altavista.com/
GOOGLE	www.google.com/language_tools
FREETRANSLATION	www.freetranslation.com
TRANSLATIONEXPERTS	www.tranexp.com
BABYLON	www.babylon.com

For the Babylon bilingual dictionary we submitted search keywords to be translated word-by-word. In response, the Babylon system provided not only one but several translations in an unknown order. In our experiments, we consistently selected the first available translation (labeled “Babylon 1”), the first two terms (labeled “Babylon 2”) or the first three available translations (labeled “Babylon 3”).

Table 3 depicts the MAP obtained by the Okapi probabilistic model when handling the German and Spanish languages. In this case, the original query was submitted in English and then automatically translated with the eight different strategies. In the first row we report the performance achieved by manually translated queries, and then we use this performance level as a baseline. As can be seen, the human translation differences for both languages were statistically significant.

An examination of the machine-based query translation strategies shows that for German and Spanish the Reverso system was the best approach. In Spanish, the best translation tool clearly provided better performance than did the other translation resources. In the German collection however the performance difference between the best translation tool (Reverso in this case) and the second best (Google) was rather small (0.3074 vs. 0.2999, -2.4% in relative effectiveness).

	MAP (% change)	
Language Translation tool	German 99 queries	Spanish 99 queries
Okapi	0.3975	0.5478
Reverso	**0.3074** (-22.7%)	**0.4613** (-15.8%)
BabelFish	0.2971 (-25.3%)	0.4046 (-26.1%)
Google	0.2999 (-24.6%)	0.4033 (-26.4%)
FreeTranslation	0.2651 (-33.3%)	0.4098 (-25.2%)
TranslationExpert	0.2223 (-44.1%)	0.3821 (-30.2%)
Babylon 1	0.2715 (-31.7%)	0.3897 (-28.9%)
Babylon 2	0.2674 (-32.7%)	0.3437 (-37.3%)
Babylon 3	0.2552 (-35.8%)	0.3234 (-41.0%)
Best	0.4040 (+1.6%)	0.5293 (-3.4%)

Table 3: Mean average precision of various translation approaches (Okapi model, TD queries)

The last line of Table 3 labeled “Best” shows the MAP results achieved when always selecting the best available translation on a per query basis. This must be viewed as a theoretical upper bound, based on an oracle that always selects, without any error, the best translation for each query. These values indicate that for Spanish, the performance level achieved is close although lower than that obtained by manually translated queries, and the performance differences are not statistically significant. For the German corpus, the “Best” translation scheme resulted in better MAP than did the manually translated queries. However, the performance difference cannot be viewed as statistically significant.

	Number of queries	
Language Translation tool	German 99 queries	Spanish 99 queries
Reverso	22 / **17**	**29** / **20**
BabelFish	**23** / 10	20 / 1
Google	**23** / 9	22 / 2
FreeTranslation	16 / 13	23 / 16
TranslationExperts	9 / 7	16 / 12
Babylon 1	11 / 9	16 / 13
Babylon 2	12 / 10	7 / 6
Babylon 3	11 / 9	6 / 5

Table 4: Number of queries showing best performance (Okapi model, TD queries)

Using mean values to represented general statistics may hide performance irregularities among queries. Table 4 depicts the number of individual queries providing the best retrieval effectiveness, for both languages and each query translation resource, based on our query sets. For the German collection Google provided the best translation effectiveness for 23 queries out of a total of 99, while for the Spanish corpus this same translation tool produced 22 queries having the best performance. However, these values do not fully reflect the query translation quality because more than one translation system may provide the same response and thus for a given query, more than one translation device may be viewed as the best one.

In Table 4, we added a second figure to indicate the number of *unique* queries having depicted this best query translation for the corresponding translation device. For example with the German collection and for nine queries, Google was the only tool to provide the best translation. In fact for the German language, both the Google and BabelFish tools often produced very similar translations. Although their results were not identical their retrieval performances were very similar. Generally, the data shown in Table 4 demonstrates that none of the automatic query translation tools show a clear and definitive advantage over the other translation schemes. If we consider the worst translation resource, Table 4 indicates that this translation tool provides the best translation for seven queries (German language, TranslationExperts) or five queries (Spanish language, Babylon 3).

