Evaluating translation quality and CLIR performance of Query Sessions

Xabier Saralegi[1], **Eneko Agirre**[2], **Iñaki Alegria**[2]
[1]Elhuyar Fundazioa, Usurbil
[2]IXA NLP Group, The University of the Basque Country, San Sebastian
x.saralegi@elhuyar.com

Abstract

This paper presents the evaluation of the translation quality and Cross-Lingual Information Retrieval (CLIR) performance when using session information as the context of queries. The hypothesis is that previous queries provide context that helps to solve ambiguous translations in the current query. We tested several strategies on the TREC 2010 Session track dataset, which includes query reformulations grouped by generalization, specification, and drifting types. We study the Basque to English direction, evaluating both the translation quality and CLIR performance, with positive results in both cases. The results show that the quality of translation improved, reducing error rate by 12% (HTER) when using session information, which improved CLIR results 5% (nDCG). We also provide an analysis of the improvements across the three kinds of sessions: generalization, specification, and drifting. Translation quality improved in all three types (generalization, specification, and drifting), and CLIR improved for generalization and specification sessions, preserving the performance in drifting sessions.

Keywords: Cross-Lingual Information Retrieval, Less-Resourced Languages, Machine Translation, Session-based Information Retrieval

1. Introduction

The successful strategies for query translation in CLIR depend on resources such as Machine Translation systems or parallel corpora, but many languages can not rely on that kind of resources. Thus, they need other strategies based on less or more easily available resources, such as bilingual dictionaries. The translation is usually done from the language of the query to language of the target collection, mainly due to scalability reasons.

Lately, some authors have underlined the importance of using session information for obtaining better rankings (Carterette et al., 2011). They claim that users often try to solve their information need by submitting more than one query, reformulating the initial query. Thus, a process regarding to an information need is often composed by several related queries. After entering an initial query, users tend to reformulate the query in different ways such as specification (e.g., q_i=*"scyhe"* q_r=*"scythe mythology"*), generalization (e.g., q_i=*"computer worms"* q_r=*"malware"*) or drifting (e.g., q_i=*"sun spot activity"* q_r=*"sun spot earthquake"*). Studies on web search query logs showed that half of all Web users reformulated their initial query: 52% of the users in the 1997 Excite data set and 45 % of the users in the 2001 Excite data-set (Wolfram et al., 2001). This paper studies the use of this session context in order to improve the translation quality of the queries and the corresponding retrieval process.

We propose to use the previous queries of the same session in order to improve the query translation step in a CLIR system. Our hypothesis is that queries corresponding to the same session can be used as adequate additional context for improving the translation selection process. For instance, let us assume a session $s=\{q_i,q_r\}$ involving two queries in Basque: the initial query q_i=*"Neil Young diska"*, and its reformulation q_r=*"Neil Young bira data"* (*"Neil Young album"* and *"Neil Young tour date"*, respectively). Translation of q_r without using any context would be wrong, *tr(q_r)="Neil Young turn date"*, but using q_i as context we are able to produce the correct translation, *tr(q_r|q_i)="Neil Young tour date"*, because *"diska"* helps to disambiguate and select the correct translation for *"bira"*.

2. Related work

Several methods are proposed in the literature to deal with the query translation problem. The various techniques can be grouped depending on the translation-knowledge source as follows: MT-based, parallel corpus-based, and bilingual dictionary-based. For the last two groups different statistical frameworks are proposed: cross-lingual probabilistic relevance models and cross lingual language models. The first framework offers operators to treat the ambiguous translations and it is usually used along with dictionaries. The second one incorporates translation probabilities on a more formal and unified framework which is obtained from parallel corpora (Hiemstra, 2000). Although the results depend on the quality of the resources, usually better results are achieved with cross-lingual language models (Xu et al., 2001).

