A Comparative Study of Online Translation Services for Cross Language Information Retrieval

Ali Hosseinzadeh Vahid, Piyush Arora, Qun Liu, Gareth J. F. Jones
ADAPT Centre / CNGL
School of Computing
Dublin City University
Dublin 9, Ireland
{avahid,parora,qliu,gjones}@computing.dcu.ie

ABSTRACT

Technical advances and its increasing availability, mean that Machine Translation (MT) is now widely used for the translation of search queries in multilingual search tasks. A number of free-to-use high-quality online MT systems are now available and, although imperfect in their translation behaviour, are found to produce good performance in Cross-Language Information Retrieval (CLIR) applications. Users of these MT systems in CLIR tasks generally assume that they all behave similarly in CLIR applications, and the choice of MT system is often made on the basis of convenience. We present a set of experiments which compare the impact of applying two of the best known online systems, Google and Bing translation, for query translation across multiple language pairs and for two very different CLIR tasks. Our experiments show that the MT systems perform differently on average for different tasks and language pairs, but more significantly for different individual queries. We examine the differing translation behaviour of these tools and seek to draw conclusions in terms of their suitability for use in different settings.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Search process; I.2.7 [**Natural Language Processing**]: Machine Translation

Keywords

Cross language information retrieval; machine translation; variable retrieval effectiveness

1. INTRODUCTION

The World Wide Web is increasingly polyglot in nature. While in its early days it was dominated by English content, according to recent statistics, the proportion of Arabic content on the web is now increasing 10 times faster than English, and only 29% of web users are English speakers. In order for users to access the maximum amount of information, search technologies need to handle content written in different languages[1], and to facilitate entry of queries to search for content in multiple languages. Enabling search using queries in one language to find content is another one has for many years been the focus of Cross-Language Information Retrieval (CLIR) research.

Much research has concentrated on translation tools for CLIR to cross the language barrier between queries and documents. However, in recent years, Machine Translation (MT) has become increasingly popular as the default translation option in CLIR. This trend has been strongly influenced by the increasing availability of high quality free online MT services, such as Google translate[2] and Microsoft Bing translator[3]. These online MT tools are often used as black boxes to provide translation in CLIR evaluation campaigns. This is perhaps not surprising since they generally provide high quality translation of search queries which produce high retrieval effectiveness in CLIR, often close to that of monolingual IR, without requiring any development cost. Users of these freely available MT systems in CLIR tasks generally assume that they all behave similarly in CLIR applications, and the choice of system is often made on the merely basis of convenience.

In this paper, we compare and analyze the performance of two well-known and popular freely available MT services: Google translate and Bing translate, for query translation across multiple language pairs and for two very different CLIR tasks. The results of our experiments show that although on average the MT approach usually provides a high quality translation for queries that consequently leads to high retrieval effectiveness in CLIR, the different MT systems can result in quite different retrieval behaviour for individual queries for different language pairs and applications. In particular we show that choosing an MT system with respect to handling Out-Of-Vocabulary (OOV) terms, Multi-Word Expression (MWE) extraction and translation, and Named entity translation and transliteration issues, can yield statistically significant improvements in mean average precision of CLIR system. Note we cannot expect to be able to observe these behavioural differences for CLIR from

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743008.

[1] http://www.internetworldstats.com/stats7.htm

[2] https://translate.google.com/

[3] http://www.bing.com/translator/

measurements of MT effectiveness since MT focuses on generating translations that are semantically, morphologically, and syntactically correct, while IR focuses on retrieving documents that match the query on the conceptual level regardless of the surface form of words.

The remainder of the paper is organized as follows: section 2 reviews previous work on the comparison of translation services based on MT and IR metrics, section 3 describes the design of our experimental test sets in which two free online translation services are compared in two different CLIR task environment, we discuss the results of our experiments in section 4, and conclusions of our finding and some ideas for the extension of this study are discussed in section 5.