This translation variability is also encountered when studying various manual translations, as noted by D. Knuth concerning a study of various Bible translations:

> "Well, my first surprise was that there is a tremendous variability between the different translations. I was expecting the translations do differ here and there, but I thought that the essential meaning and syntax of the original language would come through rather directly into English. On the contrary, I almost never found a close match between one translation and another. ... The other thing that I noticed, almost immediately when I had only looked at a few of the 3:16s, was that no translation was consistently the best. Each translation I looked at seemed to have its good moments and its bad moments." (Knuth, 2001, p. 56)

In order to illustrate this problem, we extracted two sample topics from our test-collection. As listed in Tables 5a and 5b, we only indicated the title and descriptive sections of the topic formulation together with their manual translation into German. Given that Reverso is commonly considered a good translation system (see Table 3), we selected two cases in which it failed to provide the best answer. For Topic #42 depicted in Table 5a, for example, Reverso translates the acronym "UN" (or "US") into its extended form in German, namely "Vereinten Nationen" (and "Vereinigte Staaten"). The morphological variations attached to the adjectives or nouns (ending in '-en' or '-er') will be removed by the German stemmer and did not seem to hurt the MAP for the underlying translated query. In German newspapers however the acronyms "UN" or "US" occur more frequently and cause more pertinent documents to be retrieved. Thus, as this example illustrates, knowing that two languages use the same acronyms could make the translation between two closely related languages more simple (and effective). This is not always the case however, as for example in French the acronym for UN is ONU. An inspection of the translation of Topic #42 produced by the two MT systems (last lines of Table 5a) reveals that both sentences are not grammatically correct and contain incorrect spellings (e.g. "fund"). Moreover, the answer proposed by the Reverso system ("Finden Sie Dokumente" instead of "Fund dokumentiert") contains fewer grammatical errors than the sentence produced by FreeTranslation. In this latter case the acronym "UN" is preserved and this has a positive impact on retrieval effectiveness (MAP of 0.7214 instead of MAP of 0.4007). Furthermore, the MT system encounters a problem with the acronym "US", recognizing it as a pronoun (we). The system thus translates it into its German form ("uns"), and writes it with uppercase letters ("UNS"). Clearly, the presence of a word in uppercase does not indicate that the corresponding form could be an acronym.

Translation tool	Topic description	MAP
English (original)	<TITLE> U.N./US Invasion of Haiti. <DESC> Find documents on the invasion of Haiti by the U.N./US soldiers	
German (original)	<TITLE> UN/US-Invasion Haitis <DESC> UNO/USA entsenden Truppen nach Haiti.	
Reverso	<TITLE> Invasion der Vereinter Nationen Vereinigter Staaten Haitis. <DESC> Finden Sie Dokumente auf der Invasion Haitis durch Vereinte Nationen Vereinte Nationen Vereinigte Staaten Soldaten	0.4007
FreeTranslation	<TITLE> U N/UNS Invasion von Haiti. <DESC> Fund dokumentiert auf des Invasion von Haiti durch U N UNS Soldaten.	0.7214

Table 5a: Examples of Topics #42 (in English and its German translations)

In a second example, for Topic #93 as reported in Table 5b, are the original versions in English and its manually translated form in the German language. For this topic, the Google MT system provided a better translation and a better MAP (0.5919) than did the Reverso system (MAP: 0.0476). An inspection of the corresponding translations shows that the word "Eurofighter" is preserved by Google's while the Reverso system suggests translating the word "fighter" into its German equivalent ("Kämpfer"). In actual fact German newspaper articles do not usually translate foreign names (such as "Gorbachev," "Computer," "Chat" or "Awards") into German. These terms are usually kept in their original form (such as "Eurofighter," which remains as is).

