DOI: 10.1145/1562764.1562799

BY JACQUES SAVOY AND LJILJANA DOLAMIC

How Effective is Google's Translation Service in Search?

IN MULTILINGUAL COUNTRIES (Canada, Hong Kong, India, among others) and large international organizations or companies (such as, WTO, European Parliament), and among Web users in general, accessing information written in other languages has become a real need (news, hotel or airline reservations, or government information, statistics). While some users are bilingual, others can read documents written in another language but cannot formulate a query to search it, or at least cannot provide reliable search terms in a form comparable to those found in the documents being searched. There are also many monolingual users who may want to retrieve documents in another language and then have them translated into their own language, either manually or automatically.

Translation services may however be too expensive, not readily accessible or not available within a short timeframe. On the other hand, many documents contain non-textual information such as images, videos and statistics that do not need translation and can be understood regardless of the language involved. In response to these needs and in order to make the Web universally available regardless of any language barriers, in May 2007 Google launched a translation service that now provides two-way online translation services mainly between English and 41 other languages, for example, Arabic, simplified and traditional Chinese, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish (http://translate.google.com/). Over the last few years other free Internet translation services have been made available as for example by BabelFish (http://babel.altavista.com/) or Yahoo! (http://babelfish.yahoo.com/). These two systems are similar to that used by Google, given they are based on technology developed by Systran, one of the earliest companies to develop machine translation. Also worth mentioning here is the Promt system (also known as Reverso, http://translation2.paralink.com/), which was developed in Russia to provide mainly translation between Russian and other languages.

The question we would like to address here is to what extent a translation service such as Google can produce adequate results in the language other than that being used to write the query. Although we will not evaluate translations *per se* we will test and analyze various systems in terms of their ability to retrieve items automatically based on a translated query. To be adequate, these tests must be done on a collection of documents written in one given language plus a series of topics (expressing user information needs) written in other languages, plus a series of relevance assessments (relevant documents for each topic).

Evaluation Campaigns

In an effort to promote information retrieval (IR) in languages other than English and also to evaluate bilingual searches (queries expressed in one language, documents retrieved in another), there have been various evaluation

campaigns conducted over the last few years. The first was the Text REtrieval Conference or TREC[3] in 1992, another took place in 1999 specifically for Far-East languages (the NTCIR series),[5] and beginning in 2000, CLEF[6] evaluation campaigns have been held for various European languages. The outcome of all these various international efforts was several test collections, created in various languages.

For our own tests and in an attempt to objectively evaluate Google's translation service, we used collections written in French and made up of articles published in the French newspaper *Le Monde* (1994 and 1995), plus others from the Swiss news agency (*ATS, Agence Télégraphique Suisse*) published during the same period. These collections were put together during six CLEF evaluation campaigns and contain a total of 177,452 documents (or about 487MB of data, See Table 1). On average each article contained about 178 content-bearing terms (median: 126); not counting commonly occurring words such as "la," "de" or "et"). Typically, documents in this collection were represented by a short title plus one to four paragraphs of text.

These collections also contain 310 topics, each subdivided into a brief title (denoted as T), a full statement of their information need (called description or D), plus any background information that might help assess the topic (narrative or N). The topic titles consist of 2 or 3 words reflecting typical Web requests, and are represented by a set of capitalized keywords rather than a complete grammatical phrase. These topics cover various subjects (such as, "U.N./U.S. Invasion of Haiti," "Consumer Boycotts," "Lottery Winnings", "Tour de France Winner" or "James Bond Films"), along with both regional ("Swiss Referendums," "Corruption in French Politics") and international coverage ("Crime in New York," "Euthanasia").

Relevance judgments (correct answers) were supplied by human assessors throughout the various CLEF evaluation campaigns. For example, Topics #201 to #250 were created in 2004 and responses were to result from searches in the *Le Monde* (1995) and *ATS* (1995) collections, a subset representing 90,261 documents. Of the 50 queries originally available in 2004, we found that only 49 having at least one correct answer.

