Published in Journal of the American Society for Information
Science and Technology 61, issue 11, 2266-2273, 2010
which should be used for any reference to this work

Retrieval Effectiveness of Machine Translated Queries

Ljiljana Dolamic and Jacques Savoy
Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2009 Neuchâtel, Switzerland. E-mail: {Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

This article describes and evaluates various information retrieval models used to search document collections written in English through submitting queries written in various other languages, either members of the Indo-European family (English, French, German, and Spanish) or radically different language groups such as Chinese. This evaluation method involves searching a rather large number of topics (around 300) and using two commercial machine translation systems to translate across the language barriers. In this study, mean average precision is used to measure variances in retrieval effectiveness when a query language differs from the document language. Although performance differences are rather large for certain languages pairs, this does not mean that bilingual search methods are not commercially viable. Causes of the difficulties incurred when searching or during translation are analyzed and the results of concrete examples are explained.

Introduction

In crossing language barriers, the English language often plays a central role in facilitating communication among people speaking different languages. In Europe, for example, as well as in large international organizations or companies (e.g., WTO, IBM, Novartis), the quantity of information written in English tends to be growing rapidly. Additionally, accessing information on the Web (Chung, 2008) in this language is increasingly necessary (news, hotel or airline reservations, government information, statistics, etc.). Although some users are perfectly bilingual, many others can read documents written in English but cannot formulate a request, or they, at least, cannot provide reliable search terms in a form comparable to those found in the documents being searched. On the other hand, many documents contain nontextual information such as images, videos and statistics that do not need translation or can be understood regardless of the language involved.

Although English is not the language spoken by the majority of people from around the world, as an *interlingua* medium for transmitting knowledge or expressing opinions, it clearly plays a central role. The CNN success story serves as just one example of the increasing importance of this language. Moreover, English is often the first foreign language learned in Europe, India, or the Far East. It is important, therefore, to provide adequate resources for translating from other languages to English, or vice-versa, while also analyzing translation quality.

The most important commercial search engines have certainly not ignored this demand for translation resources to and from English. Google, for example, in an effort to improve the searching of Web pages available in English, regardless of the language in which the topic is written, has launched a translation service in May 2007 that provides two-way online translation services, mainly between English and 50 other languages (http://translate.google.com/). Other free Internet translation services have been made available over the last few years. Yahoo! (http://babelfish.yahoo.com), for example, also offers a freely available translation system, and after more than 10 years of research in cross-lingual information retrieval, other commercial products have also been made freely available to Internet users.

In this article, the objective is to address the following questions: How effective is a bilingual search? What is the "information retrieval cost" to Web users who formulate requests in their own language and then search the Web and find information written in English? If we compare two translation services, does their relative effectiveness depend only on the underlying information retrieval (IR) model? Does translation quality depend on the relationship between the source and target languages, with a better quality being obtained by those languages that have a close relationship with English (e.g., French, German) as opposed to Spanish or more distant languages such as Chinese? Although we will not evaluate translations per se, we will test and analyze various IR and translation systems in terms of their abilities to retrieve items written in English, based on an automatically translated query (experiments conducted in December 2008).

The rest of this article is divided as follows. The Translation Approaches section presents related works, while the Test-Collection section depicts the main characteristics of the test collection. The IR Models section briefly describes the IR models that were used during our experiments, while the Evaluation section evaluates them under different conditions and points out some of the main problems found with the automatic translation tools being applied. A query-by-query analysis will complete this evaluation, and this article's main findings are summarized in the Conclusion section.

Translation Approaches

To be effective, a bilingual search (topic expressed in one language, document retrieved in one or more other languages) must be able to cross the language barrier. One approach to this problem is to assume that one language is merely a misspelled form of the other, as, for example, "English is French, misspelled" (Buckley, Singhal, Mitra, & Salton, 1996). Such an approach based on cognate matches may work with closely related languages (and when an effective "spelling corrector" is included). An evaluation carried out by McNamee and Mayfield (2004) has, however, shown that mean average precision varied from 9% to 27%, compared with the 45% achieved by a monolingual search, thus representing a relatively high decrease.

As a first real translation tool, various researchers suggested using machine-readable dictionaries (MRDs) (Ballesteros & Croft, 1997; Hull & Grefenstette, 1996; Hedlund et al., 2004). However, when employing MRDs, we need to handle the out-of-vocabulary (OOV) problems that result from a dictionary's limited coverage. In a related issue, it could prove helpful to recognize proper nouns and acronyms and translate them by applying a special dictionary (e.g., for the English-French languages, we would find Putin-Poutine, UNO-ONU, SIDA-AIDS). Moreover, certain input words could be ambiguous and MRDs might suggest more than one translation (e.g., the word "bank" can take on a different meaning when used in the context of a river or a financial institution). Sometimes we need to automatically transform input words into base form (lemma) listed in the dictionary, although this process may result in errors and semantic shifts (e.g., the word "saw" in "I saw a man with a saw").

