Inf Retrieval (2014) 17:492–519
DOI 10.1007/s10791-013-9231-6

Studying machine translation technologies for large-data CLIR tasks: a patent prior-art search case study

Walid Magdy · Gareth J. F. Jones

Received: 21 March 2013 / Accepted: 12 September 2013 / Published online: 21 November 2013

Abstract Prior-art search in patent retrieval is concerned with finding all existing patents relevant to a patent application. Since patents often appear in different languages, cross-language information retrieval (CLIR) is an essential component of effective patent search. In recent years machine translation (MT) has become the dominant approach to translation in CLIR. Standard MT systems focus on generating proper translations that are morphologically and syntactically correct. Development of effective MT systems of this type 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. 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, or the linguistic structure of the output. Thus much of the complexity of MT is not required for effective CLIR. We present an adapted MT technique specifically designed 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 step leads to a significant decrease in the MT computational and training resources requirements. Experimental application of the new approach to the cross language patent retrieval task from CLEF-IP 2010 shows that the new technique to be up to 23 times faster than standard MT for query translations, while maintaining IR effectiveness statistically indistinguishable from standard MT when large training resources are used. Furthermore the new method is significantly better than standard MT when only limited translation training

The work was performed while Walid Magdy was at Dublin City University.

W. Magdy (✉)
Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
e-mail: walid.magdy@gmail.com; wmagdy@qf.org.qa

G. J. F. Jones
Centre of Next Generation Localization, School of Computing, Dublin City University,
Dublin 9, Ireland
e-mail: gjones@computing.dcu.ie

resources are available, which can be a significant issue for translation in specialized domains. The new MT technique also enables patent document translation in a practical amount of time with a resulting significant improvement in the retrieval effectiveness.

Keywords Cross-language patent retrieval · Prior-art Patent search · Cross-language information retrieval · Large-data CLIR · Machine translation

1 Introduction

Interest in patent retrieval research has shown considerable growth in recent years. The focus of research in patent retrieval has included indexing techniques for patents, query formulation, evaluation methodologies, multilingual search, and image search for patent flowcharts and diagrams (Lupu and Hanbury 2013). The interest in multilingual patent search arises from their international and multilingual nature. 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. Therefore 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 another language (Azzopardi et al. 2010; Lupu and Hanbury 2013; Piroi 2010; Piroi et al. 2012; Roda et al. 2009). 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 (DBT) methods. Current statistical MT (SMT) methods lend themselves well to patent translation since patent offices often publish patent content with parallel translations which can be used for MT system training. For example, the European patent office (EPO) makes patent text available in three languages (English, French, and German) (Piroi 2010; Piroi et al. 2012). However, translation using MT is time consuming and resource intensive for cross language patent retrieval (CLPR), where the query text can often take the form of a full patent application running to tens of pages. In addition, a very large parallel training corpus is usually needed to achieve acceptable translation quality (Stroppa 2006). The translation time and the large resources required for translating large data collections such as that in patent prior-art search task has not received much attention to date. Besides, some language pairs have very limited suitable training resources available, meaning that it is not possible to train an effective SMT system for these language pairs leading to low translation quality, and consequentially usually low retrieval effectiveness.

In this paper, we present and analyse a new technique for adapting MT for CLIR. This translation technique addresses the high computational cost and resource requirements of MT for the large data collections in CLPR (Magdy and Jones, 2011a; b). The technique is demonstrated to be up to 23 times faster than standard MT techniques in both the training and decoding phases when tested on the patent search task from CLEF-IP 2010. Retrieval effectiveness using the new translation output is shown to be statistically indistinguishable from results obtained using standard MT. Furthermore, the retrieval effectiveness is found to be statistically significantly better than standard MT techniques when a small amount of training data is used to train the systems. A full analysis of the adapted MT system for CLIR is presented and extensively discussed. In addition, this MT approach enables document translation for CLPR task in a practical amount of time with a resulting

improvement in retrieval effectiveness. While document translation has been shown to be useful in CLIR, its application using standard MT approaches has been impractical to date due to the very large amount of time required for translation of patent documents (Chen and Gey 2004; Parton et al. 2008).

The remainder of this paper is organized as follows: Sect. 2 provides background on CLIR and patent retrieval; Sect. 3 presents the approach for training the MT system for CLIR; Sect. 4 describes the experimental setup and explains the construction of baselines and compares use of MT and DBT for CLPR; Sect. 5 reports, analyses, and discusses the results of the new technique compared to using standard MT techniques for patent query translation; Sect. 6 presents experiments using the new technique for document translation in CLPR; and finally Sect. 7 concludes the paper and provides possible future directions.

2 Background

2.1 Patents

A patent is an exclusive right granted for an invention by a patent office, which is a governmental or intergovernmental organization responsible for granting patents. The language of patents is characterized by the complexity and ambiguity of their contents, where the contribution of a patent is often intended to be unclear (Krier and Zacc, 2002; Lupu and Hanbury 1 2013; Magdy 2012). Unlike normal publications or technical reports, the authors of a patent try to generalize the coverage of their invention and focus on emphasizing the novelty of the ideas disclosed, rather than to help the reader to understand their technique. This leads to the usage of unusual expressions that makes understanding or even finding the patent a difficult job. Research comparing patent text to general English text (Verberne et al. 2010) has shown that very few new tokens are introduced by patent authors. Using the British national corpus as a general English corpus and 4,00,000 patents from four different patent offices as a patent corpus, it was found that 96 % of the terms in the patents were already covered in the general English text. However, the research also showed that the length of sentences in patents is much longer than general English, and that the frequency distribution of terms also differs significantly. Moreover, the word combinations found in patents are uncommon in general text. These findings illustrate the challenging nature of patent language where the same words are often used with totally different meanings. Furthermore, it is very common in patents to find citations to patents from different patent offices that use different languages. This is important since when a patent is being checked for novelty, it should be ensured that the idea has not been disclosed before by any means (patent, publication, article, or webpage) or in any other language (Magdy 2012). The absence of multilingual citations emphasizes the importance of CLIR in patent search.