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

	2001	2002	2003	2004	2005	2006
Source	Le Monde 94 ATS 94	Le Monde 94 ATS 94	Le Monde 94 ATS 94-95	Le Monde 95 ATS 95	Le Monde 94-95 ATS 94-95	Le Monde 94-95 ATS 94-95
Size	243 MB	243 MB	331 MB	244 MB	487 MB	487 MB
# docs	87,191	87,191	129,806	90,261	177,452	177,452
# topics	49	50	52	49	50	49
Topics	#41 - #90	#91 - #140	#141 - #200	#201 - #250	#251 - #300	#301 - #350

In all, 11 queries were removed because they did not have any relevant information, meaning only 299 (310 minus 11) topics were used in our evaluation. Upon an inspection of these relevance assessments, the average number of correct responses for each topic was 30.57 (median: 16), with Topic #316 ("Strikes") obtaining the greatest number of correct responses (521).

Information Retrieval Models

To search for pertinent items within this corpus, we used a vector-space model based on the classical *tf idf* scheme.[1] In this case the weight attached to each indexing term in the document (or in the query) was the product of the term occurrence frequency (or *tf*) and the inverse of the document frequency (or *idf*). Based on this formula, greater importance is attached to terms occurring frequently in the document (*tf* component), and in relatively few different documents (*idf* component).

We also applied the Okapi probabilistic model[7] in which a term's weight also depends on its discriminating power (the fact that this term occurs mainly in the relevant or non-relevant items) and on document length (weights attached to longer items are reduced).

Finally, we also applied an approach based on a statistical language model (LM),[4] which tries to estimate the occurrence probability of words, or in more sophisticated models, sequences of two words. In our experiments, the underlying estimates were based on a linear combination of occurrence frequencies both within the document and within the entire corpus.

To measure the retrieval performance obtained with these three IR models, we adopted a method known as the mean reciprocal rank (MRR).[2] For any given query, r is the rank of the first relevant document retrieved and the query performance is computed as $1/r$ or the reciprocal rank (RR). This value varies between 1 (the first retrieved item is relevant) and 0 (no correct response among the top 1,000 documents). It should be noted here that ranking the first relevant item in second place instead of first would seriously reduce the RR value, making it 0.5 instead of 1. Similarly, ranking the first relevant item in the 20th position (0.05) or lower would produce a very small RR. To measure the retrieval performance resulting from several queries, we simply computed the mean over all the queries. This value served as a measure of any given search engine's ability to extract one correct answer and list it among the top-ranked items. We thus believe that MRR value closely reflects the expectation of those internet surfers who are looking for a single good response to their requests.

In IR, not only do we want to measure a search system's ability to rank one relevant item, but also to extract all relevant information from the collection.[2] Users want both high precision (fraction of retrieved items that are relevant) and high recall (fraction of relevant items that have been retrieved). In other words they want "the truth, only the truth (precision), and nothing but the truth (recall)." To meet this need we compute the average precision for each query by measuring the precision achieved at each relevant item extracted and then computing an overall average. Then for a given set of queries we calculate the mean average precision (MAP), which varies between 0.0 (no relevant items found) and 1.0 (all relevant items always appear in the top of the ranked list). Higher MAP values are thus more difficult to obtain than higher MRR values, due to the fact that the MAP accounts for the rank of all relevant items, and not just the first one.

Using the mean to measure a sys-

tem's performance signifies that equal importance has been attached to all queries. Comparisons between two different IR strategies would therefore not be based on a single query but rather demonstrates that a single IR approach should not be rejected. Our approach is thus based on the importance of conducting experiments involving a large number of observations (in this study there were 299).

Finally, in an effort to statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology[8] in our statistical tests. With this method the null hypothesis H0 stated that both retrieval schemes produced similar MRR (or MAP) performance, and the null hypothesis would be accepted if two retrieval schemes returned statistically similar retrieval performance, otherwise it would be rejected. In this study our experiments detected statistical significant differences by applying a two-sided non-parametric bootstrap test (significance level = 5%).

Evaluation of Monolingual and English to French Searches

To define a baseline, we tested three IR models by submitting queries to search our corpus using the 299 topics written in the French language. The resulting MRR for topic titles only are depicted in the second column of Table 2 (labeled "Monolingual") and the corresponding MAP in the fourth column. We then took this value as a baseline and compared its retrieval effectiveness with other search models, while applying the same conditions. For both MRR and MAP, the Okapi model always provided the best retrieval results, and these results were significantly better than that of other search approaches.

In a second experiment, we took the English language topics and had them translated into French using Google's translation service, and then searched the French corpus with the translated topics. Through applying these three IR models, our MRR evaluations produced the results shown in the third column of Table 2 (labeled "From EN") or in the fifth column when using the MAP as retrieval effectiveness measures. In all cases, the Okapi approach performed significantly better than did the two other IR models.