Translation tool	Topic description	MAP
English (original)	<TITLE> Eurofighter <DESC> Find documents which report about the "Hunter 90" or the "Eurofighter" project.	
German (original)	<TITLE> Eurofighter <DESC> Finde Dokumente, die über das Projekt "Jäger 90" bzw. den "Eurofighter" berichten.	
Reverso	<TITLE> Eurokämpfer <DESC> Finden Sie Dokumente welche über den Jäger 90 oder das Eurokämpfer Projekt.	0.0165
Google	<TITLE> Eurofighter <DESC> Finden Sie Dokumente die über das Jäger 90 oder Projekt Eurofighter.	0.7859

Table 5b: Examples of Topics #93 (in English and its German translations)

5. Automatic Query Translation Selection

In order to overcome the inconsistencies in translation effectiveness described previously, various researchers suggest combining two or more query translation sources. For example, by concatenating query translations provided by different MT systems (Kwok *et al.*, 2001; Chen & Gey, 2004) or more generally by using various translation tools (Braschler, 2004; Savoy, 2004a; 2005).

The approach we propose for this paper is different. Instead of always combining the same two or more translation tools independently of the query, we tried to automatically identify the best single or multiple query translations on a per query basis. This type of translation-based selection results from an analysis of the characteristics and statistical properties of the translated query, and hopefully should allow us to verify whether or not the given query translation could be viewed as good or bad.

To achieve this goal, our system sent the English request to various translation tools (see Section 4) in order to obtain a set of translated queries in the target language. In a second step, for each query translation we computed a translation effectiveness estimate. This estimate was not based directly on a human assessment, as for example the assumption that the most fluent translation would be the best one (Callison-Burch & Flournoy, 2001). In our approach, we based our selection procedure on the query's ability to retrieve relevant documents. In a similar but not identical vein, different works suggest analyzing the statistical properties of the current query before any subsequent processing. For example, Savoy *et al.* (1996) proposed automatically selecting the most effective search engine based on the *k*-nearest neighbor model. Within the robust TREC 2003 track, Kwok *et al.* (2003) suggested a prediction model that could decide whether or not a blind query expansion might improve retrieval effectiveness. In a related paper, Cronen-Townsend *et al*. (2002) showed that in monolingual IR the *idf* average value of a query could be a good prediction of its retrieval effectiveness (other examples can be found in (Yom-Tov *et al.*, 2005; Voorhees, 2006)).

In this paper, the selection procedure identifying the best query translations was based on logistic regression (Hosmer & Lemeshow, 2000). This statistical approach allowed us to estimate the probability of a binary outcome variable, according to a set of explanatory variables. This type of predictive model has been proposed in different contexts, such as an informetric phenomena (Bookstein *et al.*, 1992), as an IR model (Gey, 1994), as a merging strategy for meta-search engines (Le Calvé & Savoy, 2000) or as a multilingual merging scheme (Savoy, 2004a; 2005).

In order to obtain a probability estimate using the logistic regression model, we needed to define a set of explanatory variables for each query translation alternative. To achieve this goal, the name of the translation device that produced the translation (variable denoted *source*) was taken into account, together with the number of query terms (variable denoted *concepts*). Moreover, for each translated search term, we computed its occurrence frequency and the corresponding *idf* value. Finally, based on a predefined list of terms, we would identify (with errors) the corresponding term as a proper noun (e.g., Clinton), a geographical term (e.g., Nice, France) or other proper names (for words beginning with an uppercase letter and not belonging to one of the two previous lists, e.g., Nirvana). Thus, for each available query translation we estimated the probability that the underlying translation may or may not be the best, according to the following explanatory variables:

- translation tool used to produce the translated query, variable denoted *source*;
- number of query terms, variable denoted *concepts*;
- minimal *idf* value over all search terms, variable denoted *minidf*;
- maximal *idf* value over all search terms, variable denoted *maxidf*;
- *idf* average value over all search terms, variable denoted *avgidf*;