2. RELATED WORK

MT has actually been used for both query and document translation since the early years of CLIR [5]. While in the early years much CLIR research focused on the development of translation resources and methods specifically focused on the CLIR task, the rapid advances in Statistical Machine Translation (SMT) techniques in recent years means that it has played an increasing role in the improvement of CLIR systems [10], [4], [8].

The easy availability of online translation services such as Google and Bing translate has encouraged researchers in CLIR domain to investigate the potential of these tools in CLIR applications. Chen and Bao [2] evaluated the performance of the online Google SMT service and the rule-based SYSTRAN MT system for Title (short sentence) and Description (long sentence) queries, as compared with a human translator and found that although Google translate, an SMT system, worked well for short queries, SYSTRAN MT achieved better results for the long queries. Zhang et al. [14] carried out a comparative case study of Google and Bing translators focused on five different aspects: sentence order, separation of semantic groups, choice of polysemous words, sentences with partial negation and attributive clauses. They carried out their experiments with the Chinese-English language pair. They reported that Google translator was better at translating phrases and semantic groups, but that human modification was still required after MT. Surprisingly, the result of similar studies done by Dhakar et al.[3] applied on Hindi-English language pair reported that Bing translator performed better than Google translate. They compared the two systems based on different parameters: missing words, word order, incorrect words, unknown words and punctuation errors.

Savoy and Dolamic [12] evaluated the effects of Google's online translation service on mean average precision (MAP) and precision at rank 5 cutoff for a CLIR system with French-English and German-English language pairs. They showed that on average, a translated query may retrieve the relevant documents. However, they also illustrated that similar to monolingual IR, there are difficult queries in CLIR for which the systems are unable to find even one relevant answer.

In parallel to the black box use of MT systems for CLIR tasks, other studies have focused on exploring novel MT methods specifically designed for CLIR tasks [9][13] . While this line of enquiry is clearly relevant to future developments in CLIR, in this paper we focus on exploring the behaviour of standard MT systems as used in the majority of current work on CLIR.

3. EXPERIMENT DESIGN

To evaluate and compare the behaviour of the two most well-known freely available translation services, Google translate and Microsoft Bing translator for CLIR, we carried out two set of experiments: one on retrieving English news stories with queries in six different languages and the other on retrieving similar Hindi news stories with English news stories as queries. For detailed comparison, we carried out query-specific analysis on some systems. This section describes the experimental setup for the study including the selection of the test collection and IR system.

3.1 Test collection and IR system

3.1.1 CLEF CLIR ad hoc news retrieval

For the first experiment, we used test collections created by the Cross Language Evaluation Forum (CLEF) in 2000 for retrieving English news stories with queries in six different languages. The collection contains 113,005 English news articles from the *LA Times* in 1994, 33 topics in different languages, and binary relevance judgments created using a pooled assessment methodology. We removed English terms contained in the stopword list provided with the open source Terrier IR Platform[4] from the document collection and performed Porter stemming using the same tool that we used for processing the document collection. We then created a document index based on stemmed English terms. We formulated queries using the title (T) and description (D) fields, denoted TD queries which formed the standard query form for the original task. The original query was applied to the online MT system for the appropriate language pair[5], and then performed post-translation stopword removal using the same stopword list provided for documents. All our experiments were run using Terrier 4.0, based on BM25 weights. In the BM25 formula [11], we used k1 = 1.2, b = 0.75, and k3 = 7 as has been commonly used.

3.1.2 FIRE CL!NSS task

For the second set of experiments, we report experiments carried out for the Cross Language !ndian News Story Search (CL!NSS) task at the Forum for Information Retrieval Evaluation (FIRE) [1]. The CL!NSS task [6] is an edition of the PAN@FIRE task which focuses on addressing news story linking between English and Indian languages. In this experiment, English news stories were used as queries to retrieve similar documents from Hindi news story collection. The target documents were 50,691 news documents in the Hindi language with three main fields: title of the news document, date when the news was published and the content of the news article. We used two query sets from the same task, CL!NSS 2012 and CL!NSS 2013, that have 50 and 25 news stories in English respectively. Each document in the query dataset also has the same three fields as the data collection. We used the open source Lucene 4.4.0[6] inbuilt Hindi Analyzer for stopword removal and stemming over the documents. A stopword list was obtained by concatenating different standard stopwords list for the Hindi language: i) the FIRE Hindi stopword list[7], ii) the Lucene internal stop-

[4] http://terrier.org/

[5] All our translation were got through Online MT systems between 12/01/2015 and 16/01/2015.