When comparing original with translated topics, the performances decreased due to the automatic translation process. For the MRR, this difference was around 12% when using the Okapi search model (0.6631 vs. 0.5817) while with the MAP, this difference was slightly larger (0.4008 vs. 0.3408, or -15% in relative value). Taking the column labeled "Monolingual" as the baseline, retrieval performance differences for the translated queries are always statistically significant for both the MRR or the MAP, and for all three retrieval models.

Although we know that the mean is a useful method for representing an entire distribution of observations, it may hide certain underlying irregularities. An inspection of the MRR performance obtained using the Okapi model for monolingual queries shows that out of 299 cases, 166 (55.5%) ranked the first relevant document highest, while for English queries this value was lower (142 queries or 47.5%). Second, a count of the number of queries ranking a good response among the top five shows that there were 241 monolingual vs. 213 English queries. A count of the number of hard queries (those having no relevant document ranked among the top twenty) shows that when comparing monolingual 30 vs. 60 with English queries, there was a relatively large difference. Clearly the automatic translation was not perfect and thus the retrieval quality had been decreased.

The good news was that when using the Google's translation tool to search a French corpus based on English queries, the performance difference was not large (-12%) when compared to the original French queries. There are several possible explanations for this finding. First, the two languages are related with many words have similar meanings and some even the same spelling (such as, "soldiers" and "soldats", "success" and "succès", "quota", "immigration"). Proper names also have comparable spellings (such as, "Clinton", "Israel", "Airbus", "Bosnia" vs. "Bosnie", "Iraq" vs. "Irak", "Alps" vs. "Alpes"). As an extreme example, Topic #280 appears the same in both languages ("Crime in New York" and "Crimes à New-York"). Secondly, acronyms tend to be well translated by Google (such as, "UN" into "ONU", "EU" into "UE", "US" into "USA"). In certain cases English topics even improved the RR performance, such as with Topic #117 "European Parliament Elections" which is translated as "Élections du Parlement européen", while the original form is "Elections parlementaires européennes". This latter version is more readable in French but includes two adjectives and only one noun ("élections"). For this query the IR system did not choose the same stem for the noun "parlement" and the adjective "parlementaires" and thus the translated query provided better retrieval performance.

Generally speaking a translated topic does not perform as well as the corresponding original French topic, and based on our experiments with the Google's translation service, there are three main reasons for this. First a word's semantic coverage may differ from one language to the other. For example, in Topic # 113 "European Cup", the word "cup" was translated into the French "tasse" (in the sense of "coffee cup") instead of "coupe" (the winner's trophy). As another example, the word "court" in Topic #75 "Euskirchen Court Massacre" could be translated into "tribunal" or "cour" in French. For this search the most efficient term was "tribunal", which in French is used more frequently than "cour." These examples demonstrate that Google tends to provide the same translation, regardless of the context. As another example, if we ask Google to translate "the ink is in the pen" or "the pig is in the pen", the term "pen" would always be translated into French as "stylo", an

Table 2. Mean reciprocal rank (MRR) and mean average precision (MAP) for both monolingual and bilingual searches (299 title-only queries)

	MRR		MAP	
	Monolingual	From EN	Monolingual	From EN
Okapi	**0.6631**	**0.5817**	**0.4008**	**0.3408**
Language Model	0.5948	0.5093	0.3647	0.3085
tf . idf	0.5072	0.3895	0.2591	0.2091

instrument for writing.

Second, Google is case sensitive and thus it distinguishes between uppercase and lowercase. For example a request for "made in turkey" and "Made in Turkey" would not return the same results when translated into French. In the first case Google selects the animal and in the second the country name. In some topics however Google may incorrectly tag certain terms beginning with an uppercase letter. With Topic #192 "Russian TV Director Murder" for example, the system assumes "Murder" is a personal name and thus does not translate it into French ("Directeur russe Murder de TV" vs. "Assassinat d'un directeur de la télévision russe"). The fact that words appearing in topic titles beginning with an uppercase letter may thus induce error into the translation system, causing it to wrongly assume that a proper name is present. A similar case occurs with Topic #244 "Footballer of the Year 1994" in which the term "Footballer is tagged as a proper name, or as a word not appearing in the dictionary. In this case therefore the translation into French contains a spelling error.