- presence or not of personal proper nouns, binary variable denoted *person*;
- presence or not of geographical proper nouns, binary variable denoted *geo*;
- presence or not of other proper names, binary variable denoted *other*;

For each translation, we grouped the values of the previously described variables within a vector **X** = [x_1, x_2, … x_k], and then use it to estimate the probability that the underlying translation was a good one, according to the logistic regression model.

$$\text{Prob[good translation} \mid \mathbf{X}] = \frac{e^{\alpha+\beta_1 \cdot x_1+\beta_2 \cdot x_2+\cdots+\beta_k \cdot x_k}}{1+e^{\alpha+\beta_1 \cdot x_1+\beta_2 \cdot x_2+\cdots+\beta_k \cdot x_k}} \qquad (3)$$

within which the underlying parameters α, β_1, β_2, … β_k had to be estimated using a training set. In our experiments, these estimations were done using the R software (Venaples & Ripley, 1999).

In order to evaluate our selection procedure for queries submitted in English, we chose the Spanish and German collection, given that these two languages reflect different linguistic families. For the German corpus, the best translation tool (Reverso, see Table 3) provides slightly better retrieval effectiveness than the Google system. For the Spanish corpus, the Reverso system is clearly the best translation resource when compared to the other alternatives.

As a baseline, we decided to always select the translation resulting from the Reverso MT system that achieved a MAP of 0.4613 (Spanish), respectively 0.3074 (German). A query-by-query analysis as shown in Table 4 also indicates that the Reverso provided the best translation for 29 queries out of a total of 99 for the Spanish language and 22 for the German. In order to obtain retrieval effectiveness superior to that produced by the Reverso system, we evaluated three selection models.

Model A (using all information). In this model we took all available information into account, with the remaining explanatory variables for Spanish being: *source*, *minidf*, *person*, *geo* and *other*. For German we based the selection on: *source*, *maxidf*, *avgidf* and *person* explanatory variables. We observed that this selection model did not explicitly include all variables, because we had selected the most predictive variables using the `stepAIC` procedure, a stepwise selection procedure that minimizes the AIC criterion[1] (Venaples & Ripley, 1999). During this variable selection procedure, the system retained the name of the translation tool (variable *source*). When searching within the Spanish or German corpus, the fact that Reverso produced the translation could be viewed a priori as a good indication that this query translation would be effective. As additional evidence related to translation effectiveness, the system accounted for the *idf* values (either using the *minidf* (Spanish) or *avgidf* and *maxidf* variables (German)), and also the binary variables pertaining to the presence of proper nouns (variable *person* for both languages, in addition to *geo* and *other* when selecting the best Spanish translation).

Model B (without the translation tool name). In our second selection model, we removed the translation source from potential explanatory variables. After applying the `stepAIC` stepwise selection procedure, the system retained the following variables when selecting the best Spanish query translation: *concepts*, *minidf*, *person*, *geo* and *other*. For German, the selection

[1] The Akaike information criterion (AIC) value is computed as -2 log likelihood + 2 * number of parameters in the fitted model. This criterion maximizes the probability of obtained the corresponding sample under the given model but also penalizes model requiring a greater number of parameters (Ockham's razor principle).

procedure would be based on *concepts*, *avgidf*, and *person* variables. The main difference from Model A is that the variable *source* was replaced by the number of indexed search terms included in the translated query (variable *concepts*).

Model C (using only *idf*-based information). In this third model we wanted to evaluate the effectiveness of the *idf* value as a predictor of query performance. We thus removed all binary variables and the translation name (*source*) from the set of explanatory variables. While Cronen-Townsend *et al*. (2002) suggested using the average *idf* value, the `stepAIC` selection procedure retained the *concepts*, *minidf*, *avgidf* and *maxidf* variables for the Spanish language. For the German language, the selection model was rather simple; the only explanatory variables were *concepts* and *avgidf*. When using only the average *idf* value (variable denoted *avgidf*) with the Spanish collection, we obtained a MAP of 0.4297 vs. 0.4613 (-6.8% in relative percentage). With the German corpus, using only the average *idf* and selecting the best query translation, we obtained a MAP of 0.2505 vs. 0.3074 (or -18.5%).

