Query Language and Query Intent in Cross-Lingual Web Search

Ahmed Y. Tawfik and Ahmed S. Kamel
Microsoft Research
atawfik|a-ahmedk@microsoft.com

ABSTRACT

This paper examines query intent (or domain) classification as a proxy for user information need and its use in improving the performance of cross-lingual information retrieval (CLIR). Through several human relevance studies, information needs that cross language and culture barriers are identified.

Therefore, we propose to use query intent classifiers as a filter to pick queries that may benefit from CLIR. The analysis presented in this paper are based on experiments conducted on three populations with a high percentage of bilingual internet users: the Arabs, the French Canadians, and Spanish speakers in the US. While these users differ significantly in their usage of search engines, the effect of CLIR on the quality of the search experience for individual intents was largely similar. This leads us to expect a wider applicability of our findings on a global scale.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Cross-Lingual Information Retrieval

Keywords

Query Intents, Relevance study

1. INTRODUCTION

Cross-lingual information retrieval (CLIR), or more specifically Cross-lingual Web Search (CLWS), aspires to make content globally available to Internet users across linguistic barriers. CLWS is a variant of CLIR that serves results written in one or more language other than the language of the original issued query. Multilingual users can readily benefit from this content and monolingual users may use online translation services to gain some understanding of the content in a foreign language.

While English is the dominant content language on the web with more than half the global web content still in English, content in other languages has been growing rapidly. Over the last decade Arabic content grew from 0.2% to close to 0.8%, and content in other languages like Russian, French, and Polish, has also been growing faster than English content[1] However, there is a clear mismatch between the language that the web content is written in and the native languages of current Internet users. For example, more than half the web content is in English, while only a quarter of the users are native English speakers. On the other hand, languages like Chinese and Spanish are spoken by 24% and 8% of Internet users respectively [2] but less than 5% of the global web content is in each of these two languages. This disparity continues to be observed, to a lesser extent, when we consider the language distribution within the most popular 1000 sites. CLWS is more crucial for users seeking authoritative content (as opposed to user generated content), as in many domains of knowledge such content is predominantly in English.

In Section 2, we review some related work, then introduce our intent-based approach in Section 3. We discuss the experiments we performed in Section 4 and conclude with our findings in Section 5.

2. RELATED WORKS

Earlier experiments with CLWS showed that these systems tend to have a mean average precision (MAP) in the 60% to 70% range [6]. Therefore, it is important to carefully pick the queries that are highly likely to benefit from augmented multi-lingual results, otherwise these results may hurt relevance. The spectrum of user involvement in the decision of using CLWS ranges from user driven [5] to fully automated. First iterations on the query picking task in our group Framed the CLWS query picking problem as a supervised binary classification problem where the goal is to distinguish queries that are likely to benefit from the other queries, given an input feature set. The features consid-

[1]W3Tech Web Technology Survey. (2014). Usage of content languages for websites. from `http://w3techs.com/technologies/overview/content_language/all/`

[2]Internet World Stats. (2014). Top Ten Languages Used in the Web: Number of Internet Users by Language . Bogota, Columbia: Miniwatts Marketing Group.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIRIP'2014 Gold Coast, Australia
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

ered include features that try to predict the quality of the translation and others that try to predict the quality of the results. To predict the quality of the translation, relevant features include: the relative difference in length between the original query Q_O and the translated query Q_T, source coverage: The percentage of words in Q_O that have counterparts in QT, target coverage: The percentage of words in Q_T that have counterparts in Q_O, a mutual agreement score. To predict the quality of results, the maximum Jaccard coefficient for the similarity of translated query Q_T to existing target language query logs, the Q_T frequency in the target language query logs, and the average rate of user clicks on results of translated query Q_T. These features implicitly assume that for a query Q_O to benefit from the results of its own translation Q_T, the latter must be similar to queries that users have directly issued in the same language as Q_T.

One aspect of the interaction among query region, its language, and intent is determining the proper intent of a query that may shift intent from one language to another or one region to the other [11]. Taking this interaction into account helps improve the quality of search results presented. Leveraging the multi-regional and multi-lingual intent for some queries is a useful ranking signal [3]. The suitability of particular intents to CLWS has been largely unexamined while issues like user interfaces, ranking, and measuring relevance of MLWS attracted considerably more attention (refer to [9] for a detailed review).

