Should MT Systems Be Used as Black Boxes in CLIR?*

Walid Magdy and Gareth J.F. Jones

Centre for Next Generation Localisation,
School of Computing, Dublin City University, Dublin 9, Ireland
{wmagdy,gjones}@computing.dcu.ie

Abstract. The translation stage in cross language information retrieval (CLIR) acts as the main enabling stage to cross the language barrier between documents and queries. In recent years machine translation (MT) systems have become the dominant approach to translation in CLIR. However, unlike information retrieval (IR), MT focuses on the morphological and syntactical quality of the sentence. This requires large training resources and high computational power for training and translation. We present a novel technique for MT designed specifically for CLIR. In this method IR text pre-processing in the form of stop word removal and stemming are applied to the MT training corpus prior to the training phase. Applying this pre-processing step is found to significantly speed up the translation process without affecting the retrieval quality.

1 Introduction

Cross-language information retrieval (CLIR) is concerned with searching a collection of documents that are in a different language from the user's query. Two main techniques have been used for the translation step in CLIR; bilingual dictionaries and machine translation (MT) systems [4]. In recent years, MT has become the most commonly used technique in CLIR due to the increasing availability of high quality free MT systems, such as Google translate[1], Bing translate[2], and Yahoo Babel Fish[3]. In addition, some open source statistical MT (SMT) libraries are available for research purposes, such as MaTrEx [7] and Moses [1].

Since the MT approach usually provides a high quality translation for queries that consequently leads to high retrieval effectiveness in CLIR close to that of monolingual information retrieval (IR), it has been always used as a black box for the translation process in CLIR. Less attention was directed toward the fact that MT and IR have two different perspectives in measuring the quality of a sentence. 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.

* This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at Dublin City University.

[1] http://translate.google.com/

[2] http://www.microsofttranslator.com/

[3] http://babelfish.yahoo.com/

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 683–686, 2011.

In this paper we open the black box of the MT system and we present a novel technique for using it in a much more efficient way for the purpose of CLIR. The approach introduced utilizes the fact that the surface form and the sentence structure is generally unimportant in standard IR application, To do this, the workflow of the MT process is adapted to focus only on the conceptual meaning of text and neglect its structure. The new setup of the MT system is demonstrated to be five times faster than the standard MT techniques in both the training and decoding phases when tested on the cross language patent search task from the CLEF-IP 2010. The retrieval effectiveness using the new technique of translation is proven to be statistically indistinguishable from results obtained using standard MT.

2 New Approach for Using MT in CLIR

An overlooked issue in CLIR systems when MT is used for translation is that MT systems take significant time selecting proper sentence structure for the output, which is unused later by the IR system. Conventional MT focuses on generating a translation that is human readable, therefore it seeks to select the proper pronouns, verb tenses, and word ordering for the translated text. This requires a huge amount of processing power and time for executing an effective algorithm for selecting the proper words since pronouns and verb tenses are generally found to be the most confusing terms in any translation. On the other hand, IR cares more about the conceptual meaning of the word regardless to its surface form and tense. In addition, all the pronouns are considered insignificant to the translated text for IR purposes and are filtered out of the query and the documents prior to their entry into the IR system.

The basic idea in our new approach is to train the MT system for translation of topics or documents in CLIR using training data pre-processed for IR. The pre-processing of IR data uses the standard stages performed by most of the IR systems, which includes case folding, stopword removal, and stemming. These three operations aim to improve retrieval efficiency and effectiveness by removing insignificant words and matching different surface forms of words. While these are standard processes in IR, for MT, applying these operations in the pre-processing stage would appear to be destructive of the quality of the translated sentence. However, since the objective of the translation process here is retrieval effectiveness, the quality of the text structure of the translated content is unimportant. Our hypothesis is that training an MT system using corpora pre-processed for IR can lead to similar or possibly improved translated text from the IR perspective, which consequently can lead to better retrieval effectiveness. In addition, the training of the MT system and subsequent translation is expected to be much faster and more efficient, since a large portion of the text which represents the stopwords will be removed, and the remaining content will be normalized creating a smaller vocabulary, and that a smaller processed training corpus can be as effectively as a larger unprocessed corpus for translation in CLIR.