2.2 Patent retrieval

Evaluation of patent retrieval was proposed in NTCIR-2 in 2001 (Leong 2001). Since then patent retrieval has featured as a research track in all NTCIR campaigns. Similar patent retrieval tasks were introduced at CLEF[1] in 2009 under the name CLEF-IP (CLEF Intellectual Property) (Roda et al. 2009). These tasks have been of interest to IR

[1] http://www.clef-campaign.org/.

researchers since their introduction due to the challenging nature of patent search (Leong 2001; Roda et al. 2009). Various patent search tasks have been created in these campaigns including ad-hoc search, invalidity search, and prior-art search.

In patent ad-hoc search, topics are used to search a patent collection with the objective being to retrieve a ranked list of patents that are relevant to this topic (Iwayama et al. 2003). For invalidity search, the claims of a patent application are used as the topics, and the objective is to search for all relevant documents (patents and others) to find whether the claim is novel or not (Azzopardi et al. 2010; Piroi et al. 2012). All relevant documents are needed since missing only one document can lead to later invalidation of the claim or the patent itself. Prior-art patent search is concerned with finding all relevant patents that have the potential to invalidate the novelty of a patent application or at least that have common parts to that patent (Fujii 2007; Piroi 2010). The full patent application submitted to the 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 is to find these citations of patents automatically. Prior-art search in patent retrieval focuses on finding any kind of patents relevant to the patent application in hand; this is different from invalidity search which focuses on finding any type of document that proves that a given claim in a patent application is not novel.

Reported results for different tasks of patent search show lower retrieval effectiveness compared to other IR applications (Lupu and Hanbury 2013; Magdy 2012). For example, it is generally expected to achieve a mean average precision (MAP) less than 0.1, which is still regarded as an acceptable level of effectiveness. This can be seen clearly from the results of various evaluation campaigns (Azzopardi et al. 2010; Piroi 2010; Piroi et al. 2012; Roda et al. 2009), which illustrates the challenging nature of the patent search task.

2.3 Cross-language information retrieval

CLIR is concerned with searching a collection of documents that are in a different language from the user's query (Nie 2010). Two main approaches are available to cross the language barrier between queries and domains in CLIR: translation of documents to the query language prior to the search stage or query translation to the document language at search time (Oard 1998; Parton et al. 2008). In practice the latter is the most common, since it is more practical in most operational systems, moreover it provides more flexibility to expand the query with multiple possible translations for each term to help prevent problems that can arise from incorrect translations (Darwish and Oard 2003; Wang and Oard 2006).

Considering the translation stage itself, two common techniques have been used for query translation in CLIR: bilingual dictionaries and SMT systems (Oard and Diekema 1998). Bilingual dictionaries are sets of entries of words in one language and possible translations in the other language (Darwish and Oard 2003; Leveling et al. 2011; Wang and Oard 2006). MT systems are optimised for translating whole sentences from one language to another, with the target translated sentence being created in a correct morphological, semantic, and syntactic form. MT has become the most commonly used technique for translation in CLIR in recent years due to the increasing availability of high quality MT systems. Therefore, much CLIR research now uses freely available online tools such as Google Translate,[2] Bing translate,[3] and

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

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

Yahoo Babel Fish.[4] Furthermore, some open source SMT libraries are also available freely for research purposes, e.g., Moses[5] (Koehn et al. 2007) and MaTrEx[6] (Stroppa 2006). The typical procedure adopted in CLIR using SMT technologies is to translate the query using one of the available free MT systems into the target language of the collection, and then to perform search in the document language. Thus most CLIR research has treated the translation stage as a black box without any control over the translation process (Magdy and Jones 2011; Ture et al. 2012). There are other translation techniques for CLIR such as corpus-based translation which applies statistical analysis of words or phrases in parallel or comparable corpora in different languages to obtain probabilities of translations (Oard and Diekema 1998).

2.4 Cross language patent retrieval

Cross-language patent retrieval has featured as one of the tasks in existing patent evaluation campaigns (Azzopardi et al. 2010; Piroi 2010; Piroi et al. 2012; Roda et al. 2009). Like standard CLIR research, most of the research in CLPR used the SMT systems as black box. In addition, little attention has been directed towards the time taken for the translation process. This is a significant issue for the very large topics encountered in patent search, where topics that require translation can be full documents, which can make the query translation time significantly longer than the search time. In addition, the different nature of the patent text requires domain-specific training resources for the MT system, which may not be available for some language pairs. The valuable feature of the presence of parallel translations for a large number of international patents is generally neglected by most CLPR researchers. For example, patents such as those provided by the EPO are published in three languages and translations are often effectively aligned at the sentence level. This parallel translation data can be used to build translation models for multilingual search of patents. One research study (Jochim et al. 2010) utilized these parallel corpora from EPO patents in order to translate queries for patent search. This investigation used the data to build domain-specific translation dictionaries rather than using it for MT training. The word alignment tool Giza++ was used to build a word-to-word translation dictionary for the language pairs English-French and English-German. The highest probability translation for each word was used in the translation process. The reported results of this study for CLPR are considerably lower than those reported when MT is used (Piroi 2010). This observation may stem from the use of different query formulation techniques for patent topics or from using the DBT method which fails to utilize the rich context of long patent queries that can help in the section of appropriate translations when using MT.