Third, when idioms or other compound terms are written with a hyphen, Google and other automatic translation tools tend to produce a word-by-word translation. With Topic #261 "Fortune-telling" for example the proposed translation "Fortune-dire" (with to tell = "dire") is far from being the correct translation ("Diseurs de bonne aventure"). Again, in the case of certain idiomatic expressions (such as, "from the horse's mouth"), incorrect translations could occur when using Google or other automatic translation tools.

Using Other Translation Resources

The evaluations and explanations mentioned above are limited to the Google translation service and also to very short query formulations pertaining to a limited number of topic titles. In fact, during the last few years other freely available machine-based translation services have become available. We thus decided to compare performances achieved by the Google translation service (limited to the Okapi model), with the alternative translation systems Babelfish and Promt, when automatically translating English topics into French.

Table 3. Mean reciprocal rank (MRR) for title (T) and title & descriptive (TD) topics using monolingual and bilingual searches (Okapi, 299 queries)

	T Query	TD Query
Monolingual	**0.6631**	**0.7360**
Google	0.5817	0.6551
Babelfish	0.5653	0.6426
Promt	0.5704	0.6457

The resulting MRR values are listed in Table 3 and display a larger query construction. This combination includes the title and descriptive (TD) sections of the topic formulation, mandatory during the CLEF evaluation campaigns.[6] Although the title is sometimes ambiguous, the descriptive part may help the translation system by providing a complete sentence and context, both being useful in the automatic translation process. For example, Topic #91 is titled "AI in Latin America" and its descriptive section consists of the following "Amnesty International reports on human rights in Latin America." This description indicates that the acronym AI does not mean "Artificial Intelligence." Adding the descriptive part increases the mean query length to 10.78 content-bearing terms, when with the title section is limited to 2.86 content-bearing terms.

The data in Table 3 shows that the performance difference between the three translation tools are small, around 1% to 3%. For example, using the title-only topics the Google translation system produces an MRR of 0.5817 vs. 0.5704, or -1.9% in relative value for the Prompt system. Using the performance obtained by Google as baseline, we did not find any statistically significant difference when compared to other translation resources. Note however that the performance difference between the monolingual (second row in Table 3) and the three query translation approaches are always statistically significant and in favor of the monolingual search. As mentioned previously, we knew that both the Babelfish and Google systems are based on the same translation technology. When inspecting the MRR achieved by the title-only query formulation, we found that performances were different for only 27 queries out of 299 when comparing the Google and Babelfish translation services. When comparing the Promt and Google translated queries, the retrieval performance was different for 117 queries.

Evaluation of German to French Searches

We decided that the previous findings should be compared to another language, and thus we selected German for the query source language. Using the Google translation tool we automatically translated the queries into French. As shown in Table 4 under the column labeled "From DE", when compared to monolingual searches retrieval performances were shown to decrease significantly. In mean, the relative difference was around 30%, and there was a statistically significant performance difference between queries written in German and those written in French.

An inspection of the Google translation results for German shows that poor retrieval performances are for the most part caused by the factors cited above, and also by the inadequate processing of German compound words. Such linguistic constructions also occur in English (such as, viewpoint, handgun) but in German they are more frequent, and also occur in various forms (such as, "Friedensnobelpreis" = "Frieden" (peace) + "Nobel" + "Preis" (prize) or "Nobelpreis für den Frieden"). The fact that many German compound words were not translated had a very real impact on retrieval performance. For the topics written in French, we found that only 16 queries without having a correct answer ranked among the top 50 retrieved items while for German this value increased to 61.

As a final experiment, we used the queries written in German and then automatically translated them into English, and from this pivot language we translated them into French. This

Table 4. Mean reciprocal rank (MRR) for both monolingual and bilingual searches (Title-only queries)

	Monolingual	From EN	From DE	From DE-EN
Okapi	0.6631	0.5817	0.4631	0.5273
Difference %		-12.3%	-30.2%	-20.5%

evaluation thus reflects commonly occurring situations in which one language is defined as a pivot language (*interlingua*) and serves as an intermediary between all possible language pairs. There are several advantages to using this translation strategy. For direct translations, n languages would require $n \cdot (n\text{-}1)$ possible translation services. In the European Union with its 23 official languages, this means that $23 \cdot 22 = 506$ possibilities would have to be covered. Thus, instead of a direct translation for all possible language pairs we can limit the resources to $2 \cdot (n\text{-}1)$ translation pairs (or 44 in our European example), namely $(n\text{-}1)$ from all languages to the pivot language, and $(n\text{-}1)$ from the pivot language to all the others.

