Is there Hope for Interlingua methods?
A CLIR Comparison Experiment between Interlingua and Query Translation

Marta R. Costa-jussà[1], Rafael E. Banchs[2]

[1] Instituto Politècnico Nacional, Mexico
[2] Institute for Infocomm Research, Singapore

marta@nlp.cic.ipn.mx, rembanchs@i2r.a-star.edu.sg

Abstract. A comparison of interlingua and query translation is proposed in a particular cross-language information retrieval (CLIR) application which consists on retrieving a book from the collection by using one of its chapters in a different language as a query. The experiments are performed in three languages (English, Chinese and Spanish) and all the possible combinations. It is shown that interlingua is able to outperform the query translation approach in some cross-language tasks. Results are further analysed and it is found that, for this particular task, the quality of translation (in terms of BLEU and PER) is not directly correlated with the query translation performance.

Keywords: Interlingua, machine translation, information retrieval.

1 Introduction

Cross-language information retrieval (CLIR) allows users for accessing to documents or information in a different language from their queries. CLIR is becoming more popular as the availability of information in languages different from English increases in the Internet [12]. This paper complements our previous works [3, 4] and compares the performance of different CLIR methodologies.

Research in CLIR has been significantly encouraged by three well-known evaluation campaigns: a cross-language information retrieval track at TREC [1], the Cross-Language Evaluation Forum (CLEF) [2] and the NTCIR [3] Asian Language Evaluation. There are CLIR applications available such as the cross-language search by Google [4] and the Europe Media Monitor [2]. Additionally, there have been recent projects such as Buceador that performs research on integration of all of them in a multilingual and multimodal information retrieval system [1] or XLike that develops technology to monitor and aggregate knowledge that is currently spread across mainstream and

[1] http://trec.nist.gov

[2] http://www.clef-initiative.eu

[3] http://research.nii.ac.jp/ntcir/

[4] http://translate.google.com/translate_s

pp. 81–87; rec. 2014-03-27; acc. 2014-05-04

social media, and enables cross-lingual services for publishers, media monitoring and business intelligence [15].

Given a query in a source language, the aim of CLIR is retrieving most similar and related documents in a target language. [13] identified four types of strategies for matching a query with a set of documents in the context of CLIR by: cognate matching, document translation, query translation or interlingua techniques. From these techniques the most commonly used is query translation.

Query translation is the approach where queries are translated into the document language. It is the most popular approach due to its tractability. Query translation methods translate user queries to the language of the document collection. Query translation has been applied by most CLIR experimental systems because of its convenience and the translation has been mainly addressed by using dictionary-based (i.e. using machine-readable dictionaries, MRD), machine translation (MT) and/or parallel texts techniques [5].

On the other hand, interlingua methods transforms both documents and queries into a language-independent representation. An interlingua method aims at associating related textual contents among different languages by means of language-independent semantic representations. The conventional interlingua-based CLIR approach uses latent semantic indexing (LSI) for constructing a multilingual vector-space representation [8, 9, 6] of a given parallel document collection. Vector-space representations are known to be noisy and sparse. That is why in order to obtain more efficient representations, space reduction techniques such as latent semantic indexing and probabilistic latent semantic indexing [10] are applied. The new reduced-space dimensions are supposed to capture semantic relations between words and documents in the collection. Recent approaches have achieved interesting results by using regression canonical correlation analysis (an extension of canonical correlation analysis) where one of the dimensions is fixed and demonstrate how it can be solved efficiently [14]. Also the use use of nonlinear semantic mapping techniques have been proposed in our previous works [4].

The query translation approach has been considered as the only state-of-the-art approach for CLIR applications. However, in a N-lingual environment the number of required systems reaches the $\frac{N(N-1)}{2}$. The main advantage of the interlingua-based strategy in a highly multilingual environment is that, compared to the query translation strategy, it reduces cross-language information retrieval number of systems to N. That is why, we propose to test interlingua versus query translation for one particular application which consists on retrieving a book from the collection by using one of its chapters in a different language as a query.