MT-systems and parallel corpora are a scarce for many languages (e.g., Basque). Dictionaries are more accessible for this kind of languages, but they are not free of problems: ambiguous translations must be dealt with. For the translation selection, Pirkola (1998) proposed to use structured queries along with probabilistic relevance models. In this approach all translations of a source word are treated as the same token when *TF* and *DF* statistics are calculated for the translations of that source word. Other authors propose to use the target collection as a kind of language model to solve more precisely the translation selection problem (Monz and Dorr, 2005; Ballesteros and Croft, 1998). Both kind of approaches were studied for the case of English-Basque pair on (Saralegi and Lopez de Lacalle, 2010).

Research in Information Retrieval has traditionally focused on serving the best results for a single query. In practice however users often enter queries in sessions of reformulations. The different editions of Sessions Tracks at TREC, implement experiments to evaluate the effectiveness of retrieval systems over query reformulations. In the TREC 2010 Session track (Kanoulas et al., 2010) sessions were made up of two queries and three types of reformulations were considered:

1. *Generalization*: The user reformulates a more general query when the results are too narrow for him.
2. *Specification*: The user reformulates a more specific query when the results are too broad for him.
3. *Drifting*: The user reformulates another query with the same level of specification but moved to a different aspect or facet.

Overall, systems appeared to perform better over the generalization and drifting sessions than the specification ones (Kanoulas et al., 2010). However only one team achieved a significant statistical improvement. The topics for next editions were collected from real user sessions with a search engine. Sessions were longer and reformulation types were not annotated. In 2011 and 2012 tracks about half of the submitted runs improved the baseline (no information about the session) by using the information about prior queries or using information about prior queries and retrieved results. In 2013 and 2014 editions most of the submitted runs were able to improve the baseline.

There are no papers dealing with query translation by using session information. The most similar works to this topic are those which exploit query logs and web click-through data for generation of cross-lingual query suggestions (Gao et al., 2007) and mining translation of web queries (Hu et al., 2008). Gao and others (2007) introduced a method of calculating the similarity between source language query and the target language query by exploiting, in addition to the translation information, a wide spectrum of bilingual and monolingual information, such as term co-occurrences and query logs with click-through. They used a discriminative model to learn the cross-lingual query similarity from a set of manually translated queries. Hu and other (2008) proposed a methodology for mining query translation pairs from the knowledge hidden in the click-through data. In a first step they identified bilingual URL pair patterns in the click-through data. In a second step they matched query translation pairs based on user click behaviour. Finally, query pairs are generated based on co-occurrence analysis of the click-through data.

3. Experimental setup

As mentioned before, our objective is to improve the performance of the query translation by using previous queries of the same session as context. In order to carry out the experiments we need a test collection including query sessions. We used the query session set built for TREC 2010 Session track. This set includes 150 pairs of initial and reformulated queries (q_i,q_r), grouped by their reformulation type (48 generalization, 52 specification, and 50 drifting). The query pairs were constructed from TREC 2009 and 2010 Web Track diversity topics by using the aspect and main theme of them in a variety of combinations to simulate a session composed of an initial and a second query. These topics correspond to the Clueweb09 collection.

Topic	*Initial query q_i*	*Reformulation q_r*	*Reformulation type*
2	*“hoboken”*	*“hoboken nightlife”*	specification
5	*“low carb high fat diet”*	*“list of diets”*	generalization
9	*“wooden fence”*	*“chain link fence”*	drifting

Table 1: Examples of reformulation types from TREC 2010 Session track’s dataset.

All the 150 query pairs were translated to Basque manually because the objective of the experiment was to evaluate the Basque to English retrieval process. Query pairs extracted from a Basque to English CLIR system’s log would be more realistic. However, such log data was not available. In addition, by using 2010 Session track’s test data we obtained a more standarized test-data. Due to effort limitations we choose the 2010 Session track’s dataset over the next Session tracks’ datasets because it provided reformulation types which offered us more information for the analysis stage.