As a second translation tool, we might make use of easily accessible machine translation (MT) systems (Chen & Gey, 2004). However, such devices tend to perform poorly when translating entire documents, in part, because translation is a semantic-based operation. Moreover, the best translation is not always produced by simply following the syntactical structure of the source language (Mel'èuk & Wanner, 2006), even for closely related languages. For example, the translation of the road sign "slow men at work" into French would give "ralentir, travaux" (slow, works), and this illustrates the need to process idiosyncratic transformations between the source and the target syntactic structures.

As a third possibility of identifying proper translation candidates, we might apply a corpus-based translation method in conjunction with a statistical translation model (Nie Simard, Isabelle, & Durand, 1999). In this case, we would need to access the corpora at hand and automatically build data structures from which direct translations or related term generations could be obtained (Sheridan & Ballerini, 1996), using the most probable match or the best k matches (Braschler & Schäuble, 2001). This presume, of course, that parallels or comparable corpora would be available for certain domain-specific language pairs, yet such corpora would clearly be more difficult to find. The performance of these statistical translation approaches would depend on very important factors, such as source quality (e.g., extracted Web sites) and size (Nie & Simard, 2002), along with the role played by cultural, thematic, and time differences in such methods (Kwok, Grunfeld, Dinstl, & Chan, 2001). Finally, assessing translation probabilities could be problematic and may result in disappointing performance levels, particularly when a lot of query terms and their correct translations cannot be found in an aligned corpus (Hiemstra, Kraaij, Pohlmann, & Westerveld, 2001).

Test-Collection

In an effort to promote IR in languages other than English and to evaluate bilingual searches (requests expressed in one language, documents returned in others), over the last years, various evaluations have been conducted during the TREC (Harman, 2005), NTCIR, and CLEF campaigns (Peters et al., 2007).

In our evaluation work on the retrieval effectiveness of bilingual searches involving the topic descriptions written in various languages and retrieving documents written in English, we made use of a corpus created during the various CLEF campaigns. This collection comprised articles published in 1994 in the *Los Angeles Times* newspaper, as well as documents extracted from the *Glasgow Herald* and published in 1995, comprising a total of 169,477 documents (or about 579 MB of data). On average, each article contains about 250 (median, 191) content-bearing terms (not counting commonly occurring words such as "the," "of," or "in"), and the documents in this collection are typically represented by a short title plus one to four paragraphs of text.

This collection contains 310 topics, each subdivided into a brief title (denoted as T), a full statement of the information need (called description or D), plus any background information that might help assess the topic (narrative or N). As shown in Table 2, the topic titles comprise two or three words (before stopword removal). Such short forms reflect typical Web search requests and are represented by a set of keywords that begin with capitals rather than a grammatically complete phrase. These topics cover various subjects (e.g., "El Niño and the Weather," "Chinese Currency Devaluation," "Eurofighter," "Victories of Alberto Tomba," "Marriage Jackson-Presley," or "Computer Animation"), and

TABLE 1. General statistics on our test-collection for each year.

	2001	2002	2003	2004	2005	2006
Source	*LA Times*	*LA Times*	*LA Times, Glasgow H.*	*Glasgow H.*	*LA Times, Glasgow H.*	*LA Times, Glasgow H.*
Size	425 MB	425 MB	579 MB	154 MB	579 MB	579 MB
No. docs	113,005	113,005	169,477	56,472	169,477	169,477
No. topics	47	42	54	42	50	49
Topics	#41–#90	#91–#140	#141–#200	#201–#250	#251–#300	#301–#350

TABLE 2. Examples of CLEF topic formulation.