The rest of the paper is structured as follows. Next section describes the methodologies that are compared in this paper. Section 3 reports several CLIR experiments performed on a trilingual document collection. The LSI methodology is compared with that of a standard IR system and the query translation CLIR approach showing that, in the case of cross-language tasks, the proposed approach is able to outperform the conventional one. Results are analysed in order to find out if the quality of translation can predict the quality of the query translation approach. Finally, Section 4 includes the most relevant conclusions derived from the experimental results are presented and some future research actions for continuing the present work are depicted.

2 CLIR methodologies

This section briefly describes the LSI-interlingua and the query translation methodologies. The LSI-interlingua methodology basically uses the singular value decomposition (SVD) of a tf-idf (term frequency - inverse document frequency) matrix, which considers that a rectangular matrix X of dimensions $M \times N$ can be factorized:

$$X = U \varSigma V^T \tag{1}$$

U and V are unitary matrices of dimensions $M \times M$ and $N \times N$, respectively, and $\varSigma$ is a $M \times N$ diagonal matrix containing the singular values associated to the decomposition. Consider M the number of terms and N the number of documents. According to [9], a low-dimensional representation of a given document vector x can be obtained as follows:

$$y^T = x^T U_{M \times L} \tag{2}$$

y is the L-dimensional document vector corresponding to the projection of an M-dimensional document vector x, and $U_{M \times L}$, is a matrix containing the L first column vectors of the unitary matrix U that is obtained from (1) given X. This rank reduction has been proven to preserve most important semantic information in the collection of documents while reducing noise. This LSI methodology can be extended to the cross-language case [9], where the main difference is that X is a term-document matrix constructed with parallel documents in two languages:

$$[X_a; X_b] = U_{ab} \sum_{ab} V_{ab}^T \tag{3}$$

$[X_a; X_b]$ is a bilingual term-document matrix obtained by concatenating monolingual term-document matrices for a parallel document collection. In this case, low-dimensional representations for given document vectors x_a and x_b in languages a and b, can be obtained:

$$y_a^T = [x_a; 0]^T U_{abM \times L} \tag{4}$$

$$y_b^T = [0; x_b]^T U_{abM \times L} \tag{5}$$

If we assume that similar terms in multiple languages have approximately the same occurrence patterns, then, we could find a close representation in the multilingual reduced space for semantically related terms and documents. In this case, documents across languages could be compared in the reduced space. Finally, a measure of similarity could be used to compute the similarity among documents.

In contrast, the query translation methodology simply translates the query using a standard machine translation system (MT) and it uses a monolingual information retrieval system (IR). Notice that in this case, errors from the first step (MT) are concatenated with errors from the second step (IR). In this work, we are not considering query expansions and k-bases translations of queries that are more sophisticated ways of performing query translation.

3 Experiments

The data has been extracted from the Chinese, Spanish and English versions of the *Holy Bible*, which has been proved to be a good resource for CLIR experimenting [7]. The basic characteristics of the collection are described in Table 1.

Table 1. Basic characteristics of the experimental dataset.

Language	Chapters	Books	Vocabulary	Non singletons
Chinese	1189	66	12,670	6,286
Spanish	1189	66	26,251	13,632
English	1189	66	13,216	7,265

The task consists in retrieving a book from the collection by using one of its chapters in the same or different language as a query. We have randomly selected 200 chapters from the 1189 total chapters to be used as test set.

For evaluation purposes, the LSI-interlingua method is compared with standard CLIR approach. In the former, the training set is variable (5 to 300 chapters) and the test set are 60 books. The retrieval space dimensionality is equal training size. The number of performed runs is 100. The query translation was implemented concatenating the Google translation API and the monolingual information retrieval system, which was implemented by using Solr. Solr is an XML-based open-source search server based on the Apache-Lucene search library[5].

In order to find out the correlation between why one system is better than the other, Table 2 shows the quality of translation using standard measures such as BLEU (i.e. Bilingual Evaluation Under Study) and PER (i.e. Position Error Rate).

Table 2. Translation quality. BLEU and PER metrics.