A short description of our modified MT technique for CLPR presented here was published in Magdy and Jones (2011). Here we discuss the technique in more detail and analyse the performance to better understand the reasons behind its superior performance in terms of both retrieval effectiveness and computational requirements. Moreover, we compare the MT technique to DBT techniques, which are known to be computationally less expensive. Furthermore, we apply our technique to the translation of the non-English patent documents in the retrieval collection to enrich the English content of these patents, which leads to further improvement in results.

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

[5] http://www.statmt.org/moses/.

[6] http://www.openmatrex.org/.

2.5 Related work

While not widely explored, some existing work has been reported on adapting the translation stage in CLIR for the purpose of achieving high retrieval effectiveness. In TREC-2002 for the Arabic/English CLIR task, a bilingual dictionary built by aligning Arabic and English stems was provided to participants (Oard and Gey 2002). The idea of aligning stems was based on the fact that stems are often more valuable in IR than the full words forms used in grammatical text, which is similar to what is proposed in our study. However, in the TREC 2002 work, stop words were not filtered out which led to them being treated as possible translations in the bilingual dictionary. Similar work was presented in the same track by Franz and McCarley (2002), where they aligned stems instead of words for what they called the convolutional model for CLIR, which integrates the retrieval and translation models into one model. This approach used SMT-like technology, but only for predicting the "bag of words" from which an English translation of a given document can be composed rather than for actually generating translations. This approach achieved better retrieval results than when standard MT was used for translation.

Other research has been reported which aims to improve the quality of the translations for CLIR. The majority of this work has focused on finding better candidate translations using DBT approaches (Darwish and Oard 2003; Leveling et al. 2011; Wang and Oard 2006). Darwish and Oard (2003) adapted a probabilistic structured query method to allow searching with weighted candidate translations of query terms. This approach showed its effectiveness over using only the highest probable translation. A more advanced approach was introduced later by Wang and Oard (2006), where they combined bidirectional translation and synonyms for generating better dictionary-based candidate translations, and achieved higher retrieval effectiveness than the results reported in Darwish and Oard (2003). Little further work has been reported on DBT for CLIR in recent years where the MT approach has dominated due to the significant improvement of the quality of MT systems. For CLIR tasks, MT is straightforward to use and usually achieves high retrieval effectiveness. However, as demonstrated later in this paper, it has shortcomings for CLPR.

Recent work has focused on opening the black-box of the MT system and utilized the candidate translated phrases produced by the system instead of only taking the most probable translation for CLIR tasks (Ma et al. 2012; Ture et al. 2012). This work showed significant improvement in the retrieval effectiveness that achieved a 40 % increase in MAP as reported in Ture et al. (2012).

In this paper, we adapt the current high quality MT systems for CLIR to make them more efficient in computational and resource requirements, while maintaining their effectiveness and demonstrate their utility for CLPR.

3 Adapting MT system for CLIR

This section presents the method for adapting MT technologies for CLIR. The objective is to achieve high retrieval effectiveness for CLPR while using lower resource and computational requirements than are needed by standard MT systems while maintaining or improving search effectiveness.

3.1 Basic concept

The basic idea of the new approach is to train an MT system for translation of queries 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 (Manning et al. 2009). These operations aim to improve retrieval efficiency and effectiveness by matching different surface forms of words. While these are standard processes in IR, for 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 sentences will appear the same after IR pre-processing as "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, and that a smaller processed training corpus can 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 will be 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 (Nie 2010). 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 (Gao et al. 2001). The translation approach introduced here is shown in the lower part of Fig. 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 sequence 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.