4. Query translation algorithm

We adopted a bilingual dictionary based strategy for deal with the CLIR process. At first, Basque queries were translated into English. Translation candidates for query terms were obtained from Elhuyar Basque-English dictionary which includes 77,864 entries and 28,874 Basque headwords. Then, an iterative algorithm based on target language co-occurrences was applied for selecting the correct translation when multiple candidates were available. This iterative algorithm is based in (Monz and Dorr, 2005) and its application on the Basque to English CLIR task was evaluated by Saralegi and Lopez de Lacalle (2010). The algorithm selects the translation candidates that maximize the association degree between them according to a target collection.

Initially, all the translation candidates t given by the dictionary for each query term s_i are equally likely:

$$w_0(t|s_i) = \frac{1}{|tr(s_i)|} \quad (1)$$

In the iteration step, the translation weight $w_n(t|s_i)$ of each translation candidate is updated according to the translation weights $w_{n-1}(t'|s_i)$ of the rest of the candidates (*inlink(t)*) and the association degree between them *L(t,t')*:

$$w_n(t|s_i) = w_{n-1}(t|s_i) + \sum_{t' \in inlink(t)} L(t,t') \cdot w_{n-1}(t'|s_i) \quad (2)$$

Then, each translation's weight is re-computed and normalized. The iteration stops when the variations of the term weights become smaller than a predefined threshold.

The weight of the association between candidates *L(t,t')* was computed by calculating the Log-likelihood ratio association measure. We investigated extracting the frequencies (marginal and joint frequencies) required by the Log-likelihood ratio from three collections: ClueWeb09, Wikipedia, and the web (by using Bing search engine). There was not any difference between them, in terms of HTER (Translation Edit Rate). Finally, Wikipedia was used as source collection because of efficiency reasons.

5. Query translation using session as context

5.1. Evaluation of query translation

The quality of translations was evaluated by means of the HTER measure (Snover et al., 2006). This measure computes the average amount of editing that a human would have to perform to correct the output of a system. A lower HTER value means a better translation. We do not penalize wrong word order in the translation because it does not have any negative effect on the retrieval process. In addition to this, according to (Snover et al., 2006), HTER achieves higher correlations than BLEU with human judgements. A fluent speaker in Basque and English performed the minimum number of edits over the English translations provided by the strategies which will be introduced in this section, in order to compute HTER.

We analysed different strategies for combining the initial query and its reformulation in order to improve the translation of the reformulated query:

1. $tr(q_r)$: Translation of the reformulated query without previous query information. This would be the baseline.
2. $tr(q_r|q_i)$: Translation of the reformulated query using the previous query as additional context.
3. $tr(q_r|tr(q_i))$: translation of the reformulated query using the translation of the previous query as additional context.

The hypothesis behind the second strategy is that the words of the previous query contribute positively in the iterative algorithm when finding the correct translations for q_r. For performing this strategy the association degrees between the words in the initial and reformulated queries are taken into account when the translation algorithm is applied over the reformulated query. Thus, the translations of the words in the initial query are added to the *inlink(t)* set. We tested a coefficient *c* for giving more or less importance to the words of the initial query when *L(t,t')* was computed. *L(t,t')* was modified by the coefficient *c* when *t* or *t'* belongs to q_i. Best results were achieved when more weight (*c*=2) was given to the initial query words, with a one point absolute error reduction with respect to the baseline or first strategy, which does not use previous queries.

The third strategy consists on using the translation of the initial query q_i as additional context when translating q_r. The translation of q_i is performed by using the same iterative algorithm. In this case the idea is to to include less and more precise context words. The hypothesis behind this strategy is that, in spite of including some wrong translations, these contexts are more helpful for the iterative translation selection algorithm. For example, the reformulated query q_r=*"PS 2 joko berriak"* (*"new PS 2 games"*) was wrongly translated to $tr(q_r)$=*"PS 2* ***set*** *new"* with the second strategy. The third strategy, which uses the translation of the initial query as context *tr("PS 2 joku")="PS 2 game"*, provides a correct translation $tr(q_r)$=*"PS 2 game new"*.