	Chinese		Spanish		English	
	BLEU	PER	BLEU	PER	BLEU	PER
Chinese	-	-	12.01	63.5	15.68	57.72
Spanish	12.77	64.86	-	-	26.47	44.69
English	17.94	63.13	28.07	45.72	-	-

Finally, we performed analysis of the correlation between the results and the translation metrics, see Figure 2. From the aforementioned figures one can see:

1. LSI (interlingua) outperforms query translation in five situations out of twelve and in one situation both techniques obtain equal results;

[5] http://lucene.apache.org/solr/

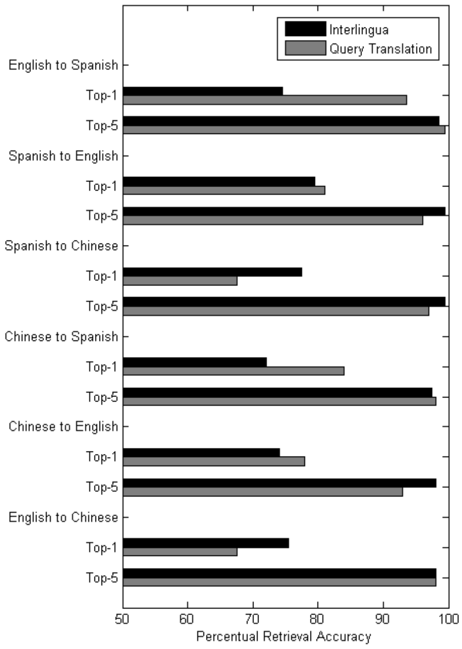

Fig. 1. Results for the cross-language information retrieval task for all pairs of languages. For each language pair, the first two bars correspond to the Top-1 results and the second two bars correspond the Top-5 results.

2. Translation metrics do not correlate well with any metric of CLIR. This means that the quality of translation evaluated with standard MT metrics does not provide information in a CLIR system.

These results differ from previous shown in [11]. We suspect that the main reason for contradicting the Kettunen conclusions is that we are working in a different CLIR task and the query is specially large; (3) the Top-1 and Top-5 metrics are correlated when using the interlingua approach but not in the query translation approach. Finally, we analysed the errors from both systems. Among the total errors only between 10% and 30% of the cases were the same errors, which indicates that a system combination could improve the task performance.

4 Conclusions and Future Work

This paper presents one particular application which consists on retrieving a book from the collection by using one of its chapters in a different language as a query. In this

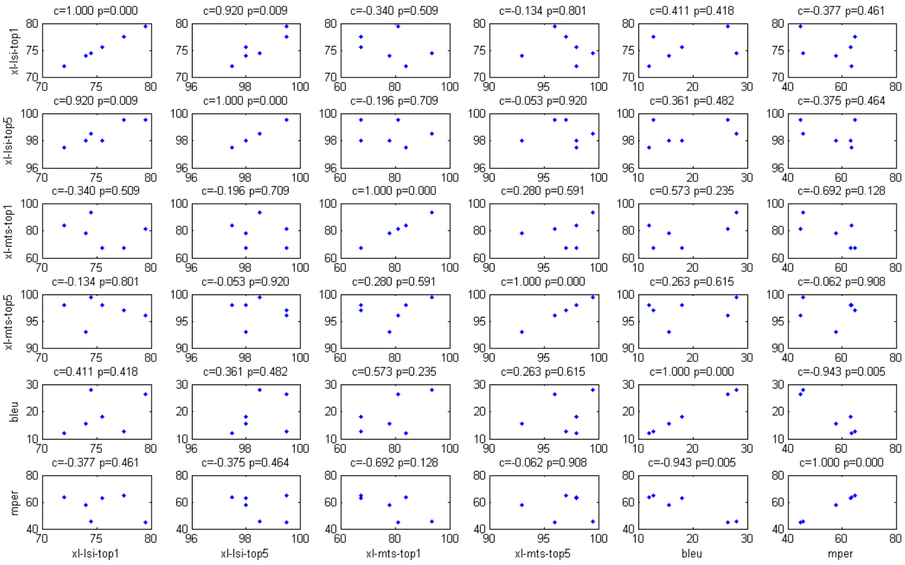

Fig. 2. Correlations (c) and p-value (p) between CLIR quality and MT quality (BLEU and PER). LSI stands for Latent Semantic Indexing and MTS stands for query translation.