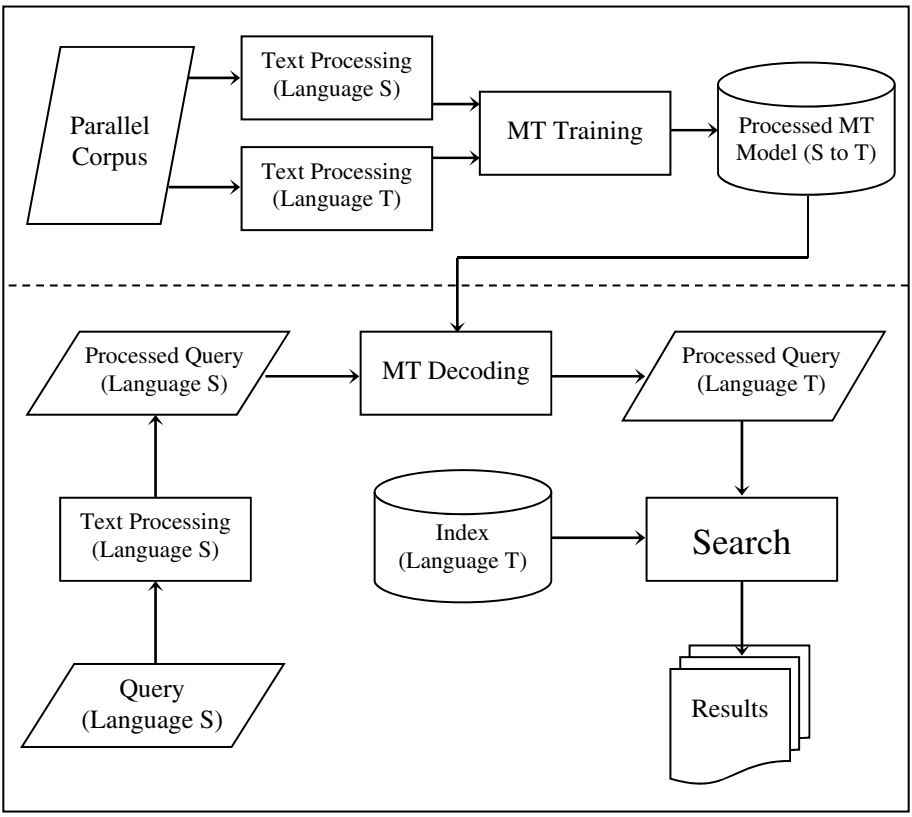

Fig. 1 Workflow of the proposed CLIR system

4 Experimental investigation

The following experimental investigation is designed to test three dimensions of the proposed approach. The first is to examine the effect of processing the words before the MT step on the quality of the translated text, which will be reflected in the retrieval effectiveness. The second is to investigate the efficiency of the proposed translation process in terms of the computational requirements for the training and decoding (translation) 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 examines the effect of using a limited amount of training data on the retrieval effectiveness.

The non-English topics in the CLEF-IP 2010 task are used for the experimentation. Retrieval effectiveness is measured using MAP and PRES (Magdy and Jones 2010a; b). PRES is an evaluation score designed for the recall-oriented tasks when the objective is to find all possible relevant documents in the highest possible ranks. PRES emphasises the ability of the IR system to retrieve a large portion of the relevant documents at relatively high ranks based on a user specific cut-off (N_{max}) (Magdy and Jones 2010). In our analysis, we focus on PRES since it is designed for measuring retrieval effectiveness in patent

search. We include MAP results here in order to allow direct comparison of our work with previously reported results for this task that did not report PRES (Jochim et al. 2010; Roda et al. 2009). Significance is tested using a 2-tailed t test and Wilcoxon tests with p-value 0.05 (Hull 1993).

4.1 Test data

The cross language search sub-task in CLEF-IP 2010 is used for our experiments (Piroi 2010). The main objective is to find relevant patents in an English collection that are related to patent applications filed in French and German languages. The collection consists of 1.35 M patents from the EPO with 69 % of them totally in English and 31 % in German and French. However, the German and French patents in the collections have various sections manually translated into English by the patent office, including the patent title, abstract, and claims. Full details of the collection and the structure of patent documents can be found in Magdy (2012). The English text of all patents in the collection was indexed to create a baseline index of documents in English only. The CLEF-IP track provided two sets of topics: 300 training topics among which 89 are German, 15 are French, and the remainder are English; and 2,000 test topics among 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 CLIR experiments, the 89 German training topics and the 134 French test topics were selected to have a close number of topics for each language (520 German topics would be too many and 15 French topics would be very few) and to allow the replication of the experiments.

4.2 Extracting parallel corpus for training translation model

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). Almost all the patents in the collection contain the title in the three languages, since these translations should be provided to the EPO from the first patent application (the A1 version). However, only some of the patents in the collection contain the claims in all three languages, since the claims translations is only provided in the granted version of a patent (the B version), which did not exist for all of the patents in the CLEF 2010 patent collection. It was easy to identify parallel translations of the patent title, since it is only a short field in the patent application which usually only consists of a small number of words. Identifying the parallel sentences in the claims section was a more difficult task, since the claims section is formed from several separate claims, each one of which is formed from smaller parts [called "claim-text" in the patent XML file (Magdy 2012)], see Fig. 2.

Sometimes the number of claims does not match for the three languages, and sometimes, even if the number of claims matches, the number of sentences ("claim-text") for a given claim does not match for all three languages due to rephrasing in translations. To overcome these problems, only the parts of the claims section which have the same number of claims and claim parts were used to create the parallel corpus, since this is an indication that they are parallel. Any other parts where the number of claims or claim parts do not match were ignored. A more complex sentence alignment strategy could have been adopted; however this simple method enabled us to collect sufficient parallel training data for our experiments.

We applied two additional filtering steps to the sentences to assure the quality of the parallel extracted sentences:

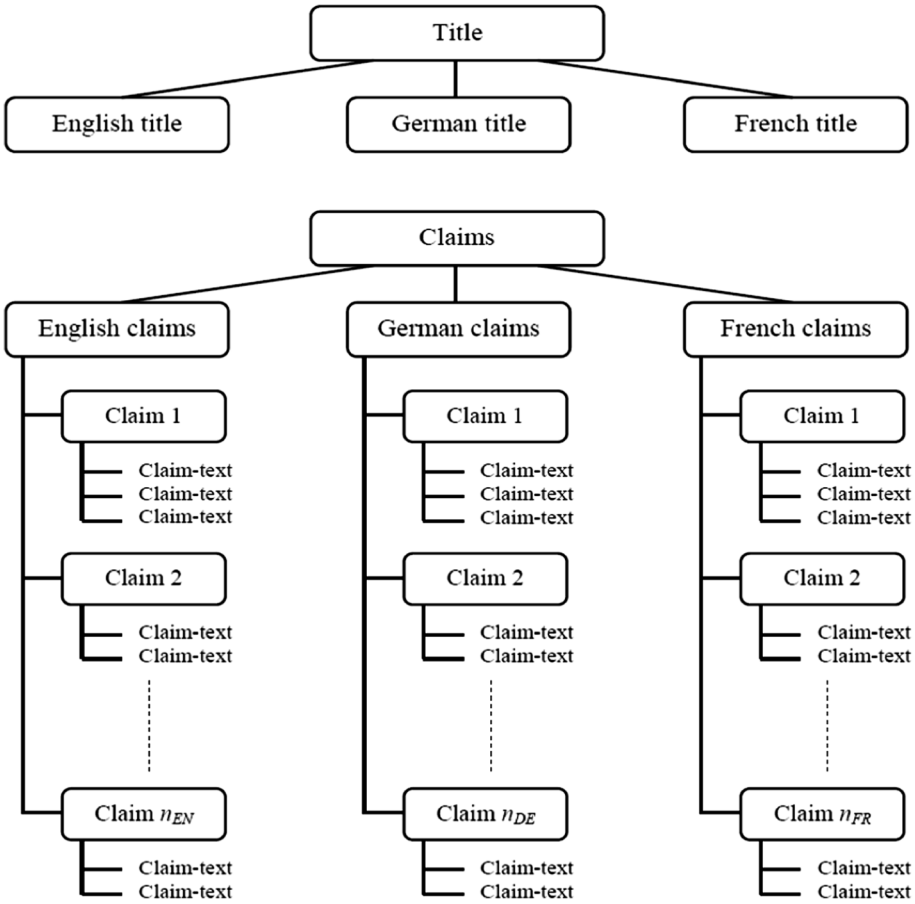

Fig. 2 Structure of patent sections which contain parallel translations

1. Deleting parallel sentences in the three languages if any two sentences differ significantly in length. The maximum allowed ratio between the lengths of two parallel sentences was 1:9, as recommended in Stroppa (2006).
2. Deleting very long sentences that contain more than 60 terms, since long sentences may not be exact translations. Additionally, they require alignment at the word level, which takes significant time. Deleted long sentences represented less than 5 % of the extracted corpus.

Finally, a set of more than 8.15 M parallel sentences in the three languages was extracted from the CLEF-IP 2010 patent collection (referred to later as 8 M corpus). 1.33 M sentences were extracted from the title section of the patents and the remainder from the claims section. This does not overlap with any of the patent applications of the test topics, which do not exist in the patent collection and do not include translations. The average length of the extracted English sentences in the corpus was 28 words. All extracted sentences were converted to lower case to reduce the vocabulary when generating translation models as suggested in Wang et al. (2006). The translation quality was not measured using MT evaluation scores such as the Bleu score (Papineni et al. 2001), since no manual

translation is available for the CLEF-IP topics. Moreover, the main objective here is the retrieval effectiveness, which is reported and analysed in detail below.

For the stemming and stop word removal in our later experiments, the Snowball[7] toolkit was used to stem to the English, French, and German text. Snowball contains state-of-the-art implementations of stemming algorithms for many European languages (e.g. Porter stemming for English). The stop word lists[8] used for each of the three languages contained: 571, 463, and 603 stop words for English, French, and German languages respectively.

4.3 The MaTrEx MT system

The MT experiments were performed using the MaTrEx (Machine Translation using Examples) MT system developed by the MT group at Dublin City University (Stroppa 2006).

The default configuration built within the MaTrEx system was used for the experiments. The configuration of the workflow of the MT system was as follows:

1. Building language model: a standard trigram LM was built from the target language corpus.
2. Building translation model was performed as follows:
    a. Word alignment using GIZA++, which learns the translation tables of IBM model-4 for word alignment (Och and Ney 2003).
    b. Building lexical translation tables for bidirectional translation for both languages.
    c. Building aligned statistical phrases, using the *grow-diag-final* algorithm to produce the final phrase aligned table (Stroppa 2006). The maximum phrase length used was 7-grams, which is the default value of the MaTrEx system.
    d. Building lexicalized reordering model.
    e. Building the generation models, where forward and backward probabilities are computed.

The generated language model and translation model were used in the decoding step to produce the highest probable translation based on the context of the sentence.

4.4 Baseline construction

Query formulation from the patent topic is one of the main challenges in patent search (Azzopardi et al. 2010; Lupu and Hanbury 2013). To construct our baseline run, we tested a number of query formulation approaches based on the best runs submitted to CLEF-IP 2010 (Lopez and Romary 2010; Magdy and Jones 2010; Piroi 2010; Teodoro et al. 2010). For the query formulation title, abstract, description, claims, and classification sections were used according to the best runs in CLEF-IP 2010 (Lopez and Romary 2010; Magdy and Jones 2010; Piroi 2010). We performed several experiments to test different combinations of these fields for constructing the query. The best result was achieved by using terms in the topic after translation that appeared more than two times in these sections combined and all bigram terms that appeared more than three times, with the term frequency acting as weight for these terms (Magdy and Jones 2010). The Indri search toolkit was used for indexing and searching the collection. The Indri retrieval model combines

[7] http://snowball.tartarus.org/.

[8] http://members.unine.ch/jacques.savoy/clef/index.html.

Table 1 Baseline runs for the 89 German topics and 134 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

inference networks and statistical language modelling approaches to retrieval (Strohman et al. 2004). Both Lopez and Romary (2010), Magdy and Jones (2010) showed that automatic citation extraction from patent topics can lead to significant improvement in the retrieval effectiveness. Lopez and Romary (2010), Magdy and Jones (2010) used different information extraction techniques using text patterns and regular expressions to automatically extract citations cited in the text of patent applications, where some of these citations are part of the set of relevant documents to the patent topic. However, in our experiments we did not include this technique since the focus here is specifically on the IR process and not on information extraction techniques.