3. INTENTS IN CROSS-LINGUAL SEARCH

3.1 Intents

An assumption underlying the present work is that queries that are likely to benefit share an underlying global, cross cultural or at least bi-cultural intent. For instance, if a user is looking for a global celebrity, a health condition, or features of a new smart phone, such user may consider foreign results relevant. The determination of query intents is a classification problem that groups queries into classes such as "navigational", "informational", or "transactional" [2] or into finer grained classes such as "job search" and "product search" [8]. In the present study, a fine grained intent classification is used that includes 40 intents spanning global ones such as technology, health, and science as well as some cultural-dependent ones like cultural (books and events), and religious/spiritual. Table 1 lists the set of query intents used for this analysis. Given that CLWS could have either a positive or a negative impact on relevance and the search experience in general, it is useful to identify groups of queries that can benefit from CLWS. For an intent to be a suitable CLWS intent it must cross two barriers: the machine translation (or language barrier) and the content availability (or culture barrier). In this subsection, we present an analysis of the factors affecting the suitability of various intents to cross each of these two barriers. This analysis is based on a smaller sample of queries from three regions: the Arabic region characterized by a fast growth in internet connectivity and a reported preference for browsing in English [4], the Province of Quebec as a mature highly bilingual region, and Spanish queries in the United States as another large bilingual population.

3.2 Crossing the Language Barrier

The performance of machine translation is not uniform

Academic	Adult	Art
Auto	Banking	Brands
Celebrity	Commerce	Computers/Technology
Culture	Employment	Entertainment
Event	Forums	Gambling
Games	Gardening	Government
Home Renovation	Image	Informational
Legal	Medical	Music
Navigational	News	People
Phonebook	Political	Recipes
Science	Social	Religious/Spiritual
Sports	Travel	TV
Video	Vocabulary	Weather
Web		

Table 1: List of Intents

across the various intents. For example, in the Arabic market, medical and technology queries were generally better translated than entertainment, music, or news. The greater obstacle in the latter group being named entities getting translated which accounted for nearly 17% of translation problems. However, 15% of the problems were attributed to spelling issues in the original queries (misspellings and word concatenation without spaces). Table 2 summarizes the problems in query translation that we identified in the Arabic and French queries.

Problem	Arabic Queries(%)	French Queries(%)
Poor translation	14.85%	10.00%
Entity translation	17.00%	4.00%
Colloquial	2.20%	-
Spelling	13.20%	4.50%
Word Concatenation	2.75%	-
Spelling-alteration	1.65%	-

Table 2: Distribution of problems in the machine translated Arabic and French queries.

3.2.1 Use of Intents in Translation Verification

The quality of the translation can be considered good if it maintains the same query intent classification. In other words, query intent classifiers can be used in the CLWS pipeline to play two distinct roles. First, they could help determine if a particular query could benefit from CLWS. For example, a scientific query like image enhancement using Fourier transform would be classified as having a scientific and image intents. These intents are indicative of a good CLWS candidate. The query will be translated, but before issuing it, the translated query intent can be checked for a match with the original query intents. Preserving the intent is a signal that the translation is correct. However, a shift in intents does not necessarily indicate a bad translation. For example, the query "Longe da Agua", or "Away from the water" in Portuguese, would trigger a book intent. A perfect English translation may trigger a music intent. The reason behind this discrepancy is the similarity between the translation of the book title and an English song. However, this shift in intents still indicates that the query may not be an ideal candidate for CLWS, as a user seeking information about the book may obtain irrelevant English results.

3.3 Crossing The Content Barrier

A straightforward implementation of CLWS would serve the user results in the target language (say English) in addition to results found in the language of the original query. Consequently, multilingual users stand to benefit more from CLWS. However, these same users are also likely to issue query in the target language when they expect more relevant results in that language. Published studies that relied on observing bilingual users [7] and [10] have found that users often query in the language that matches their information needs. For example, English-Chinese speakers used English to find information about smart phones, and Chinese to find information about Chinese religious holidays. There is a difference though between expecting relevant content in a given language and getting the most authoritative content, as opposed to user generated content. For several intents, authoritative content is often more easily or at times only available in English. Serving multi-lingual users such content is desirable even if the user thought she could find relevant content in another language. Examples of intents whose authoritative content is mostly in English include medical intents, and scientific intents. On the other hand, cultural content has most of its authoritative content in the local language. The level of mastery of the target language introduces another dimension to consider when serving multi-lingual results. A study involving three European language pairs [1]showed that users with moderate or inactive skills in the target language stand to benefit the most from CLWS. These users will have difficulty formulating a query in the target language but can understand the results in that language.

3.3.1 Use of Intents in Content Availability

Some intents have relevant content almost exclusively in the language of the original query in the regions that we studied. These intents include news, entertainment, TV, cultural, and events (local events) intents. Some intents exhibited somewhat different behavior in the Arabic region than in Quebec like the musical, religious, and recipes intents where the English relevant content was relatively scarce compared to the content found for French queries.

