An Efficient Method for Using Machine Translation Technologies in Cross-Language Patent Search

Walid Magdy and Gareth J. F. Jones
Centre for Next Generation Localization
School of Computing, Dublin City University, Dublin 9, Ireland
{wmagdy, gjones}@computing.dcu.ie

ABSTRACT
Topics in prior-art patent search are typically full patent applications and relevant items are patents often taken from sources in different languages. Cross language patent retrieval (CLPR) technologies support searching for relevant patents across multiple languages. As such, CLPR requires a translation process between topic and document languages. The most popular method for crossing the language barrier in cross language information retrieval (CLIR) in general is machine translation (MT). High quality MT systems are becoming widely available for many language pairs and generally have higher effectiveness for CLIR than dictionary based methods. However for patent search, using MT for translation of the very long search queries requires significant time and computational resources. We present a novel MT approach specifically designed for CLIR in general and CLPR in particular. In this method information retrieval (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 of the MT system. Applying this step leads to a significant decrease in the MT computational and resource requirements in both the training and translation phases. Experiments on the CLEF-IP 2010 CLPR task show the new technique to be 5 to 23 times faster than standard MT for query translation, while maintaining statistically indistinguishable IR effectiveness. Furthermore the new method is significantly better than standard MT when only limited translation training resources are available.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms
Algorithms, Performance, Experimentation.

Keywords
Patent Retrieval; Cross-Language Information Retrieval; Machine Translation.

1. INTRODUCTION
Interest in patent retrieval research has shown considerable growth in recent years. The focus of most of this research has mainly been on exploring methods for monolingual patent search tasks, where the emphasis has been on indexing techniques for patents and query formulation for topics. However, an important and largely overlooked topic in patent retrieval is international and hence multilingual patent search. Patents on the same topic may be published in different countries in different languages, and it is important for patent examiners to be able to locate relevant existing patents whatever language they are published in. Hence an important topic in patent retrieval is cross-language information retrieval (CLIR), where the topic is a patent application in one language and the objective is to find relevant prior-art patents in other languages [2, 6]. In recent years machine translation (MT) has become established as the dominant technique for translation in CLIR. This has largely come about due to the increased availability of high quality MT systems, which usually achieve better CLIR effectiveness than dictionary-based translation methods. Standard MT systems focus on generating proper translations that are morphologically and syntactically correct. Development of effective MT systems requires large training resources and high computational power for training and translation. This is an important issue for patent CLIR where queries are typically very long, sometimes taking the form of a full patent application; meaning that query translation using MT systems can be very slow and computationally demanding. However, in contrast to MT, the focus for information retrieval (IR) is on the conceptual meaning of the search words regardless of their surface form. Thus much of the complexity of the standard MT process is not required for effective CLIR. The significant time and resources required for translation of patent topics in cross language patent retrieval (CLPR) has not received much attention to date. In addition, some language pairs have limited suitable training data available, meaning that it is not possible to train an effective MT system for these language pairs leading to low CLPR effectiveness.

In this paper, a novel adaptation of MT for CLIR is presented which addresses the high computational cost and resource requirements of MT for CLPR. The is demonstrated to be up to 23 times faster than standard MT in both the training and decoding phases for the CLEF-IP 2010 patent search task. Retrieval effectiveness using the new approach is shown to be statistically indistinguishable from that obtained using standard MT. Furthermore, it is found to be statistically significantly better than standard MT when only a small amount of data is used to train the system.

2. PATENT SEARCH
In recent years, several IR evaluation campaign have included tracks exploring recall-orientated tasks. Two of these are the NTCIR [2] and CLEF [6] patent search tracks, which have examined ad-hoc search, invalidity search, and prior-art search.

In this paper, we focus on the prior-art patent search task, which is concerned with finding all relevant patents that can invalidate the novelty of a patent application or at least that have common parts to that patent [6]. The full patent application submitted to the

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

patent office is considered as the topic, and patent citations that are identified by the patent office are taken as the relevant documents, therefore the objective in prior-art patent search is to find these citations of patents automatically.