Two baseline runs were prepared for each 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 (Magdy and Jones2010; Piroi 2010; Teodoro et al. 2010). For the second and main baseline, we used the MaTrEx MT system (Stroppa 2006). The 8 M extracted sentences were used to train the MaTrEx MT system to create two translation models: (French → English) and (German → English). The translation models were then used to translate the 89 German topics and 134 French topics into English, prior to searching the collection. 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 Google and MaTrEx MT systems achieved similar retrieval effectiveness. However for German topics Google Translate achieved lower performance for 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. The results are very similar to those achieved by CLEF-IP 2010 participants when IR is used without citation extraction (Piroi 2010).

Although Google Translate is fast, since it is powered over a cluster of very powerful machines, the translation for our experiments took several days since it only allows the translation of a limited number of sentences at a time. For the translation time using MaTrEx, it was found that the average translation time per topic was 31 min for the French patent topic (contains 7,058 words on average) and 12 min for the German patent topic (contains 3,571 words on average) on a server machine (Intel Xeon quad-core processor, 2.83 GHz, 12 MB cache, and 32 GB RAM). However, the average search time using all the translated text as a query was 42 s for French topics and 14 s German topics on a desktop machine (Intel Core2Due, 3 GHz, 6 MB cache, 3 GB RAM). This highlights the importance of developing faster translation techniques for patent topics.

4.5 Dictionary-based translation baselines

Although SMT is currently the most commonly used technique for translation in CLIR and is the main focus of this paper, we also report results for an investigation of DBT of search queries. These are included for two reasons: the computational and development costs of

DBT are significantly lower than those for SMT, while good CLIR results have been reported for DBT methods (Darwish et al. 2003; Leveling et al. 2011; Wang and Oard 2006). Thus, it is interesting to know if comparable retrieval effectiveness can be achieved using DBT for CLIR patent search. Furthermore, very limited investigation has been reported for this technique for the CLEF-IP prior-art patent search task to date. Only straightforward usage of the technique was reported in Jochim et al. (2010) where poor query formulation for patent topics led to low retrieval effectiveness compared to a monolingual baseline.

In our experiments, DBT was applied to the French and German topics using Giza++. The experiments here tested use of the highest probable translation as in Jochim et al. (2010), and using multiple weighted translations as in the approaches described in Darwish and Oard (2003), Leveling et al. (2011), Wang and Oard (2006). The experiments using the approaches described in Darwish and Oard (2003), Wang and Oard (2006) showed that working with large numbers of low probability translations yields low effectiveness and higher computational cost. Based on this, we imposed a cumulative probability threshold for translations of 0.99 as suggested in Wang and Oard (2006) and 0.6 which achieved the best results in Darwish and Oard (2003). However, cumulative probabilities led to some words with large numbers of possible translations in some cases consisting of thousands of terms, which can correspond to the vague and ambiguous meaning of some terms in the patent text. Therefore, to reduce the computational cost for our experiments, we used a cumulative probability of 0.99 and also applied a hard threshold to allow not more than N candidate translations for a given term. We tested values of N between 2 and 10. The Indri "wsyn" operator was used to allow the presence of multiple weighted alternatives for each term in the query (Strohman et al. 2004). In the Indri "wsyn" operator, the retrieval score of each of the terms is computed and the geometrical mean is calculated for all terms according to the weight of each term (Strohman et al. 2004).

Table 2 shows the results of using DBT for CLPR using the same data as was used to train the MaTrEx MT system (8 M parallel corpora). All the results were found to be statistically lower than those for the MaTrEx MT shown in Table 1, which confirms that MT is a better method for translation in CLPR. Surprisingly, it was found that adding more alternative translations in the query always leads to worse results, which contradicts many reported results in CLIR, including Darwish and Oard (2003), Wang and Oard (2006). In addition, we found that the search time significantly increases when using multiple translations since the query length increases significantly. The average time of search when $N = 10$ for the French topics was 19 min (28 times slower than when using one translation/term), and for the German topics was 10 min (45 times slower). These results show the advantages of using MT in CLPR, where it is better at selecting the correct translation based on the rich context in patent topics rather than using weighted candidate translations,

Table 2 Retrieval effectiveness when using DBT for the 89 German topics and 134 French topics

Translation/word	French		German	
	MAP	PRES	MAP	PRES
1	0.078	0.384	0.061	0.449
2	0.066	0.379	0.055	0.453
3	0.060	0.362	0.049	0.428
5	0.056	0.338	0.045	0.407
10	0.038	0.288	0.045	0.357

which creates more ambiguity to the long query. DBT methods were reported to be effective for standard ad hoc CLIR tasks, where the queries are typically a small number of words that lack context to assist in selecting the proper translation of a term in an MT system. This may be why adding multiple weighted translations for query terms of these tasks showed effectiveness in the retrieval results.

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 by removing stop words and applying stemming ("processed MT"). This was then compared to the standard MT system without pre-processing the data ("ordinary MT"). Two additional MT training setups were also investigated to understand the effect of each pre-processing stage in the processed MT. The system "stemmed MT" trained the MT system using the same data but after applying only stemming without removing the stop words; and "stopped MT" applied the training using the parallel corpus after removing stop words but without applying stemming to the words. The idea behind these two additional MT training sets was to investigate the effect of each of stemming and stop word removal individually on the efficiency and effectiveness of the MT in the CLPR task. Table 3 summarizes the MT training setups used in our experimentation.