3.4 Query Picking

The simplest way to leverage the dependence of the quality of the CLWS on intents is to pick queries whose intents have been identified as benefiting from the process. Out of the 40 intents that we used as intents for the set of queries, the intents that showed the gains were technology, product searches, games, non-local health, science, and travel. Another set of intents has incurred a loss in relevance while a third group remained unaffected. As a query can trigger multiple intents, some combinations of intents were also considered. For example, the combination of "local" and "health" intents predicts a relevance loss as this pair classifiers fires on queries seeking local clinics, doctors, or hospitals and such content is typically available in the local language. Therefore, we identify individual intents and intent combinations that show relevance gains and use them to pick CLWS potential queries.

However, using this strategy alone did not yield a good user experience because it did not address spelling and translation issues that can harm the experience. Hence,we adopted a probabilistic model that takes as inputs the outputs of a spelling correction module to determine if a spelling mistake exists in the given query, and the intent classifiers and calculates a probability of relevance gain. The model uses the quality of translation as a latent variable. To build the model we use a dataset of queries, their intent classifications, a speller derived correctness score, a human judgement of the quality of translation, and a human judgment whether they benefit from CLWS (computed as cumulative gain from inserting top foreign results and removing the worst monolingual results). We compute the probability distribution of translation quality $P(T|S, I_1, I_2, ..., I_n)$, where T is a binary random variable representing the translation quality, S represents the spelling correctness of the query and $I_1, I_2, ..., I_n$ represent the intents. Subsequently, we compute the probability of cumulative gain CG given $T, S, I_1, I_2, ..., I_n$. Only queries that are more likely to produce a gain are translated and issued as CLWS queries.

3.5 Result Merging

The pipeline envisioned for CLWS issues multiple queries to the search servers and merges the results obtained from the different queries. While there could be alternative approaches, we believe that issuing multiple queries stands to bring the best set of results from each language as it exploit all pre-existing query expansion and alteration specific to the language. Hence, the problem of merging ranked results needs to be addressed properly. Ranking scores in different languages are not directly comparable due to large differences in click volumes, number of hyperlinks and many other ranking signals. Simple solutions to result merging like normalizing or scaling the ranking scores generally yield poor results [9]. Here again, we use the probabilistic model to estimate the probability of NDCG gain if we were to insert a foreign result at position i given the spelling correctness score, predicted quality of translation, number of returned results, respective ranker scores, and intents.

4. EXPERIMENTS

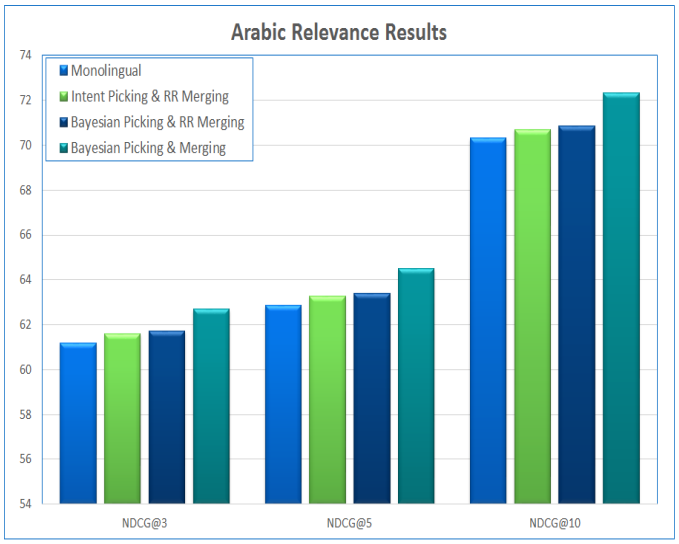

Figure 1: NDCG Results for Arabic Region Queries

To evaluate the intent-based approach to CLWS described here, we used 1500 tail queries from each of the three regions.