<num> C092 </num>
<title> U.N. sanctions against Iraq </title>
<description> What measures has Iraq taken to effect the lifting of the U.N. economic embargo and political sanctions imposed after its invasion of Kuwait in 1990? </description>
<narrative> Documents must include ways in which Iraq has attempted to get the sanctions lifted. Mere descriptions of the sanctions or rhetoric against the sanctions are not relevant. Expressions of regret for invading Kuwait by Iraqi officials are relevant. </narrative>
<num> C147 </num>
<title> Oil Accidents and Birds </title>
<description> Find documents describing damage or injury to birds caused by accidental oil spills or pollution. </description>
<narrative> All documents which mention birds suffering because of oil accidents are relevant. Accounts of damage caused as a result of bilge discharges or oil dumping are not relevant. </narrative>
<num> C208 </num>
<title> "Sophie's World" </title>
<description> Find documents about the editorial success of the book "Sophie's World" by Jostein Gaarder. </description>
<narrative> Relevant documents should describe the topic "Sophie's World", and should mention its sales success. </narrative>

they include both regional ("Films Set in Scotland," "Area of Kaliningrad") and international coverage ("Oil Prices," "Sex in Advertisements").

These topic sets were compiled during the various CLEF evaluation campaigns by various groups of experts that represent all the available languages in which the articles were written (e.g., around 8 to 10 topics per collection). These selected topics were then manually translated by native speakers, who produced complete sets of topics in English and in all the other languages. During this stage, the human translators did not have access to relevance assessments in any language (Braschler & Peters, 2004).

For each language, relevance assessments were done by the same human assessors who helped compile the topic set. During this process, the narrative part of topic formulations made it easier to obtain a better understanding of the user's real information need and also helped to identify both relevant and irrelevant documents (Voorhees & Harman, 2005). The topic descriptions listed Table 2 are good examples.

As shown in Table 1, the entire corpus was not used during all evaluation campaigns and, thus, different parts of the corpus had to be searched for pertinent articles. For example, Topics #201 to #250 were created in 2004 and the responses resulted from searches in the *Glasgow Herald* (1995) collection, a subset representing 56,472 documents. Of the 50 topics originally available in 2004, we found that only 42 contributed to at least one correct answer.

From the original set that comprised topics, 26 were removed because they had no relevant documents in the corpus, meaning only 284 topics were used in our evaluation. Upon an inspection of these relevance assessments, the average number of correct responses for each topic was 22.46 (standard deviation [SD] 28.9, median, 11.5), with Topic #254 ("Earthquake Damage") obtaining the greatest number of relevant documents (229).

The topics were manually translated in different languages and the German, French, Spanish, and simplified Chinese topic descriptions were used in this study. The European language topics were encoded in ISO-8859-1 and the Chinese in the GB2312 format.

IR Models

To obtain a broader view of the relative merit of the various retrieval models, we used one vector-space scheme and three probabilistic models. First, we adopted the classical *tf.idf* model, a well-known vector-processing scheme used since the late 1970s in various applications such as document clustering or automatic text categorization (Sebastiani, 2002). In this case, the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document d_i) and its inverse document frequency (or idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights (Manning, Raghavan, & Schütze, 2008).

In addition to this classical vector-space scheme, we also considered probabilistic models such as that of Okapi (or BM25; Robertson, Walker, & Beaulieu, 2000). As demonstrated in various evaluation campaigns and under different search tasks (Peters et al., 2007), this method usually results in highly effective retrieval. In this case, the importance of term t_j in describing the content of document d_i is weighted by:

$$w_{ij} = [(k_1 + 1) \cdot tf_{ij}]/(K + tf_{ij}) \text{ with } K = k_1 \cdot [(1 - b) + ((b \cdot l_i)/mean\ dl)] \quad (1)$$

where *mean dl* indicates the average document length, and K and b are two constants set to 1.2 and 0.55, respectively.

The Okapi model does not always produce the best performance results when compared with a second probabilistic

approach we implemented in the I(n_e)C2 model, as applied within the *Divergence from Randomness* (DFR) framework (Amati & van Rijsbergen, 2002). In this case, the two information measures are combined as follows:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1–Prob^2_{ij}) \quad (2)$$

where $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document and $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, provided that tf_{ij} occurrences of this term had already been found. The I(n_e)C2 model was based on the following formulae:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1)/(n_e + 0.5)] \quad (3)$$

with $n_e = n \cdot [1 - [(n-1)/n]^{tc_j}]$ and $tfn_{ij} = tf_{ij} \cdot \ln[1 + ((c \cdot mean\ dl)/l_i)]$

$$Prob^2_{ij} = 1 - [(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))] \quad (4)$$

where tc_j is the number of occurrences of term t_j in the collection, df_j the number of documents in which the term t_j occurs, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (= 271), the average document length, and c a constant (fixed at 1.5). This type of IR model tends to produce very effective results when applied to a variety of corpus or tasks (Peters et al., 2007).