To evaluate our three query translation selection models, we might use the 99 queries to estimate the coefficients α, β_1, β_2, … β_k of the logistic regression and the same set of queries to compute the MAP. Such an evaluation methodology is biased, but could be employed to provide an upper bound value related to the underlying model's performance. To produce an unbiased estimation of real performance we therefore preferred using the leaving-one-out approach. In this case, the training set was taken from all queries except one, and the last one was used to compute the average precision for this single query. Finally, we iterated this procedure over the query samples, generating 99 different training sets (composed of 98 queries) and 99 query evaluations from which a MAP could be computed.

	Mean average precision (% change)		
	Model A	Model B	Model C
Reverso (Spanish)	0.4613	0.4613	0.4613
Single translation	0.4613 (0.0%)	0.4085 (-11.4%)	0.4016 (-12.9%)
tolerance 10%	0.4613 (0.0%)	0.4780 (+3.6%)	0.4549 (-1.4%)
tolerance 15%	0.4609 (- 0.1%)	**0.4875** (+5.7%)	0.4713 (+2.2%)
tolerance 25%	**0.4622** (+0.2%)	0.4858 (+5.3%)	**0.4824** (+4.6%)
Reverso (German)	0.3074	0.3074	0.3074
Single translation	0.3037 (-1.2%)	0.2625 (-14.6%)	0.2671 (-13.1%)
tolerance 10%	0.3216 (+4.6%)	0.2869 (-6.7%)	0.2877 (-6.4%)
tolerance 15%	0.3202 (+4.2%)	0.3200 (+4.1%)	0.3118 (+1.4%)
tolerance 25%	**0.3309** (+7.6%)	**0.3406** (+10.8%)	**0.3384** (+10.1%)

Table 6: Mean average precision achieved by our three selection models (leaving-one-out)

In a first experiment, out of the eight available query translations the system was able to select only a single query translation. Table 6 shows the MAP achieved by this strict selection procedure in the line labeled "Single translation." When inspecting the selection made by Model A, the decision was rather simple: for the Spanish collection, we should always select the translation returned by the Reverso system. For both Models B and C, our evaluation resulted in a MAP decrease of between 11.4% and 12.9% with the Spanish corpus or between 13.1% and 14.6% with the German corpus.

Instead of trying to select the single best query translation, we might select a small set of good query translations. As seen from the data depicted in Table 4, the best translation for a given query may be obtained from more than one translation resource. To account for this, we assumed that if the probability estimated by Equation 3 resulted in a relative decrease of 10% (15% or 25%, corresponding to the tolerance values depicted in Table 6) when compared to the highest probability, we might assume that the underlying query translation would also be good. In such cases, the computer might select not only one, but a few query translations as being effective translations of the current English topic. Finally, when the computer selects more that one translation, they are concatenated and thus form the final translation to be used for retrieval.

Table 6 shows an evaluation of our proposition, where tolerances of 10%, 15% and 25% were used, both for the Spanish and German languages. This table shows that the most effective retrieval performances were usually obtained when a tolerance of 25% and selection Model B are used. The resulting performance differences were not always statistically significant (values underlined in Table 6). In order to compare the MAP values depicted in Table 6 with another baseline, for each query we randomly selected both the number of query translations (following a uniform distribution between 1 and 4) and the name(s) of the translation resource to be concatenated (following an uniform distribution between 1 and 8). With the Spanish collection, this random query translation selection produced a MAP of 0.3929, and 0.2724 for the German corpus. These performance levels were clearly below those achieved by our suggested query translation selection.