framework, the LSI (or interlingua) method is compared with state-of-the-art CLIR approach. Evaluation results show that the proposed method is able to outperform the reference system in the case of cross-language information retrieval (specially when considering the Top-5 results). Additionally, comparing the errors of both system outputs, there is a maximum of 30% matching. This result was somehow expected given the different nature of both systems. Taking advantage of this information, we could try to perform system combination methods. Finally, we found that for this particular task, the translation quality (in terms of BLEU and PER) is not correlated with the CLIR quality. This may be explained because, for our task, the query is specially large. As future research in this area, we will focus on finding a correlation between the quality of translation and the interlingua and query translation performances.

With these results, we showed that there is hope for interlingua methods. Specially taking into account that, if we get similar results to a query translation approach, in a N-lingual environment, interlingua approaches reduce the number of systems from $\frac{N(N-1)}{2}$ to N.

Acknowledgements. This work has been supported in part by Spanish Ministerio de Economía y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER).

References

1. Adell, J., Bonafonte, A., Cardenal, A., Costa-Jussà, M.R., Fonollosa, J.A.R., Moreno, A., Navas, E., Banga, E.R.: Buceador, a multi-language search engine for digital libraries. In: Chair), N.C.C., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
2. Atkinson, M., der Goot, E.V.: Near real time information mining in multilingual news. In: Proceedings of the 18th international conference on World wide web. pp. 1153–1154 (2009)
3. Banchs, R.E., Costa-jussà, M.R.: A non-linear semantic mapping technique for cross-language sentence matching. In: Proceedings of the 7th International Conference on Advances in Natural Language Processing. pp. 57–66. IceTAL'10, Springer-Verlag, Berlin, Heidelberg (2010)
4. Banchs, R.E., Costa-jussà, M.R.: Cross-language document retrieval by using non-linear semantic mapping. Applied Artificial Intelligence Journal 27(9), 781–802 (2013)
5. Chen, J., Bao, Y.: Cross-language search: The case of google language tools. First Monday 14(3) (2009)
6. Chew, P., Abdelali, A.: Benefits of the passively parallel Rosetta stone? Cross-Language information retrieval with over 30 languages. In: Proc of the 45th Annual Meeting of the Association for Computational Linguistics. vol. 45, pp. 872–879 (2007)
7. Chew, P.A., Verzi, S.J., Bauer, T.L., McClain, J.T.: Evaluation of the bible as a resource for cross-language information retrieval. In: Proceedings of the Workshop on Multilingual Language Resources and Interoperability. pp. 68–74 (2006)
8. Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
9. Dumais, S.T., Landauer, T.K., Littman, M.L.: Automatic cross-linguistic information retrieval using latent semantic indexing. In: SIGIR96 Workshop on Cross-Linguistic Information Retrieval. pp. 16–23 (1996)
10. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of Uncertainty in Artificial Intelligence, UAI99. pp. 289–296 (1999)
11. Kettunen, K.: Choosing the best MT programs for CLIR Purposes: Can MT metrics be helpful? In: Proc. of the 31th European Conference on IR Research on Advances in Information Retrieval. pp. 706–712 (2009)
12. Nie, J.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010)
13. Oard, D.W., Diekema, A.R.: Cross-Language information retrieval. Annual Review of Information Science and Technology (ARIST) 33, 223–256 (1998)
14. Rupnik, J., J., S.T.: Multiview canonical correlation analysis and cross-lingual information retrieval. In: http://videolectures.net/lms08_rupnik_rcca/ (2008)
15. Wang, Z., Li, J., Zhao, Y., Setchi, R., Tang, J.: A unified approach to matching semantic data on the web. Knowl.-Based Syst. 39, 173–184 (2013)