[6] http://lucene.apache.org/core/

[7] http://www.isical.ac.in/~fire/resources.html

		MAP	P@5	Rel_Ret
Monolingual		0.3739	0.4061	549
Italian	Google	0.3468	0.4	519
	Bing	0.3556	0.4061	510
Spanish	Google	**0.3578**	**0.4182**	**534**
	Bing	0.3256	0.3697	510
German	Google	0.3558	0.3939	505
	Bing	0.353	0.4061	509
Finnish	Google	0.3217	0.3879	514
	Bing	0.3354	0.4061	511
Swedish	Google	0.3428	0.4121	531
	Bing	**0.3673**	**0.4182**	**532**
Dutch	Google	0.3597	0.4303	536
	Bing	0.3454	0.4182	524

Table 1: Comparison of online MT system's impact on different CLIR performance of CLEF task

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
2012 Query Set				
Translation				
Bing	0.520	0.477	0.498	0.514
Google	**0.581**	**0.518**	**0.521**	**0.549**
Translation + Named Entities Transliteration				
Bing	0.469	0.495	0.508	0.523
Google	**0.584**	**0.523**	**0.529**	**0.556**
2013 Query Set				
Translation				
Bing	**0.780**	**0.734**	**0.748**	**0.747**
Google	0.760	0.673	0.689	0.691
Translation + Named Entities Transliteration				
Bing	**0.780**	**0.736**	**0.749**	**0.751**
Google	0.760	0.673	0.689	0.691

Table 2: Comparison of online MT system's impact on CLIR performance for the CL!NSS 2012 and 2013 query sets

word list, and iii) a stopword list created by selecting all the words with document frequency (DF) greater than 5,000 in the target document collection. For queries, we applied the original query to the translator, and then again performed post-translation stopword removal using the same stopword list as used for the documents. We used Lucene's default scoring function[8] for these experiments which is a variant of a standard TF-IDF function. We observed that the Hindi target documents contained both words in the translated and transliterated forms of input queries as shown in Table 3. The use of the translated or transliterated forms in the documents was not predictable, and thus we performed transliteration of named entities using Google transliteration[9].

To find conditions or errors in translations which cause a decrease in CLIR system performance, We also performed query performance analysis for both experiments.

4. RESULTS AND ANALYSIS

Named Entities		
English Word	Translated Word	Transliterated Word
Commonwealth	राष्ट्रमंडल	कामनवेल्थ
Games	खेल	गेम्स
OOV Words		
Ex Dantewadas,	Ichapuram,	Thipsay
Transliterations		
”LTTE”	एलटीटीई	लिट्टे
”PLGA”	प्लग	पीएलजीए

Table 3: Examples of errors in translation for English-Hindi FIRE experiments

In this section, we report our experimental results for the CLEF and FIRE test sets. It is quite common in CLIR evaluation to compare the effectiveness of a CLIR system against a monolingual baseline where query translation is not required. For the CLEF task this was easy to obtain for the IR test system described in section 3.1, using the parallel English language version of the queries provided in the test set. In order to check that our system was comparable to that used in previous work using these test collection, we compared our monolingual baseline with than reported for the best performing previously published results and found them to be comparable. The FIRE test collection does not include a parallel Hindi version of the test news stories, and thus we were not able to carry out monolingual runs for this task[10].