Finally, we also considered a language model (LM; Hiemstra, 2000), known as a nonparametric probabilistic model (the Okapi and DFR are viewed as parametric models). With this latter LM approach, we had covered a large majority of the possible probabilistic retrieval model implementations. In this case, probability estimates were not based on any known distribution but rather were estimated directly and based on occurrence frequencies in document d_i or the entire C corpus. Within this language model paradigm, various implementations and smoothing methods might also be considered, and in this study, we adopted a model proposed by Hiemstra as described in Equation 5. It combines an estimate based on document ($P[t_j|d_i]$) and corpus ($P[t_j|C]$).

$$P[d_i|q] = P[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j|d_i] + (1 - \lambda_j) \cdot P[t_j|C]]$$
$$\text{with } P[t_j|d_i] = tf_{ij}/l_i \text{ and } P[t_j|C] = df_j/lc$$
$$\text{with } lc = \sum_k df_k \quad (5)$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the underlying corpus C.

Evaluation

Given the relatively large number of topics (284), we were in a position to evaluate the retrieval effectiveness of queries that were submitted in various languages to search a corpus written in English. In doing so, we wanted "the truth, the whole truth (recall), and nothing but the truth (precision)." The precision would represent the proportion of relevant retrieved items, while the recall would indicate the proportion of relevant documents retrieved (Braschler & Peters, 2004). Instead of using these two measures to compare the performance obtained by two search systems, we would prefer adopting a unique performance value, one that would also account for the rank of both the retrieved and relevant documents. To do so, we have adopted the mean average precision (MAP), computed by the `trec_eval` software (based on a maximum of 1,000 retrieved records; Buckley & Voorhees, 2005). This performance measure represents the mean of every average precision (AP) value, obtained by each query on the test-collection. To define this average precision for a given query, we measured the precision after each relevant document was retrieved and then computed the mean. The AP, thus, accounts for the precision, the recall and the rank of the relevant items.

To determine whether a given search strategy would be better than another, we applied the nonparametric bootstrap test (Savoy, 1997), a statistical test in which the null hypothesis H_0 states that the two retrieval schemes used in the comparison produce similar retrieval performance. Thus, in the experiments presented in this article, statistically significant differences were detected by a two-sided test (significance level 5%), and the corresponding computations were done using the R language.

To complete this type of overall evaluation based on both precision and recall and to obtain a better understanding of the effect of a given search strategy, we analyzed the retrieval performance of certain queries. This query-by-query inspection was intended to provide detailed information on the drawbacks of the underlying IR scheme.

In the following sections, we begin by presenting the retrieval effectiveness of various IR models in a monolingual context, which are to be used as a baseline in our further experiments. Then, we evaluate these IR models, using different query languages, followed by describing a query-by-query analysis, revealing the main translation difficulties.

Monolingual Evaluation

To define a baseline, we tested four IR models, using the topics written in the English language (monolingual search). In our experiments, we considered only the topic titles (T, mean number of search terms = 2.8, median, 3, minimum, 1, maximum, 7, *SD* 0.86). Short topic formulations such as these more closely reflect the current practice of Web users of sending their requests to commercial search engines. In this case, the mean number of search terms per request was estimated as 2.4 (Jansen & Spink, 2006). These resulting MAP (see Table 3), will be our baseline when evaluating bilingual searches.

As shown in Table 3, the I(n_e)C2 model provided the best retrieval results, which were statistically significant and better than those achieved by either the LM or *tf.idf* approaches

TABLE 3. Mean average precision (MAP) for monolingual searches (284 title-only queries).

Information retrieval model	MAP
$I(n_e)C2$	**0.4053**
Okapi	0.4044
Language model	0.3708*
tf·idf	0.2392*

(as denoted by an "*"). On the other hand, the performance difference between the Okapi and $I(n_e)C2$ cannot be viewed as significant.

Although the performance that resulted from these topics forms the baseline, we must mention that these formulations do not form a gold standard. Other people may write slightly different requests that correspond to the same user information need (e.g., instead of having "Oil Accidents and Birds" as in Table 2, we may encounter "Birds and Oil Pollution"). Moreover, when translating manually from one language (e.g., French) to another (e.g., English), various human translators may propose different translations. Savoy (2003) shows that based on full topic descriptions (three examples are given in Table 2), the relative performance differences could be as high as 30% (Okapi model, 34 queries) when the topic titles written in English are manually translated into French. In this study, however, the experiments were conducted with students and not with expert human translators.

Bilingual Evaluation

In bilingual searches, the documents are written in one language (in English in our case), while the search topics are written in another language. To obtain a broader viewpoint, we selected four different topic languages: two from the Latin family (French [FR] and Spanish [SP]), and one from German (DE) to have another language related to English. To include a language with a very different morphology and writing system, we chose the Chinese (ZH) language.