To explore the behaviour of the MT systems and CLPR performance when less training examples are available, which will be the case in practice for some language pairs, a number of subsets of the 8 M training dataset of different sizes were selected at random and used to train the MT system. In addition to the full 8 M training set, subsets of the following sizes: 800k, 80k, 8k and 2k sentences were extracted at random from the full corpus and used to train the MT systems for additional experiments.

In all experiments, we use the best candidate translation generated from SMT only. The results obtained in the DBT baseline when using multiple translations did not motivate us to apply multiple translations with the MT systems. In addition, a study by Magdy and Jones (2011) showed that expanding patent queries with synonyms, which can be seen as multiple translation in CLPR, did not improve the retrieval effectiveness for prior-art patent search. Furthermore, another study by Jones el al. (1999) showed that using the best candidate translation using MT is more effective than using multiple translation candidates. For all aforementioned reasons, we did not find the motivation for testing using multiple translations with MT systems.

5.1 Experimental results

Figure 3a, b present the retrieval effectiveness when translating the French and German topics using the ordinary MT system compared to using the alternative MT systems for

Table 3 Pre-processing applied to different MT systems in our experimentation

MT	Pre-processing applied to the training corpus
Ordinary MT	Punctuation and digits filtered out, and text in lower case
Stemmed MT	Ordinary MT + applying stemming to text
Stopped MT	Ordinary MT + filtering out stop words
Processed MT	Ordinary MT + applying stemming to text + filtering out stop words

different sizes of training data, evaluated using MAP and PRES. It can be seen that the difference in retrieval effectiveness using these translation methods is not significant for almost all training sizes. However, with small size training sets (2k), it can be observed that the processed MT and the stemmed MT achieved significantly better retrieval effectiveness than the ordinary MT and the stopped MT when compared using PRES for both query

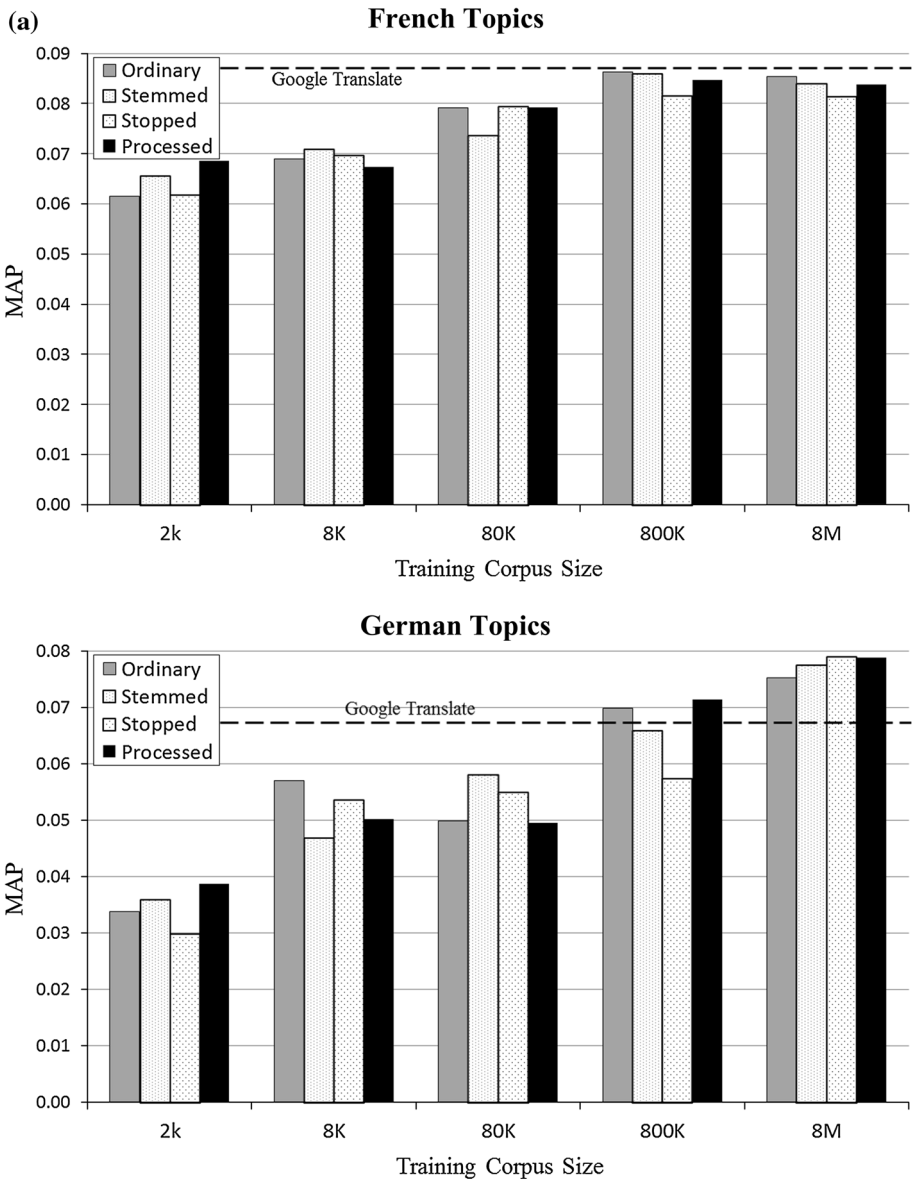

Fig. 3 **a** MAP for French and German topics when using the four MT systems. **b** PRES for French and German topics when using the four MT systems. Retrieval effectiveness for French and German topics compared when using ordinary MT, stemmed MT, stopped MT, and processed MT for the cross language patent search task