CLPR has featured as a task at both NTCIR and CLEF. The typical procedure adopted for CLPR has been to translate the query into the target collection language using one of the available free MT systems, and then to perform search in the document language. Thus this research has treated the translation stage as a black box without any control over the translation process. In addition, little attention has been directed toward the time taken for the translation process.

3. ADAPTING MT FOR CLIR

3.1 Basic Concept

The basic idea of the new approach is to train an MT system for translation of topics or documents in CLIR using training data pre-processed for IR. The pre-processing uses the standard stages performed by most IR systems. specifically case folding, stop word removal, and stemming. These operations aim to improve retrieval efficiency and improve effectiveness by matching different surface forms of words. While these are standard processes in IR, for standard MT applying these operations would be destructive to the quality of the translated output. For example, the translated sentence "*he are an great idea to applied stem by information retrieving*" instead of "*It is a great idea to apply stemming in information retrieval*" would be considered a very bad translation from an MT perspective. However, from an IR perspective this output is fine since it contains all the information needed for the retrieval process, since both are the same after IR pre-processing: "*great idea appli stem informat retriev*".

Our hypothesis is that training an MT system using corpora pre-processed for IR can lead to similar or improved translated text from the IR perspective, which consequently can lead to better retrieval effectiveness. In addition, the training of the MT system is expected to be much faster and more efficient, since a large proportion of the training text represented by the stop words will be removed, and the rest will be normalized creating a smaller vocabulary. Further this reduced vocabulary should mean that a smaller training corpus will be found to be as effective as a larger unprocessed one for translation in CLIR.

3.2 MT Training and Decoding

Figure 1 presents the workflow of the proposed CLIR system. The upper part represents the MT training which produces the translation model used for the translation step in the CLIR. The new "Text Processing" step introduced for both languages in the parallel corpus works by applying the standard IR pre-processing steps. The resulting translation model is in the "Processed" form, where words are in their stemmed form and no stop words are present. For consistency, the terms "Processed" and "Text Processing" in the remainder of the paper refer to "case folding", "stop word removal" and "stemming".

For query translation in CLIR when using MT, a query in source "S" language is translated into target "T" language; the translated query is then processed in language "T" for search. Actually, when using MT for CLIR, longer queries are preferable since they tend to be more grammatical, therefore better translation can be achieved using an MT system taking context into account, leading to better retrieval effectiveness. The novel translation approach introduced here is shown in the lower part of Figure 1. It can be seen that the "Text Processing" step has been moved to be a step prior to translation instead of a posterior step in the standard CLIR workflow. Therefore, the processing is applied to the source language query which produces a much shorter input with a reduced vocabulary to be translated using the processed MT model. The output from the translation process is in the processed form, and therefore no additional processing of the query is required. This query is used directly to search the index of documents and produce a list of retrieved results.