For the title descriptions available in these languages, we had them automatically translated into English, using the Google translation service, and then used the translations (done in December 2008) to search our newspaper corpus. The resulting MAPs are listed in Table 4, while Table 5 shows the same retrieval measures obtained from the Yahoo! translation service.

TABLE 4. Mean average precision (MAP) for both monolingual and bilingual searches, using the Google translation service (284 title-only queries).

	Mono	From ZH	From DE	From FR	From SP
$I(n_e)C2$	**0.4053**	**0.3340***	0.3618*	**0.3719***	0.3741*
Okapi	0.4044	0.3327*	**0.3625***	0.3692*	**0.3752***
LM	0.3708	0.3019*	0.3305*	0.3400*	0.3426*
tf·idf	0.2392	0.1920*	0.2266*	0.2294	0.2256*
Mean difference %		−18.2%	−9.3%	−7.3%	−7.1%

TABLE 5. Mean average precision (MAP) for both monolingual and bilingual searches, using the Yahoo! translation service (284 title-only queries).

	Mono	From ZH	From DE	From FR	From SP
$I(n_e)C2$	**0.4053**	**0.2286*†**	**0.2951*†**	**0.3322*†**	**0.2897*†**
Okapi	0.4044	0.2245*†	0.2917*†	0.3268*†	0.2867*†
LM	0.3708	0.2000*†	0.2636*†	0.3006*†	0.2600*†
tf·idf	0.2392	0.1289*†	0.1846*†	0.2065*†	0.1812*†
Mean difference %		−45.1%	−26.7%	−17.5%	−27.9%

In Tables 4 and 5, the last rows list the mean percentage differences when compared with the corresponding monolingual search. The MAP differences, usually statistically significant (values denoted by an "*"), show that bilingual searches always resulted in lower retrieval performance. The only exception was the difference that was obtained using the *tf.idf* model and the Google translation service, for topics written in French. In this case, the performance difference between the monolingual and bilingual searches (0.2392 vs. 0.2294) was not statistically significant.

From comparing the different languages when using the Google system, we saw that, for Google, the translation from the French or Spanish language was easier than it was from Chinese language. Using the $I(n_e)C2$ IR model, the MAP obtained for the bilingual French or Spanish language searches was 92% of the value obtained for the monolingual search, 90% for German topics, and 82% for the simplified Chinese language. For Yahoo!, the situation was somewhat comparable; with the French language, the search obtained the best precision (82% for the monolingual search) and Chinese was the most difficult (only 55% for the monolingual search).

Moreover, when comparing the translations obtained from the Yahoo! and Google systems, Yahoo! seemed, on average, to encounter more problems. To verify this, and using the Google translation performance as a baseline (Table 4), we compared them with those obtained by Yahoo! (Table 5). Based on an analysis of performances of the two translation devices, the differences were always statistically significant (denoted by an "†" in Table 5).

Query Translation Difficulties

To learn why translation failed with some searches, we analyzed the retrieval performance of all individual queries. We hoped to identify systematically occurring types of translation error. To limit our investigation somewhat, queries considered problematic were those resulting in a decrease of more than 10% in average precision. Moreover, we should mention that both Google and Yahoo! translation services were not specifically designed to handle cross-lingual information retrievals. These systems, thus, send back only simple translations without including possible alternatives or attaching weights to words, thus indicating other translation probabilities. In an embedded IR system,

this additional information could be useful in finding more relevant documents or in ranking them higher in the result list.

The first source of translation difficulties was the presence of proper names in the request. Although, in some cases, a name did not change from one language to English (e.g., "France" or "Haiti"), usually a modification had to be made (e.g., "London" is written "Londres" in French). We also encountered various topics that depicted similar problems, such as Topic #94 ("Return of Solzhenitsyn"), which was written as "le retour de Soljénitsyne" in French, "Retorno de Solzhenitsin" in Spanish, or "Rückkehr Solschenizyns" in German. When French or German was the query language, Yahoo!'s translation system was not able to return the correct English spelling for this name. It is also interesting to note that when Spanish was the query language, both MT systems failed to translate this personal name correctly.