It was however a surprise to find that Model A did not produce the best performance, even though this selection model was based on all available information. In fact, the underlying probabilities computed using this selection model assigned too much bias to the fact that the translation was provided by the Reverso system, one that did not always produce the best translation. Finally, the performance difference between Models B and C represents the impact of using binary variables related to names. As shown in Table 6, these differences are rather small.

6. Conclusion

In this paper we propose a new query-by-query translation selection procedure based on logistic regression. From the outcome of our experiments we are able to draw several conclusions. First, whenever a query must be translated before performing a search within a corpus written in another language, human translations provide better retrieval performance for all languages studied (see Table 3). The differences are statistically significant and thus they too favor manual translation. However, performance differences between manual and machine-based translation queries vary from one language to another (e.g., from a relative percentage of 15.8% for the Spanish collection to 22.7% for the German corpus, as shown in Table 3). Moreover, the MAPs obtained by the different machine-based translation tools vary from 0.3074 to 0.2223 (German) and from 0.4613 to 0.3234 (Spanish).

Second, within a given translation resource translation quality varies widely from one query to another, and this variation is independent of the target language. For example, in our set of 99 queries (see Table 4), for the target language the best translation tool was the only one able to produce the best translation for 17 queries while for Spanish it was 20 times. Moreover, the MAPs resulting from selecting the best single translation on a query-by-query basis showed performance levels statistically similar to those obtained by manual translation.

Third, based on the statistical properties of the automatically translated queries, our logistic regression model cannot effectively predict the single best translation when compared to the best translation tool (Reverso). Our suggested selection method may however effectively predict a reduced set of best translations on a per query basis. When with this type of automatic selection procedure, the resultant retrieval performance is statistically better that the best automatic translation resource (Reverso, see Table 6). When using all available information in our selection procedure (Model A), the resultant retrieval performance was not the best. When limited to various *idf*-based measurements (Model C), the selection procedure tends to produce better MAP for both languages.

Acknowledgements

This research was supported in part by the Swiss NSF under Grants #200020-103420, and #200021-113273.

References

Bookstein, A., O'Neil, E., Dillon, M., & Stephen, D. (1992). Applications of Loglinear Models for Informetric Phenomena. *Information Processing & Management*, 28(1), 75-88.

Braschler, M., Ripplinger, B., & Schäuble, P. (2002). Experiments with the Eurospider retrieval system for CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems* (pp. 102-110). Berlin: Springer-Verlag, Lecture Notes in Computer Science #2406.

Braschler, M., & Ripplinger, B. (2004). How Effective Is Stemming and Decompounding for German Text Retrieval? *IR Journal*, 7(3-4), 291-316.

Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval. *IR Journal,* 7(1-2), 183-204.

Buckley, C., Mitra, M., Walz, J., & Cardie, C. (1998). Using Clustering and Super Concepts within SMART: TREC’6. In E.M. Voorhees & D.K. Harman (Eds.), *Proceedings TREC'6*, (pp. 107-124). Gaithersburg: NIST Publication #500-240.

Callison-Burch, C., & Flournoy, R.S. (2001). A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engine. In *Proceedings of the 8th Machine Translation Summit.*

Chen, A., & Gey, F.C. (2004). Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. *IR Journal*, 7(1-2), 149-182.

Cronen-Townsend, S., Zhou, Y., & Croft, W.B. (2002). Predicting Query Performance. In K. Jarvelin, M. Beaulieu, R. Baeza-Yates & S.H. Myaeng (Eds.), *Proceedings of the 25th International Conference on Research and Development in Information Retrieval (ACM SIGIR 2002)* (pp. 299-306). New York: The ACM Press.

Gao, J., & Nie, J.Y., (2006). A Study of Statistical Models for Query Translation: Finding a Good Unit of Translation. In E.N. Efhimiadis, S. Dumais, D. Hawking, K. Järvelin (Eds.), *Proceedings of the 29th International Conference on Research and Development in Information Retrieval (ACM SIGIR 2006)* (pp. 194-201). New York: The ACM Press.