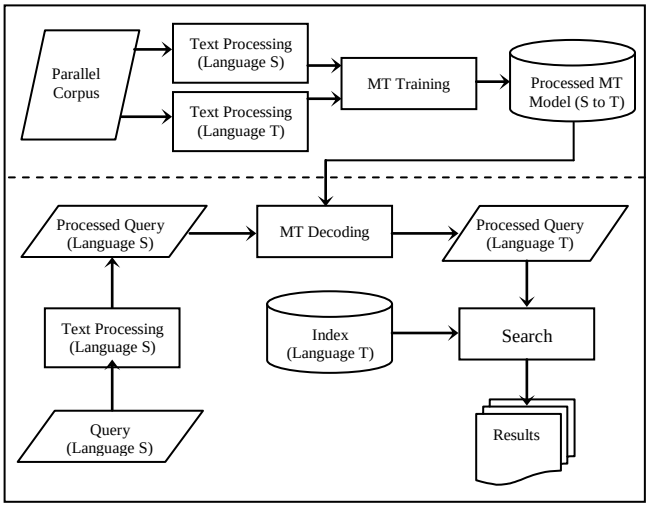

Figure 1: Workflow of the proposed CLIR system

4. EXPERIMENTAL INVESTIGATION

The experimental investigation examines three main dimensions of the proposed approach. The first is to explores the effect of processing the words before the MT step. The second investigates the efficiency of the proposed translation process according to the computational requirements for the MT training and decoding phases when compared to translation using standard MT. However, more emphasis is given to the decoding time for query translation since it is the online processing time for translating the query which is generally more significant to the user. The third dimension considers the effect of using a limited amount of training data on the retrieval effectiveness.

Retrieval effectiveness in this investigation is measured using MAP and the recently introduced patent retrieval evaluation score (PRES) [3]. PRES is an evaluation score designed for recall-oriented tasks where the objective is to find all possible relevant documents at the highest possible ranks. PRES emphasises the quality of the system in retrieving a large portion of the relevant documents at relatively high rank based on a user specific cut-off (N_{max}). In our analysis, we focus on PRES since it is specifically designed for measuring retrieval effectiveness in patent search, where it combines recall and quality of ranking in one score. Moreover, it is used in CLEF-IP track since 2010 to evaluate the performance of the submitted runs. Significance is tested using a Wilcoxon test with p-value 0.05. In addition, the times for training the MT systems and for decoding (translating) the topics are calculated for both methods.

4.1 Test Data

The cross language search task in CLEF-IP 2010 is used for our experiments. The main objective is to find relevant patents in a multilingual collection that are related to patent applications filed in French and German languages. The patent collection consists of 1.35M patents from the European Patent Office (EPO) with 69%

of them in English and 31% in German and French. The German and French patents are provided with many sections manually pre-translated into English, including the patent title, abstract and claims. The English text of all patents in the collection was indexed to create an index of documents in English only. The CLEF-IP track provided two sets of topics; 300 training topics of which 89 are German, 15 are French, and the remainder are English; and 2000 test topics of which 520 are German, 134 are French, and the rest are English. Both sets of topics are patent applications filed after those in the patent collection and do not contain translations. For the CLPR experiments, the 89 German training topics and the 134 French test topics were selected to have a similar number of topics for each query language.

Since the patent collection comes from the EPO, most of the patents in the collection have the title and claims sections translated into three languages (English, French, and German). For the MT experiments, more than 8M (~8.1M) parallel sentences in English, German, and French were extracted from the collection for use as the MT training set. The average length of the English sentences in the corpus is 28 words.

4.2 Baseline Construction

Query formulation from the patent topic is one of the main challenges in patent search [2, 6]. To construct a baseline retrieval run, we tested a number of query formulation approaches based on the best runs submitted to the CLEF-IP 2010 [6]. Based on these existing runs, our query formulation used the title, abstract, description, claims, and classification sections. We followed the our query formulation originally presented in [4], where the query is constructed using terms in the topic after translation that appeared more than two times across the sections when combined and all bigram terms that appeared more than three times, with the term frequency acting as weight for these terms. The Indri search toolkit[1] was used for indexing and search, Porter stemmer was applied for the queries and documents, and a list of 684 stop words from patent domain used in [4] was filtered out from text.

Two baseline runs were prepared for each query language: the first baseline used Google translate to translate the German and French topics into English, as was done by most of the participants in CLEF-IP 2010 [6]. For the second and main baseline, we used the MaTrEx MT system[2] [7]. The 8M extracted sentences were used to train the MaTrEx MT system to create two translation models: (French→English) and (German→English). The default configuration and training parameters of the MaTrEx system were used to generate the translation model, which was then used to translate the German and French test topics into English. Table 1 shows the MAP and PRES values for each of the baselines for the French and German topics. From these results it can be seen that, for the French topics the Google and MaTrEx MT systems achieved similar retrieval effectiveness. However for German topics Google translate achieved lower performance with respect to both MAP and PRES, this can be attributed to the many unusual compounds found in the text that require a training corpus in a similar domain in order to be translated effectively. For the translation time using MaTrEx, it was found that the average translation time was 31 mins for the French patent topic (which contain 7,058 words on average) and 12 mins for the German patent topic (which contain 3,571 words on average) on a server machine (Intel Xeon quad-core processor, 2.83GHz, 12MB cache, and 32GB RAM). However, the average search time using all the translated text as a query was 42 secs for French topics and 14 secs German topics on a desktop machine (Intel Core2Due, 3GHz, 6MB cache, 3GB RAM). This highlights the importance of developing faster translation techniques for patent topics.