Table 1 shows results for the CLEF test collection on six different European languages, including details of the number of relevant-retrieved documents, mean average precision (MAP), and precision at rank cutoff 5 (P@5). The results show that performance of CLIR systems in all language pairs (regardless to translation service) is more than 85% of monolingual baseline. For the Spanish-English language pair, queries translated by Google get better results in all evaluated metrics. However the system using queries translated from Swedish to English using Bing outperformed the system with translated queries by Google. Meanwhile, there is no significant difference in performance of CLIR systems based on MAP and P@5 for other language pairs.

Table 2 shows results of our experimental runs on the FIRE test set. Because of the task requirement, we report our results using the NDCG measure. We show results with respect to the translation of queries using both Bing and Google translation system and adding named entities translation to the translated query to handle the cases where translation fails. Surprisingly, we observe that for the experiments carried out on the 2012 query set, the CLIR system using Google translation outperforms the CLIR system using Bing translation. However, for the 2013 query set, the behaviour is reversed, where system results based on Bing translation outperform system results based on Google translation. Adding named entity transliteration always improves the results apart from 2013 query set using Google translate, where the results are exactly similar to using just the translation of queries.

For a more detailed comparison, we show query-specific analysis on the CLEF test set for the Spanish-English and

[8] http://ipl.cs.aueb.gr/stougiannis/default.html

[9] http://www.google.com/inputtools/

[10] Neither were done by original participants in this task.

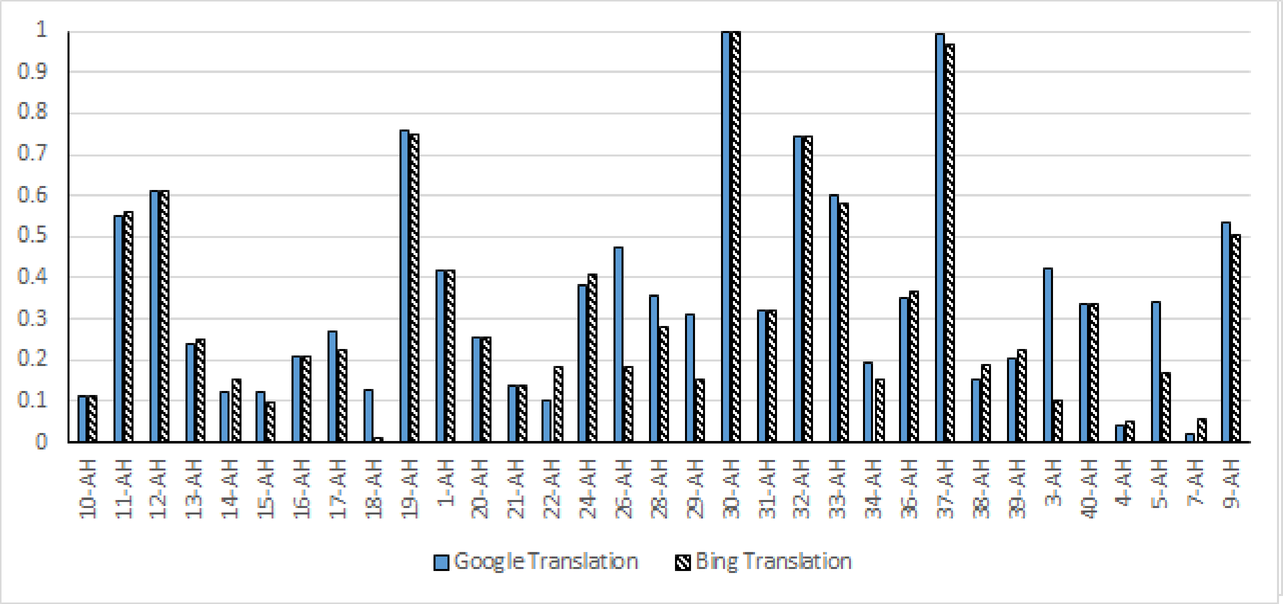

Figure 1: Impact of query translation using Google and Bing MT services on CLIR performance (Spanish-English)