Gey, F.C. (1994). Inferring Probability of Relevance using the Method of Logistic Regression. In W.B. Croft & C. van Rijsbergen (Eds.), *Proceedings of the 17th International Conference on Research and Development in Information Retrieval* (ACM SIGIR ‘94) (pp. 222-231). New York: Springer-Verlag.

Grefenstette, G. (Ed.) (1998). *Cross-Language Information Retrieval*. Amsterdam: Kluwer.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Ed., New York: John Wiley.

Knuth, D.E. (2001). *Things a Computer Scientist Rarely Talks About*. Stanford: CSLI Publications.

Kwok, K. L., Grunfeld L., Dinstl, N., & Chan, M. (2001). TREC-9 Cross-Language, Web and Question-Answering Track Experiments using PIRCS. In E.M. Voorhees & D. Harman (Eds.), *Proceedings TREC-9* (pp. 417-426). Gaithersburg: NIST Publication #500-249.

Kwok, K.L., Grunfeld, L., Dinstl, N., & Deng, P. (2003). TREC 2003 Robust, HARD and QA Track Experiments using PIRCS. In E.M. Voorhees & L.P. Buckland (Eds.), *Proceedings TREC-2003*, (pp. 201-209). Gaithersburg: NIST Publication #500-255.

Le Calvé, A., & Savoy, J. (2000). Database Merging Strategy Based on Logistic Regression. *Information Processing & Management*, 36(3), 341-359.

McNamee, P., & Mayfield, J. (2004). Character N-gram Tokenization for European Language Text Retrieval. *IR Journal*, 7(1-2), 73-97.

Oard, D.W., & Resnik, P. (1999). Support for Interactive Document Selection in Cross-Language Information Retrieval. In *Information Processing & Management*, 35(4), 363-379.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.). (2002). *Evaluation of Cross-Language Information Retrieval Systems*. Lecture Notes in Computer Science #2406, Berlin: Springer-Verlag.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.). (2003). *Advances in Cross-Language Information Retrieval*. Lecture Notes in Computer Science #2785, Berlin: Springer-Verlag.

Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magini, B., & de Rijke, M. (Eds.). (2006). *Accessing Multilingual Information Repositories*. Lecture Notes in Computer Science #4022, Berlin: Springer-Verlag.

Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P., & Hansen, P. (2004). Observing Users, Designing Clarity: A Case Study on the User-Centered Design of a Cross-Language Information Retrieval System. *Journal of the American Society for Information Science and Technology*, 55(10), 923-934.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Savoy, J., Ndarugendamwo, M., & Vrajitoru, D. (1996). Report on the TREC-4 Experiment: Combining Probabilistic and Vector-Space Schemes. In D.K. Harman (Ed.), *Proceedings of the TREC'4*, (pp. 537-547). Gaithersburg: NIST Publication #500-236.

Savoy, J. (1997). Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy, J. (2004a). Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval. *IR Journal*, 7(1-2), 121-148.

Savoy, J. (2004b). Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In C. Peters, J. Gonzalo, M. Braschler, & M. Kluck (Eds.), *Comparative Evaluation of Multilingual Access Systems* (pp. 322-336). Berlin: Springer-Verlag, Lecture Notes in Computer Science #3237.

Savoy J. (2005). Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. *ACM Transactions on Asian Languages Information Processing*, 4(2), 163-189.

Venaples, W.N., & Ripley, B.D. (1999). *Modern Applied Statistics with S-PLUS*. New York: Spinger-Verlag.

Voorhees, E.M. (2006). The TREC 2005 Robust Track. *ACM-SIGIR Forum*, 40(1), 41-48

Yom-Tov, E., Fine, S., Carmel, D., & Darlow, A. (2005). Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In G. Marchionini, A. Moffat, J. Tait (Eds.), *Proceedings of the 28th International Conference on Research and Development in Information Retrieval (ACM SIGIR 2005)* (pp. 512-519). New York: The ACM Press.