The correct translation of a proper name could be rendered more difficult whenever it might have a specific meaning in the source language. For example, in Topic #89 ("Schneider Bankruptcy"), the name "Schneider" means also "cutter" in German and this meaning was selected by Yahoo!'s translation system, producing the phrase "Cutter bankruptcy." Topic #43 ("El Niño and the Weather") demonstrates another but related difficulty. In this case, the weather phenomenon was designated as a Spanish noun that also means "the boy." For the Spanish expression, Yahoo!'s translation service returns "the boy and the time," ignoring the fact that the topic contains a particular name. When selecting Chinese as the query language, and as shown in Table 6, both MT systems often cannot translate a proper name such as this, leaving the Chinese word untouched or returning a weird expression (e.g., for Topic #89 "Schneider Bankruptcy," we obtained "史特加 bankruptcy" from Google and "Shi Tejia goes bankrupt" from Yahoo!). Moreover, knowing that the Chinese language does not employ the same set of phonemes, the pronunciation and resulting spelling forms did not have a bijective relationship with the English phonology. With Topic #121 for example, for "Edouard Balladur," Google returned "Edward Baladu" while Yahoo! returned "Edward Baladoo," and for Topic #208 "Sophie's World," Yahoo! gave us "Su Fei world."

A second main source of translation errors was the polysemy that was attached to a given word in the source language. More precisely, to find the appropriate word (or expression) in the target language (English in our case), the translation system had to consider the context. In fact, a given word in one language can be translated by various words that involved different semantics. As shown previously, in the Spanish Topic #43 "El Niño y el tiempo," the word "tiempo" could be translated as "weather" or "time," and the latter was selected by Yahoo!'s system. With Topic #341 ("Theft of 'The Scream"') written in French as "Vol" du 'Cri,"' the French word "vol" could be translated by "flight" or "theft." In this case, the translation produced by Google was "The Flight of the 'Scream,"' while by Yahoo!, it was "Flight of the 'Cry."'

This latter translation demonstrates another problem that was related to the synonymy of a given set of words, wherein the translation system was faced with different translations but related meanings. In our example, the French word "cri" could be translated using "scream" or "cry," This synonymy aspect was also found in various topics that involved the related terms "car" and "automobile." In the original English version of the topics, the term "car" is used more frequently (five times to be precise) in the topic titles (e.g., Topics #106 "European car industry," Topics #288 "US Cars Import") and 18 times in all topic formulations. On the other hand, the term "automobile" was never used in the titles and only twice in the description part of two topics (note that our evaluations were done only on the topic titles). Moreover, the semantic relationship between two (or more) alternatives is not always that close, as illustrated by Topic 67 "Ship Collisions." In this case, Yahoo! returned "Naval collisions" as the translation from Spanish.

As a third translation difference with the original English description, we found different morphologies and grammatical categories. Also, when expressing an idea, we can select different forms from the same root, (e.g., "merger," "merge," or "merging"). For example, from the original Topic #196 "Merger of Japanese Banks," the system ranked the first relevant item in the top position, while with the translation "Merging of Japanese Banks," the first relevant article appeared in the sixth position. The same problem occurred in Topic #165 "Golden Globes 1994," for which the retrieval system returned a relevant document and ranked it in first position. With the translated query "Gold Globes 1994," however, the first relevant item appeared only in the sixth position. This last example also demonstrated that using a more aggressive stemmer, such as that of Porter (1980), to conflate word variants into a common stem, tended to be the most appropriate solution. In our case, with the form "golden," the IR system was able to rank a relevant item in the first position. On the other hand, Porter's stemmer was not able to conflate the forms "merger" and "merging" into the same root ("merging" is transformed into "merg" while "merger" is left untouched).

As a fourth main source of translation problems, we found that compound constructions, such as those occurring frequently in the German language, were not always translated into English. For example, with Topic #84 "Shark Attacks," from the German formulation, we obtained "Haifischangriffe" (Google), or with Topic #105 "Bronchial asthma," we obtained "Bronchialasthma" (Yahoo!). In both cases, however, for the German topic formulation,

TABLE 6. Translation error distribution according to source language and translation systems (284 title-only queries).

	Google				Yahoo!			
	ZH	DE	FR	ES	ZH	DE	FR	ES
Name	21	2	1	2	37	11	3	13
Polysemy/synonymy	16	4	11	11	27	21	23	14
Morphology	2	2	1	2	7	8	3	7
Compound	0	4	0	1	0	15	0	0
Other			2		6		2	19

the retrieval system was not able to find any relevant items. Using the original English form the IR system ranked a relevant item in the first position for both topics.

Other sources of translation problems can be found in the various topic languages. For example, the French Topic #200 ("Innondationeurs en Hollande et en Allemagne") contains a spelling error in the word "Innondationeurs" (instead of "Inondations"). The original French Topic #259 is written as "Lions d'or" (award name of the Venice Film Festival), which is incorrectly translated from "Golden Bear," the award name for the Berlin Film Festival. Even if the translation was correct, the translated query was not able to rank any relevant item in the top 10.