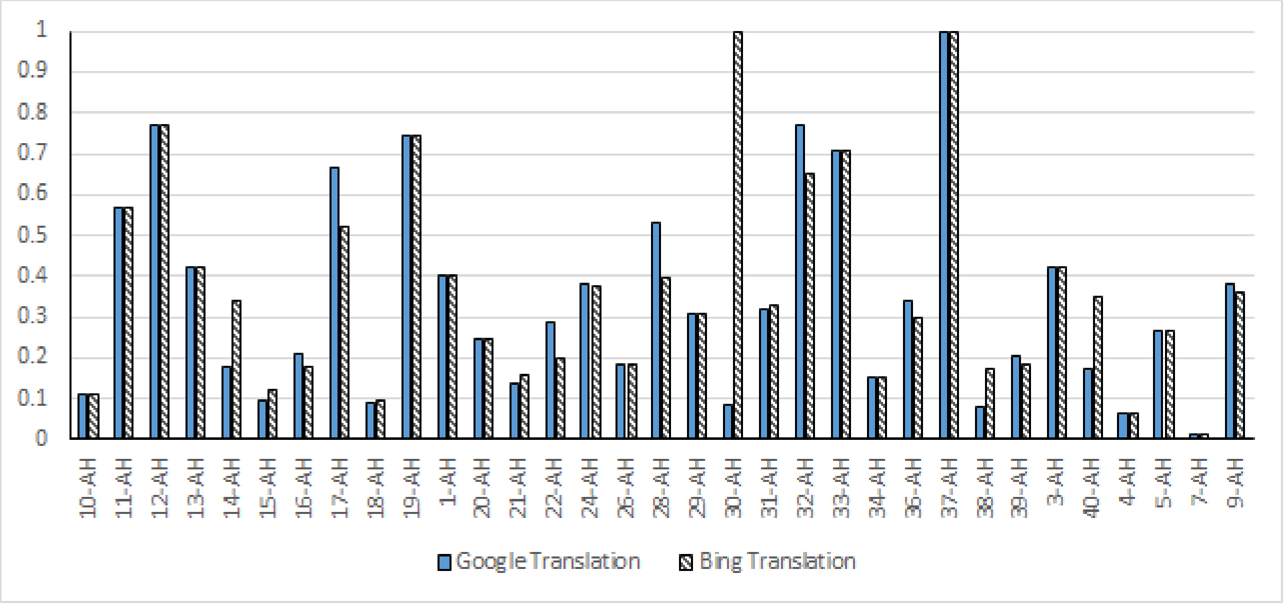

Figure 2: Impact of query translation using Google and Bing MT services on CLIR performance (Swedish-English)

Query ID	Original Spanish Query	Google Translation	Bing Translation
3-AH	Drogas en Holanda	**Holland** Drug	Drugs in the **Netherlands**
5-AH	Ingreso en la Unión Europea	**Join** in the European Union	**Income** in the European Union
18-AH	Bajas entre bomberos	Casualties among **firefighters**	**Fire** casualties
26-AH	Uso de la energía eólica	Use of wind **energy**	Use of wind **power**
29-AH	Premio Nobel de Economía	Nobel Prize in **Economics**	Nobel Prize **winner**
14-AH	Turismo en E.E.U.U.	Tourism **E.E.U.U.**	Tourism in **USA**
22-AH	Accidentes de aviones en pista	**Aircraft** accidents on track	**Planes** in track crashes

Table 4: Examples of translation problems in Spanish-English CLEF experiments

Swedish-English language pairs. Figure 1 shows comparison on a per query basis for Spanish queries. For most of queries (19 out of 33), MAP for both translation services is similar with negligible (less than 10%) difference. Bing translator provides better results for 5 of the 33 queries, while for the remaining 9 queries, Google translate gives the best results. Figure 2 shows query-by-query results for Swedish-English CLIR. Similar to the Spanish-English CLIR system, there are some queries where both translations give similar performance (21 out of 33). Of the remaining 12 queries, for half of them Google translate provides more effective performance, while for other half Bing translations work better. Since the results for queries translated by Bing are considerably better, the final performance of the CLIR system using Bing translations outperforms the one using Google translate.