3 Experimental Investigation

To demonstrate the effectiveness and efficiency of the proposed approach, a cross language search in patent retrieval task was used. The main objective is to find

relevant documents in an English collection of patents that are related to French patent applications. The data comes from the CLEF-IP 2010 task [5], where 134 French patent topics are used to search a collection of 1.35M patents that consist of English text only. Since the patent collection comes from the European patent office (EPO) most of the patents in the collection have some parts translated into three languages: English, French, and German. For the MT experiments, 8.1M parallel sentences in English and French were extracted from the collection.

For the CLIR baseline run, the 8.1M parallel sentences were used to train the MaTrEx MT system [7] without pre-processing (referred to later as "ordinary MT"), and then using the output MT model to translate the 134 French patent topics. Since the translated text is in its full form, standard IR pre-processing was applied to the translated text to filter out English stopwords and to stem the words.

The same training data set was used to train the MaTrEx MT system again, but after pre-processing the data to remove stop words, apply case folding, and stem words for both languages in the parallel corpora (referred to later as "processed MT"). The output MT model was used to translate the French topics after applying the same IR pre-processing prior to the translation. Hence the translated text output in this case is in the form of stemmed English words with no stop words.

Queries were constructed from the translated patent topics based on the best runs submitted to the CLEF-IP 2010 [5], where most of sections in the patent topics were used to formulate the query as described in [3]. The time taken for translating these long queries was found to be very long (30 mins per topic using an Intel Xeon quad-core processor, 2.83GHz, 12MB cache, and 32GB RAM), which motivates the need for a more efficient translation process to reduce the translation time. The indri search toolkit was used for indexing and searching the collection [6].

Retrieval effectiveness is measured using two scores; mean average precision (MAP) and the recently introduced patent retrieval evaluation score (PRES) [2]. PRES is an evaluation score designed for recall-oriented retrieval tasks where the objective is to find all possible relevant documents at the highest possible ranks. Significance is tested using Wilcoxon test with p-value 0.05. The time for training the MT systems and decoding (translating) the patent topics were calculated.

4 Results

Table 1 reports the results for the CLEF-IP 2010 CLIR task when using the ordinary MT vs. the processed MT as the translation process. The retrieval effectiveness results were found to be statistically indistinguishable between using the translation techniques when compared using either MAP or PRES. This result shows that processing the query text by removing stop words and stemming will lead to the same retrieval results regardless of whether it is applied before or after the translation process. The other results in Table 1 show the main benefit of applying the new "processed MT" approach, which is the MT processing time. It can be seen that the processed MT is much faster than the ordinary MT, since it is more than five times faster in both the training and decoding (translation) phases. These results confirm that adapting the MT system for IR use to be much more efficient than using it as a black box while maintaining the retrieval effectiveness.

Table 1. Retrieval effectiveness and processing time compared when using ordinary MT vs. processed MT for the CLEF 2010 cross language patent search task

		Ordinary MT	**Processed MT**
Retrieval Effectiveness	MAP	0.085	0.084
	PRES	0.413	0.419
Processing Time (*hh:mm:ss*)	Training (8M sentences)	221:31:28	44:11:16
	Decoding (134 patents)	68:18:21	13:29:39

5 Conclusion and Future Work

This paper studied the use of MT systems in CLIR with the objective of discovering whether there is a way of training MT systems specifically for IR instead of using them as black boxes. We presented an efficient technique for training MT systems for the purpose of CLIR by re-ordering the workflow of the CLIR steps to apply the standard IR pre-processing prior to the translation process instead of after it, and to train the MT system in the same fashion by processing the parallel corpus before the MT training. Testing the suggested approach on a cross language patent search task showed the new translation process to be five times faster than the ordinary MT system while preserving the same retrieval quality.

For future work, the approach needs to be further tested on different language pairs. In addition, the performance of the new MT approach is required to be investigated when only limited amount of MT training corpus is available.

References

1. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL 2007 (2007)
2. Magdy, W., Jones, G.J.F.: PRES: a score metric for evaluating recall-oriented information retrieval applications. In: SIGIR 2010 (2010)
3. Magdy, W., Jones, G.J.F.: Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. In: CLEF 2010 (2010)
4. Oard, D.W., Diekema, A.R.: Cross-Language Information Retrieval. In: Williams, M. (ed.) ARIST (1998)
5. Piroi, F.: CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In: CLEF 2010 (2010)
6. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: ICIA (2004)
7. Stroppa, N., Way, A.: MaTrEx: DCU Machine Translation System for IWSLT 2006. In: IWSLT (2006)