Table 2. presents the results, showing that better translations are obtained with this third strategy.

Strategy	*Clueweb09*
$tr(q_r)$	0.160
$tr(q_r\|q_i)$	0.157
$tr(q_r\|tr(q_i))$	**0.140**

Table 2: Average HTER depending on translation strategy (smaller is better).

We also analyzed whether any reformulation type could benefit more than the others from using the session information on the translation selection process. The evaluation of translated queries with respect to the different reformulation types shows that the drifting and specification types are more susceptible to be improved (See table 3.). In the case of drifting reformulation words of initial query contribute with new information useful to disambiguate some words of reformulation. This is the case of q_i=*"neil young* ***diska****"* (*"neil young album"*) q_r=*"neil young* ***bira*** *data"* (*"neil young tour date"*) pair, where *"album"* helps to correctly disambiguate *"bira"*. The specification reformulations are improved because initial queries are more general and involve frequent phrases, which the iteration algorithm tends to translate correctly. Using these translations as context helps translation selection process. For example. the reformulated query q_r=*"txakur adopzio erakunde"* is translated correctly to *tr(q_r)*=*"dog adoption organization"* when including the translation of the initial query (q_i=*"txakur adopzio"* translated as *"dog adoption"*), which is performed correctly because it is a frequent phrase, and thus providing useful context words.

We analysed manually some translations of reformulations that theoretically could be improved using the previous query as additional context in the translation selection stage. An analysis of the co-occurrence and Log-likelihood ratio values obtained from the collection was carried out. In some cases, we realized that the semantic relatedness we detected manually between some initial and reformulated queries was not strong enough to affect the translation selection process of the reformulation. In other cases, the manually identified semantic relatedness was not reflected adequately in the collection used for mining co-occurrences. And due to the variety of the topics in the test-set, it is difficult to build an unique collection which fits all of them.

Reformulation type	*HTER for tr(q_r)*	*HTER for tr(q_r|tr(q_i))*
generalization	0.118	**0.107**
specification	0.200	**0.179**
drifting	0.152	**0.130**

Table 3: Improvement on translation quality depending on reformulation type (smaller is better).

5.2. Evaluation of the retrieval process

Next, the retrieval process was evaluated. The translated queries were processed with the Batch Query service for Clueweb09[1] which is based on the Indri search engine. We evaluated the three strategies mentioned above: a) *tr(q_r)*: translation of qr without previous query information (the baseline), b) *tr(q_r|q_i)*: translation of qr using the previous query as additional context, and c) *tr(q_r|tr(q_i))*: translation of q_r using translation of the previous query as additional context.

The results in Table 4 show that the best result is achieved by using the translation of the initial query as context for translating the reformulated query, with up to 5.1% improvement. This improvements of *tr(q_r|tr(q_i))* over *tr(q_r)* on p@10, MAP, and nDCG@10 are significant according to the Paired Randomization Test with α=0.05. The results correspond very well to the improvement in translation quality reported in the previous section.

Strategy	*p@10*	*Impr. over tr(q_r)*	*MAP*	*Impr. over tr(q_r)*	*nDCG@10*	*Impr. over tr(q_r)*
tr(q_r)	0.148	-	0.060	-	0.157	-
tr(qr|qi)	0.152	2.7%	0.060	0%	0.162	3.2%
tr(q_r|tr(q_i))	**0.154**	4.1%	**0.063**	5%	**0.165**	5.1%

Table 4: Retrieval performance depending on query translation and building strategy.

A comparison of the IR results by reformulation type with respect to translation quality (Table 5 vs. Table 3) shows that, unlike the translation quality, the IR improvements for drifting reformulations is the weakest. Hand inspection showed that, in some cases, producing better translation does not necessarily mean that the information need is expressed better. However, this fact should be contrasted with a more extended topic-set including this type of reformulations.