In Table 3, we show examples of different translation errors for the CL!NSS task. Transliteration of named entities appears to be useful for English-Hindi cross language search, with improvements for both 2012 and 2013 query sets. We observe that news documents in the Hindi language can contain both translated and transliterated form of most frequent named entities. Words such as Commonwealth and Games as shown in Table 3 have both translated and transliterated forms in the target news documents. However, automatic transliteration can result in inappropriate matches or failures to match. Transliteration is sometimes a complex task. There might be different representations for the same transliterated word based on its pronunciation as shown in Table 3. For example, the abbreviation LTTE has two possible transliterations: एलटीटीई and लिट्टे both of which are valid and frequently used. On the other hand, Google transliteration provides some errors as it fails to handle spelling variations in the Hindi language and maps characters wrongly. For example, PLGA is transliterated as स्नग by Google transliteration where the actual transliteration is पीएलजीए. Certain challenges remain unexplored in our current study. For example, abbreviations such as "MNIK","YSR", movie names and political party names should be handled in a systematic way. In addition, handling spelling variants is a significant challenge. Stemming takes care of the affixes. However the main problem for Hindi arises with handling the diacritic marks and vowel variations. With better normalization techniques, we would be able to handle the erroneous cases and capture the missing information.

Table 4 shows some examples of such queries in Spanish and compares their translation to find conditions or errors in translation for our experiments on CLEF test collection. The top part of the table includes queries for which translations using Google result in better IR results, while the other part comprises queries for which their translation using Bing achieved better IR results. For further insight, we considered queries where different translations affect CLIR system performance for the experiments carried out on the CLEF 2000 data set. We observe that the CLIR results with respect to Google and Bing translation vary depending on the language pair. The reason for these results relates to be the output quality of the translation service, where differences in the translation result in differences in the vocabulary and coverage of the translation system for a language pair. As shown in Table 4, differences in translation eventuate more distinguishable effectiveness in IR performance of 3-AH, 5-AH, 18-AH and 22-AH queries. The Spanish query word *"holanda"* in query 3-AH which is translated as *"Holland"* by Google which achieves 4 times greater IR effectiveness compared to its translation by Bing as *"Netherlands"*. The same reason causes twice the effectiveness with Google translate for query 5-AH where the Spanish word *"Ingreso"* is translated to *"join"* by Google and *"income"* by Bing. Bing's translation of the Spanish term *"avions"* as *"planes"* has similar impact on IR performance for query 22-AH compared with its translation by Google as *"aircraft"*. On the other hand, since term *"E. E. U. U."* is not in the Spanish vocabulary list of Google, it cannot be translated correctly, but the same term can be translated by Bing as *"USA"* and affects IR system performance positively.

Another possible reason of these differences comes from various models of multi-word expression extraction and translation in different MT systems. For example, while Bing translates the Spanish bi-gram *"energiá eoćlica"* as *"wind power"*, it translates Spanish term *"energiá"* as *"energy"*. This causes IR performance to decrease for query 26-AH. The same reason results in Bing translate losing its effectiveness for IR performance on query 18-Ah where it translates the Spanish term *"bombreos"* as *"fire"* in translation of the Spanish tri-gram phrase *"Bajas entre bomberos"*, while it translates *"bombreos"* individually as *"firefighter"* which is a very effective translation for that query. The same thing happens in query 29-AH where Bing could not translate Spanish term *"Economía"* inside a n-gram phrase *"Premio Nobel de Economía"* while it has the word *"economy"* as its translation in translation vocabulary list.