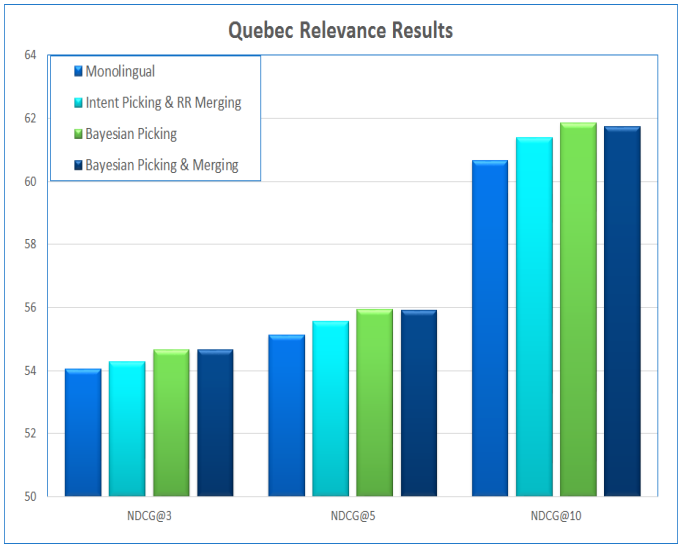

Figure 2: NDCG Results for Quebec Queries

The queries were frequency sampled from the query log of a commercial search engine. The reason for focusing on the tail queries in the evaluation, is that frequent queries are generally navigational queries that are seeking a particular web site and a technique like CLWS is not likely to improve them. The queries were classified for intent, spelling, and four sets of NDCG results are reported for each region. The first set represents the monolingual NDCG results. The second set represents the NDCG results after picking queries whose intents are among the one we designated as benefiting from CLWS and merging the results in a round robin fashion. The third set of results represents the NDCG using probabilistic cumulative gain estimation and round robin merging. The fourth set of results represents NDCG for probabilistic picking and merging as described above. Figures 1, and 2 show the NDCG results obtained for the Arabic queries and Quebec queries respectively. The chart suggest hat the proposed techniques result in a CLWS that improves relevance. The probabilistic picking showed similar performance to probabilistic picking and merging in Quebec but the latter was clearly superior for Arabic queries. One reason for this observation is that the English results were of comparable quality to the French results, so the probabilistic model ended up almost interleaving the results.

The relatively small overall relevance gains would support a statistical null hypothesis stating that the monolingual results and the CLWS results are essentially at parity. However, a closer look at the gaining and loosing queries tells a different story. While the overall NDCG gains are small (less than 1%), the gains observed for the treatment group (i.e. the queries picked for CLWS) are clear as these gains exceed 5% at various depths. However, the number of affected queries has been relatively small and varied widely from one region to another as the query intent profile changes. The arabic query sample we analyzed had a small prelevance of gaining intent like technology, non-local health, travel, product searches and e-commerce. These intents represented a larger proportion of quebec queries. These findings are based on an analysis of at least 2,000 unique tail queries in each region sampled over a period of six months. For additional confirmation, some larger scale analysis has also been conducted targeting particular intents in particular regions.

5. CONCLUSION

The cross-lingual query picking and result merging techniques introduced in this work rely on query intent identification as a triggering signal as well as a translation validation and result merging signal. This intent-based approach has showed gains in two different regions with clearly different intent profile. The result show that CLWS can be useful in both content poor languages like Arabic and more established ones like French. While the final results for Spanish were not ready as the time of this writing, we expect them to be available in the final paper.

6. REFERENCES

[1] E. Airio. Who benefits from clir in web retrieval? *Journal of Documentation*, 64(5):760–778, 2008.

[2] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. In *Advances in Information Retrieval*, pages 578–586. Springer, 2009.

[3] Y. Chang, R. Zhang, S. Reddy, and Y. Liu. Detecting multilingual and multi-regional query intent in web search. In *AAAI*, 2011.

[4] A. Dewachi and S. Aita. *Status of the Digital Arabic Content Industry in the Arab Region*. Number E/ESCWA/ICTD/2012/Technical paper.4. ESCWA: United Nations Economic and Social Commission for Western Asia, 2012.

[5] A. Hefny, K. Darwish, and A. Alkahky. Is a query worth translating: ask the users! In *Advances in Information Retrieval*, pages 238–250. Springer, 2011.

[6] B. Herbert, G. Szarvas, and I. Gurevych. Combining query translation techniques to improve cross-language information retrieval. In *Advances in Information Retrieval*, pages 712–715. Springer, 2011.

[7] W.-Y. Hong. *A descriptive user study of bilingual information seekers searching for online information to complete four tasks*. PhD thesis, University of Pittsburgh, 2011.

[8] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008.

[9] C. Peters, M. Braschler, and P. Clough. *Multilingual information retrieval: from research to practice*. Springer, 2012.

[10] H.-y. Rieh and S. Y. Rieh. Web searching across languages: Preference and behavior of bilingual academic users in korea. *Library & information science research*, 27(2):249–263, 2005.

[11] S. Vadrevu, Y. Zhang, B. Tseng, G. Sun, and X. Li. Identifying regional sensitive queries in web search. In *Proceedings of the 17th international conference on World Wide Web*, pages 1185–1186. ACM, 2008.