<table>
<tr><td colspan="4"><i>generalization</i></td></tr>
<tr><td><i>Strategy</i></td><td><i>p@10</i></td><td><i>MAP</i></td><td><i>nDCG@10</i></td></tr>
<tr><td><i>tr(q_r)</i></td><td>0.145</td><td>0.053</td><td>0.157</td></tr>
<tr><td><i>tr(q_r|tr(q_i))</i></td><td><b>0.148</b></td><td><b>0.055</b></td><td><b>0.162</b></td></tr>
<tr><td colspan="4"><i>specification</i></td></tr>
<tr><td><i>tr(q_r)</i></td><td>0.143</td><td>0.056</td><td>0.148</td></tr>
<tr><td><i>tr(q_r|tr(q_i))</i></td><td><b>0.160</b></td><td><b>0.061</b></td><td><b>0.167</b></td></tr>
<tr><td colspan="4"><i>drifting</i></td></tr>
<tr><td><i>tr(q_r)</i></td><td><b>0.155</b></td><td>0.070</td><td><b>0.165</b></td></tr>
<tr><td><i>tr(q_r|tr(q_i))</i></td><td><b>0.155</b></td><td><b>0.071</b></td><td><b>0.165</b></td></tr>
</table>

Table 5: Retrieval performance depending on reformulation type.

6. Conclusions

This work shows that: 1) The quality of query translation can be improved using previous queries as context, 2) The improvements in translation quality transfer to improvements in CLIR performance, 3) Translation quality improved in all three types of sessions (generalization, specification, and drifting), and CLIR improved for generalization and specification sessions, preserving the performance in drifting sessions. 4) The best strategy to include the initial query as context is to translate it and then to use the translation in the iterative translation selection algorithm.

The main limitation to obtain higher improvements is due to the weak semantic relatedness scores between the words in the initial query and in the reformulated query. In some cases the related words are not well represented in the collection. Future works will be focused on performing further experiments with different datasets including longer query sessions. We expect that using a

[1]http://boston.lti.cs.cmu.edu/Services/batchquery/

longer context, in terms of amount of previous queries, can mitigate the problems derived from the aforementioned limitations.

7. Acknowledgments

This work has been supported by Basque Government Elkartek program, trough Elkarola project (grants no. KK-2015/00098, KK-2016/00087).

8. Bibliographical References

Ballesteros, L., & Croft, W. B. (1998, August). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 64-71). ACM.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2011, October). Simulating simple user behavior for system effectiveness evaluation. *In Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 611-620). ACM.

Gao, W., Niu, C., Nie, J. Y., Zhou, M., Hu, J., Wong, K. F., & Hon, H. W. (2007, July). Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 463-470). ACM.

Hiemstra, D. (2001). Using language models for information retrieval. Taaluitgeverij Neslia Paniculata.

Hu, R., Chen, W., Hu, J., Lu, Y., Chen, Z., & Yang, Q. (2008, July). Mining Translations of Web Queries from Web Click-through Data. In *AAAI* (pp. 1144-1149).

Kanoulas, E., Clough, P., Carterette, B., & Sanderson, M. (2010). Session track at TREC 2010. Azzopardi et al. [3], 13-14.

Kanoulas, E., Carterette, B., Hall, M., Clough, P., & Sanderson, M. (2011). Session track 2011 overview. In *The 20th Text Retrieval Conference Notebook Proceedings* (TREC 2011).

Monz, C., & Dorr, B. J. (2005, August). Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 520-527). ACM.

Pirkola, A. (1998, August). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 55-63). ACM.

Saralegi, X., & De Lacalle, M. L. (2010). Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In *LREC*. 2010.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006, August). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (pp. 223-231).

Wolfram, D., Spink, A., Jansen, B. J., & Saracevic, T. (2001). Vox populi: The public searching of the web. *JASIST*, 52(12), 1073-1074.

Xu, J., & Weischedel, R. (2000, October). Cross-lingual information retrieval using hidden Markov models. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13* (pp. 95-103). Association for Computational Linguistics.