5. CONCLUSIONS

In this paper, we compared the CLIR effectiveness of different MT systems for CLIR tasks with different language pairs. Our experiments show that CLIR performance is affected by translation effectiveness in different language pairs but also in the same language pair for different query sets. Differences in the translation arise from differences in the vocabulary and coverage of the translation system for a language pair. The results thus indicate that it is often not possible to say that a particular MT is universally preferable for a particular CLIR task, but rather than the MT system for a multilingual access tasks on the web should be selected on the basis of a realistic document, query and relevance development collection.

Since it will apparently often not be possible to select one MT system which will always be the best for a particular task, we believe that some form of data fusion combining results of different translations would generally be more robust for CLIR performance than that of a single MT system. This is consistent with the original results obtained using different commercial MT systems for the CLEF 2001 where data fusion methods were found to improve CLIR effectiveness with less advanced MT system [7]. MT system research generally pays considerable attention to syntactic structure of this translated output. This is largely unimportant when translating queries for CLIR. Conversely, MT systems place less emphasis on the issues of translation of OOV words, which often have a significant impact on retrieval effectiveness. To boost the performance of the CLIR system, it is necessary to reach inside the translation black box and try to find parameters in MT systems to tune them for better CLIR results. Hence, applying similar experiments in different genres and domains could yield a better understanding of how to best define the factors of MT systems for CLIR and will be investigated during later experiments.

Acknowledgments

This research is supported by Science Foundation Ireland (SFI) as a part of the CNGL Centre for Global Intelligent Content at DCU (Grant No: 12/CE/I2267).

6. REFERENCES

[1] P. Arora, J. Foster, and G. J. F. Jones. DCU at FIRE 2013: Cross-Language !ndian News Story Search. In *Forum for Information Retrieval Evaluation (FIRE 2013)*, New Delhi, India, 2013.

[2] J. Chen and Y. Bao. Information access across languages on the Web: from search engines to digital libraries. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–14, 2009.

[3] B. S. Dhakar, S. K. Sinha, and K. K. Pandey. A survey of translation quality of English to Hindi online translation systems (Google and Bing). *International Journal of Scientific and Research Publications*, page 313, 2013.

[4] N. Ferro and C. Peters. CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks. Lecture Notes in Computer Science, Corfu, Greece, 2009. Springer.

[5] D. A. Gachot, E. Lange, and J. Yang. The Systran NLP browser: An application of machine translation technology in cross-language information retrieval. pages 105–118. Kluwer Academic Publishers, 1998.

[6] P. Gupta, P. Clough, P. Rosso, and M. Stevenson. PAN@FIRE: Overview of the cross-language !ndian news story search (CL!NSS) track. In *Forum for Information Retrieval Evaluation (FIRE 2012), ISI, Kolkata, India*, 2012.

[7] G. J. F. Jones and A. M. Lam-Adesina. Exeter at CLEF 2001: Experiments with machine translation for bilingual retrieval. In *Proceedings of CLEF 2001*, pages 59–77, Darmstadt, Germany, 2001.

[8] J. Leveling, D. Zhou, G. J. F. Jones, and V. Wade. Document Expansion, Query Translation and Language Modeling for Ad-Hoc IR. In *Proceedings of CLEF 2009*, Corfu, Greece, 2010. Springer.

[9] W. Magdy and G. J. F. Jones. Should MT systems be used as black boxes in CLIR? In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 683–686. Springer-Verlag, 2011.

[10] C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peas, and V. Petras. *CLEF 2009: Evaluating Systems for Multilingual and Multimodal Information Access*. Springer, Corfu, Greece, 2009.

[11] S. E. Robertson and K. S. Jones. *Simple, proven approaches to text retrieval*. Technical Report 356, University of Cambridge. Computer Laboratory, 1994.

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

[13] F. Ture, J. Lin, and D. W. Oard. Looking Inside the Box: Context-sensitive Translation for Cross-language Information Retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1105–1106. ACM, 2012.

[14] R. Zhang, Y. Pan, and Y. Yang. A comparative case study of Google and Bing translation. In *Proceedings of 5th International Conference of Education, Research and Innovation (ICERI-2012)*, pages 3669–3673, Madrid, Spain, 2012. IATED.