As shown in Table 4, with the Okapi model the retrieval performance obtained was 0.5273, resulting in a mean performance significantly lower than that of the monolingual search (0.6631) but higher than the direct translation from German (0.4631). In an effort to explain this better performance when English was selected as the pivot language, we found that translation from German to English was better than from German to French. For example, Topic #235 "Seal-hunting" is written as a compound in German ("Robbenjagd" = "Robben"(seals) + "Jagd" (hunting)) which is correctly translated into English ("Seal hunting") but not into French ("Robbenjagd"). These experiments therefore demonstrate that query translation may be effective for some language pairs yet with other language pairs certain problems may be encountered, even when using the same translation system. Moreover, compared to direct translation, the pivot language approach does not always imply less effective translation performance.

Conclusion

Writing a topic in another language and then asking Google to automatically translate it before launching a search degrades retrieval effectiveness, compared to a monolingual search in which requests and documents are written in the same language. As revealed in our evaluations based on short topic formulations, retrieval performance reductions are not always impressive (see Table 4). Applying the Google translation tool to automatically translate an English topic into French may achieve retrieval effectiveness of around 88% compared to a corresponding monolingual search. From another perspective, a monolingual search provides at least one relevant item among the first five retrieved items for 241 queries out of 299 (or 80.6%). Using the English topics and using Google to translate them into French will place a relevant item in the top five for 213 queries (or 71.2%). Clearly, in mean, a translated query may retrieve the needed information.

Using another translation service should allow us to obtain similar retrieval performance. For example, adopting the Babelfish that Yahoo! uses, 206 queries (or 68.9%) would find at least one good answer ranked among the top five, while for the Promt translation tool this number would be 212 (or 70.9%). Changing the language pairs may however degrade retrieval effectiveness. For example, using topics written in German instead of English clearly hinders retrieval performance by around 30% compared to a monolingual search (see Table 4). An inspection of the first five retrieved items among the German topics automatically translated into French shows that at least one pertinent item would be retrieved from only 174 queries out of 299 (or 58.2%). For some language pairs, the mean result obtained is around 10% lower than that of a monolingual search while for other pairs, the retrieval performance is clearly lower. In this study, we have investigated three important languages from an economic point of view, but automatic translation resources are not available for all language pairs, particularly for languages used by small numbers of users and having only modest economic importance.

For all search systems there are difficult queries for which the search engine encounters difficulties to find at least one relevant answer. These queries typically contain concepts expressed in an ambiguous way or use vocabulary that leads to incorrect identification of relevant and non-relevant items, and when adding a translation stage this phenomenon seems to increase. In our experiments for example we found 30 title-only queries for which a monolingual search was not able to extract any relevant items in the first 20 responses. With English topics and the Google translation system however this number increased to 60. Through making use of other freely available translation services, we obtained similar results (56 queries with Promt or 64 with Babelfish). C

References

1. Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press, 1999.
2. Buckley, C., and Voorhees, E. Retrieval system evaluation. In *TREC, Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005, 53-75.
3. Harman, D.K. The TREC ad hoc experiments. *TREC, Experiment and Evaluation in Information Retrieval*. MIT Press, 2005, 79-97.
4. Hiemstra, D. Term-specific smoothing for the language modeling approach to information retrieval. In *Proceedings of the ACM-SIGIR*. ACM Press, 2002, 35-41.
5. Noriko, K., (Ed) NTCIR Workshop 6 Meeting. National Institute of Informatics, Tokyo, 2007.
6. Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., and Stempfhuber, M. (Eds) Evaluation of multilingual and multi-modal information retrieval. Lecture Notes in Computer Science #4730. Spinger-Verlag, Berlin, 2007.
7. Robertson, S.E., Walker, S., and Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management 36*, 1, (2000), 95-108.
8. Savoy, J. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management 33, 4*, (1997), 495-512.

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

Jacques Savoy (Jacques.Savoy@unine.ch) is a professor in the Computer Science Department at the University of Neuchatel, Switzerland.

Ljiljana Dolamic (Ljiljana.Dolamic@unine.ch) is Ph.D. student in the Computer Science Department at the University of Neuchatel, Switzerland.