Our error classification is based on queries that resulted in clear and significant retrieval performance difference compared with the original English topics. In many other cases, the translation was not perfect, or was even incorrect, but the MAP performance was similar, or produced only a slight variation, and usually had a slight degradation. For example, with the Topics #192 ("Russian TV Director Murder"), the first relevant item was ranked third in a monolingual search. Yahoo! returned "Assassination of a director of Russian television" from the French language, and with this formulation, the IR system ranked the first relevant item in the fifth position. Using Google and French as the search language we obtained "The assassination of a head of Russian television" as the query translation. In this case, the first relevant document appeared only in the 30th position. For Chinese as the search language, we obtained "Russian television murder charge" with Google and "The Russian television station managers murder" with Yahoo!. With both MT systems, the translated search performed reasonably well compared with the original English search (with the Google translation, the first retrieved item was relevant, and with Yahoo!, the sixth, instead of the third with the English language).

Conclusion

Compared to a monolingual search, writing a topic in another language and then asking Google or Yahoo! to automatically translate it before launching a search will significantly degrade retrieval effectiveness. When using the MAP as a performance measure, this finding is valid for German, French, Spanish, or Chinese as the query language and for all IR models analyzed in this study (see Tables 4 and 5).

What does that means for a Web user? First we must recognize that the MAP values do not provide for the final users a simple and direct interpretation. Second, users are not concerned with the "mean behavior" of a search system but rather with how their particular queries behave. To take account for this second perspective and through a query-by-query analysis, we can determine that the best IR model was not able to rank a correct answer among the top 10 for 38 queries out of a total of 284 in the monolingual search. This result represents an "inaccuracy rate" of 13.4%, implying that the final user must either reformulate the submitted topics or search further down the result list. For the French requests that were automatically translated by the Google system, this value increased to 57 (or 20%), while for Yahoo!, the number of difficult queries increased to 73 (25.7%). In conjunction with the MAP values given in Tables 4 and 5, the retrieval of information that was based on a freely available translation service could be effective, showing a significant decrease in the MAP (e.g., from 0.4053 to 0.3719, a relative difference of 8.2%, using French as query language and the Google translation service). From our point of view, performance differences below 10% indicate that the prospects for this type of search system could be really good at the commercial level.

When submitting queries in the Spanish language instead of French, we could expect either similar performance levels (Google) or slightly decreased performance (Yahoo!). With simplified Chinese as the query language, however, retrieval performance decreases significantly, as does the number of topics for which no relevant items were listed among the top 10 (69 with Google, 114 with Yahoo!). With German as the query language, retrieval performances tend to lie between these two extremes.

This finding clearly indicates that the performance of a bilingual search does not depend exclusively on the source language. In our experiments, the French and Spanish languages achieve the best performance with the Google translation service, yet this finding cannot be confirmed for the Yahoo! system. Moreover, we might expect that bilingual searches that were based on closely related language pairs (such as French-English or German-English) would be easier than more distantly related languages (e.g., Chinese-English). In the absence of well-defined metrics that are able to measure the distance between languages, our study cannot easily confirm this prior assumption. Although topics written in French show high-retrieval effectiveness, those written in German always perform at lower levels than Spanish when using the Google translation system. With Yahoo!, however, the mean performance for German or Spanish topics is similar, while French topics result in much better retrieval effectiveness. When Chinese is the query language, performance levels are always lower than other languages, yet the last line in Tables 4 and 5 indicate that the mean difference for Chinese requests with the Google translation system (-18.2%, Table 4) could be lower than the expected results from the Indo-European languages using the Yahoo! system (-26.7% with German, -27.9% with Spanish, -17.5 with French, see Table 5).

When comparing two different freely available translation services, we can infer that, statistically, one system performs significantly better that the other (significant differences are denoted by an "†" in Table 5). This finding is grounded on four different query languages and four IR models.

When analyzing the performances of the various IR systems in conjunction with the different query languages and translation services, we do not see important differences when compared with the ranking defined for the monolingual search (see Table 3). The best results are achieved by either

the $I(n_e)C2$ or the Okapi model, while the LM approach ranks in second place and the classical vector-space model *tf.idf* in third place.