Table 1: Baseline runs for the German and French topics

	French		German	
	MAP	PRES	MAP	PRES
Google	0.087	0.413	0.067	0.466
MaTrEx	0.085	0.413	0.075	0.487

5. EXPERIMENTS WITH THE NEW CLIR MT APPROACH

The same training dataset of parallel sentences was used to train the MaTrEx MT system again, but after pre-processing the data ("processed MT"). This was then compared to the standard MT system without pre-processing the data ("ordinary MT"). In addition, several portions of the training data were selected and used to train alternative MT systems to explore the performance of both MT systems when less training examples are available. For these experiments subsets 800k, 80k, 8k and 2k sentences were extracted at random from the full 8M training set and used to train the additional MT systems.

5.1 Results

Table 2 shows the retrieval effectiveness measured by MAP and PRES, the out-of-vocabulary (OOV) rates when decoding the topics, and decoding time for French and German topics compared when using ordinary MT vs. processed MT for the cross language patent search task.

For the retrieval effectiveness measured by MAP and PRES, it can be seen that the difference in the retrieval effectiveness using both translation methods is not significant compared to each other for almost all training sizes. However, with smaller training sets (2k), it is found that the processed MT achieved significantly better retrieval effectiveness than the ordinary MT for both query languages when compared using PRES. For the French topics when using processed MT, results remain statistically indistinguishable from Google translate for training sizes 8M, 800k, and 80k. However, for ordinary MT, the 80k training set translation led to retrieval that is statistically worse than Google translate when compared using PRES. These results show that the new approach has higher effectiveness when limited amounts of training data are available.

To analyse the reason behind these results, the OOV percentage while translating the patent topics is also reported in Table 2. It can be seen that the stemming performed in the "Text Processing" step in the processed MT system reduces the number of OOV terms, leading to the presence of a translation. In particular, it can be seen that for small size training sets, the standard translation approach suffers from a large percentage of OOVs, while the processed MT system overcomes part of this problem. The German topics suffer from higher OOV than the French ones due to the presence of productive compounds in German.

The second main benefit of the new approach to translation is shown clearly in the last row of Table 2, which compares the average decoding time required to translate a patent topic into English using both approaches. It can be seen that the processed MT system is at least 5 times faster than the ordinary MT system

[1] http://www.lemurproject.org/indri/
[2] http://www.openmatrex.org/

Table 2: Retrieval effectiveness, OOV, and decoding time for French and German topics compared when using ordinary MT vs. processed MT for the cross language patent search task. Underlined values indicate that the result is indistinguishable from Google translate, and '*' indicates that processed MT is statistically better than ordinary MT

		French						German					
		Google	2k	8K	80K	800K	8M	Google	2k	8K	80K	800K	8M
MAP	Processed MT	0.087	0.069	0.067	0.079	0.085	0.084	0.067	0.039	0.050	0.050	0.071	0.079
	Ordinary MT		0.062	0.069	0.079	0.086	0.085		0.034	0.057	0.050	0.070	0.075
PRES	Processed MT	0.413	0.343*	0.369	0.399	0.414	0.419	0.466	0.332*	0.405	0.455	0.471	0.483
	Ordinary MT		0.323	0.360	0.396	0.412	0.413		0.260	0.394	0.445	0.484	0.487
OOV (%)	Processed MT	NA	20.7%	11.6%	5.0%	2.6%	1.6%	NA	40.7%	28.3%	13.6%	7.0%	4.2%
	Ordinary MT		28.6%	16.8%	7.3%	3.0%	1.6%		49.8%	35.8%	18.0%	8.9%	4.2%
Decoding time (*mm:ss*)	Processed MT	NA	00:19	01:05	03:06	04:44	06:03	NA	00:07	00:17	01:01	01:58	02:49
	Ordinary MT		06:43	09:30	15:09	21:31	30:35		02:33	03:46	05:47	07:58	11:24