Finally, we analyze the query translations that were produced by the two MT systems to investigate their main difficulties and find four main causes of this performance degradation (Query Translation Difficulties section). An improved translation of names (personal, geographical, product) and better processing of German compounds will clearly improve bilingual searches. In our opinion, an increase in matches for ambiguous terms would also further improve translation quality (e.g., the French word "temps" could be translated as "time" or "weather," depending on the context). For the moment, however, it is not clear how this context could be effectively taken into account when handling 2.6 terms per query, on average. The synonymy problem (e.g., film/movie, ship/boat, car/automobile) was also a source of performance variations between the original and translated topics. Finally, choosing the most appropriate word form (even from a common root) could play a role in final retrieval performances ("merging" or "merger," "prehistorical" or "prehistoric," "golden" or "gold"). In this case, however, formulation can be difficult when designing a MT system to delimit the precise boundary between good and less effective topic. On the other hand, it is worth noting that we were surprised to confirm that frequently used acronyms were usually correctly translated (e.g., "ONU" and "UNO" or "UN"), a feature that was absent a few years ago.

Acknowledgment

This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM-Transactions on Information Systems, 20(4), 357–389.

Ballesteros, L., & Croft, B.W. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the ACM SIGIR Conference (pp. 84–91). New York: ACM Press.

Braschler, M., & Peters, C. (2004). Methodology and metrics. In C. Peters, J. Gonzalo, M. Braschler, & M. Kluck (Eds.), Comparatice evaluation of multilingual information access systems. In Proceedings of the Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003). Lecture Notes in Computer Science, 2337, 7–20.

Braschler, M., & Schäuble, P. (2001). Experiments with the Eurospider retrieval system for CLEF 2000. In Proceedings of CLEF 2000. Lecture Notes in Computer Sciences, 3237, 7–20.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC-4 (pp. 25–48). Gaithersburg, MD: NIST.

Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In E.M. Voorhees & D.K. Harman (Eds.), TREC, experiment and evaluation in information retrieval (pp. 53–75). Cambridge: The MIT Press.

Chen, A., & Gey, F.C. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. IR Journal, 7(1–2), 149–182.

Chung, W. (2008). Web searching in a multilingual world. Communications of the ACM, 51(5), 32–40.

Harman, D.K. (2005). The TREC ad hoc experiments. In E.M. Voorhees, D.K. Harman (Eds.), TREC, experiment and evaluation in information retrieval (pp. 79–97). Cambridge: The MIT Press.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., & Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. IR Journal, 7(1–2), 99–119.

Hiemstra, D. (2000). Using language models for information retrieval. Unpublished doctoral dissertation, CTIT, University of Twente.

Hiemstra, D., Kraaij, W., Pohlmann, R., & Westerveld, T. (2001). Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In C. Peters (Ed.), Cross-language information retrieval and evaluation. Lecture Notes in Computer Science, 2069, 102–115.

Hull, D., & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the ACM SIGIR Conference (pp. 49–57). New York: ACM Press.

Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine large search engine transaction logs. Information In Processing & Management, 42(1), 248–263.

Kwok, K.L., Grunfeld L., Dinstl, N., & Chan, M. (2001). TREC-9 Cross-language, Web and question-answering track experiments using PIRCS. In Proceedings of TREC-9 (pp. 417–426). Gaithersburg, MD: NIST.

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, United Kingdom: Cambridge University Press.

McNamee, P., & Mayfield, J. (2004). Character N-gram tokenization for European language text retrieval. IR Journal, 7(1–2), 73–97.

Mel'èuk, I., & Wanner, L. (2006). Syntactic mismatches in machine translation. Machine Translation, 20, 81–138.

Nie, J.Y., & Simard, M. (2002). Using statistical translation models for bilingual IR. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems, Springer-Verlag, Lecture Notes in Computer Science, 2406, 137–150.

Nie, J.Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In Proceedings of the ACM SIGIR Conference (pp. 74–81). New York: ACM Press.

Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., & Oard, D.W., et al. (Eds.). (2007). Evaluation of multilingual and multi-modal information retrieval. Lecture Notes in Computer Science, 4730.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

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. (1997). Statistical inference in retrieval effectiveness evaluation. Information Processing & Management, 33(4), 495–512.

Savoy, J. (2003). Cross-language information retrieval: Experiments based on CLEF-2000 corpora. Information Processing & Management, 39(1), 75–115.

Sebastiani F. (2002). Machine learning in automatic text categorization. ACM Computing Survey, 14(1), 1–27.

Sheridan, P., & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using SPIDER system. In Proceedings of the ACM SIGIR Conference (pp. 58–65). New York: ACM Press.

Voorhees, E.M., & Harman, D.K. (2005). TREC, experiment and evaluation in information retrieval. Cambridge: The MIT Press.