when using the same training parallel corpus. In addition, with smaller sized training data sets, the speed of decoding using the new MT system reaches up to 23 times faster than the ordinary MT system. Furthermore, the decoding time needed for the processed MT system when it is trained with 8M parallel sentence is comparable to the decoding time required for the ordinary system when it is trained with only 2k examples.

Similar results to those shown in Table 2 were obtained for the training time, where the training time for the processed MT system was 5 to 13 times faster than the ordinary MT system.

5.2 Discussion

Comparing the retrieval effectiveness of the processed vs. the ordinary MT systems when a very small training corpus was used (only 2k) performance was statistically indistinguishable when compared by MAP, but statistically better for processed MT when compared by PRES. This result means that while the systems cannot be distinguished when compared with respect to finding relevant documents at very high ranks, the processed MT is noticeably better when compared to standard MT for finding a greater number of relevant documents at higher ranks. For a recall-oriented search task such as patent retrieval, PRES is a more a meaningful score, since the average number of documents to be examined for this task is often large, sometimes reaching hundreds of documents [1].

The large difference in the average translation time for a French patent compared to that of a German patent stems from the length of the patents, where the French patents are nearly double the length of the German patents on average due to word compounding in the German patents. In addition, the high percentage of the OOV terms in the German patents speeds up the translation since no translation is examined for OOV words.

Removing the stop words from the text reduces the amount of text to be translated by nearly half. However, the gain in speed is much more than the double (5 to 23 times). The reason for this comes from the nature of stop words, where the MT takes a longer time to translate them in order to select the proper translation in the proper position. Additionally, stemming reduces the vocabulary in the MT model leading to less choices of translation for terms, which leads to higher translation speed.

6. CONCLUSION

This paper has presented a novel technique for adapting MT systems for the purpose of CLIR. Although the technique mainly comprises a re-ordering of the workflow of the steps in CLIR, the impact was shown to be significantly more efficient in the resource and computational requirements of the MT process. The new technique was tested on the patent search task that usually requires a large amount of training data and for which the query translation time that can reach more than 50 times the search time. Experimental results show that processing the text by stop word removal and stemming before MT training and decoding leads to speeding up the translation process by up to 23 times. In addition, this technique proved to be much more effective when a limited amount of data is available.

For future work, the approach should be tested for different types of CLIR tasks including ad hoc and web search, especially for languages where limited MT training resources are available.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

1 Azzopardi L., H. Joho and W. Vanderbauwhede. A Survey on Patent Users Search Behavior, Search Functionality and System Requirements. *IRF Report* 2010-0001, (2010)

2 Fujii A., M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-4. In *Proceedings of NTCIR-4*, (2004)

3 Magdy W. and G. J. F. Jones. PRES: a Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. In *Proceedings of SIGIR'10*, (2010)

4 Magdy W. and G. J. F. Jones. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. In *Proceedings of CLEF-2010*, (2010)

5 Oard, D. W. and Diekema, A. R. Cross-Language Information Retrieval. In: *WILLIAMS, M. (ed.) Annual Review of Information Science ARIST*, (1998)

6 Piroi F. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In *Proceedings of CLEF-2010*, (2010)

7 Stroppa, N. and A. Way. 2006. MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of IWSLT*, (2006)

8 Wang J., D. W. Oard. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In *Proceedings of SIGIR'06*, (2006)