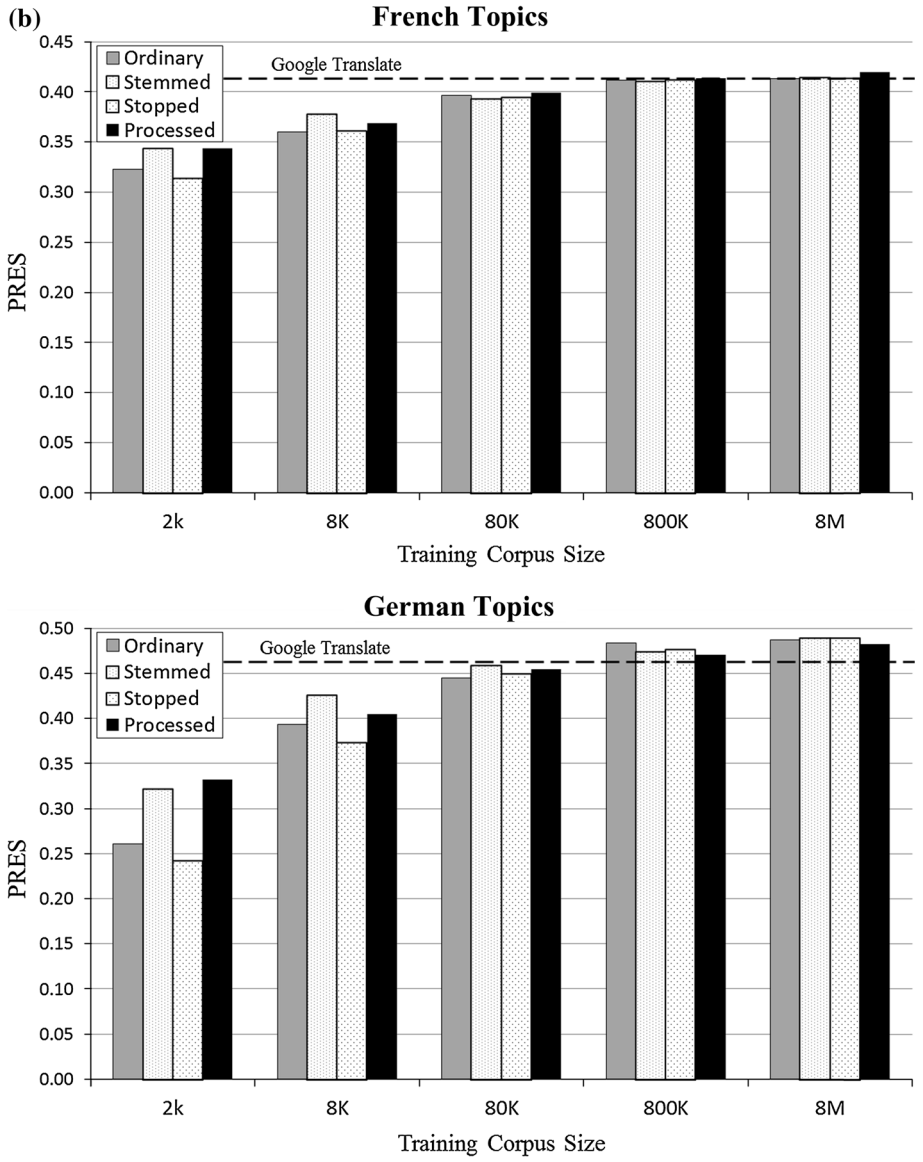

Fig. 3 continued

languages. In addition, for the French topics when using processed MT, results remain statistically indistinguishable from Google Translate for training sizes 8 M, 800k, and 80k. However, for the ordinary and other MT systems, the 80k training set translation led to a retrieval result that is statistically worse than Google Translate when compared by PRES. These results highlight the effectiveness of stemming on MT in CLPR when smaller sizes of training data are available.

One of the possible reasons for these results may be the presence of out-of-vocabulary (OOV) terms when attempting to translate words which do not appear in the MT training data.

To explore the issue of OOV terms while translating the patent topics, the OOV percentage for each method is reported in Fig. 4. The OOV rate of the stemmed MT is the same as that of the processed MT, and the OOV rate for the stopped MT is the same as that of the ordinary MT. Hence, Fig. 4 reports only the OOV rates for the ordinary and processed MT systems. It can be seen that the stemming helps to overcome some of the OOV terms, which leads to the presence of a translation. Also, it can be seen that for small size training sets, the ordinary translation approach suffers from a large percentage of OOVs, while the processed MT and stemmed MT systems overcome part of this problem. The German topics suffer from higher OOV than the French ones due to the presence of word compounds in German. German decompounding is common state-of-the-art solution for the OOV problem in German language. In other work, it has been shown to be effective for overcoming part of the problem applications such as CLPR (Leveling et al. 2011; Magdy 2012). It was shown in these studies that decompounding for German is useful when only limited training data is available, while with large training datasets, decompounding fails to improve effectiveness. However, the same findings for efficiency in computational cost were achieved for processed compared to ordinary MT system when German decompounding was applied.

The second main benefit of the new approach to translation is shown clearly in Fig. 5, which compares the average decoding time required to translate a patent topic into English using these different MT systems. It can be seen that the processed MT and stopped MT systems are at least 5 times faster than the ordinary MT and stemmed MT systems when using

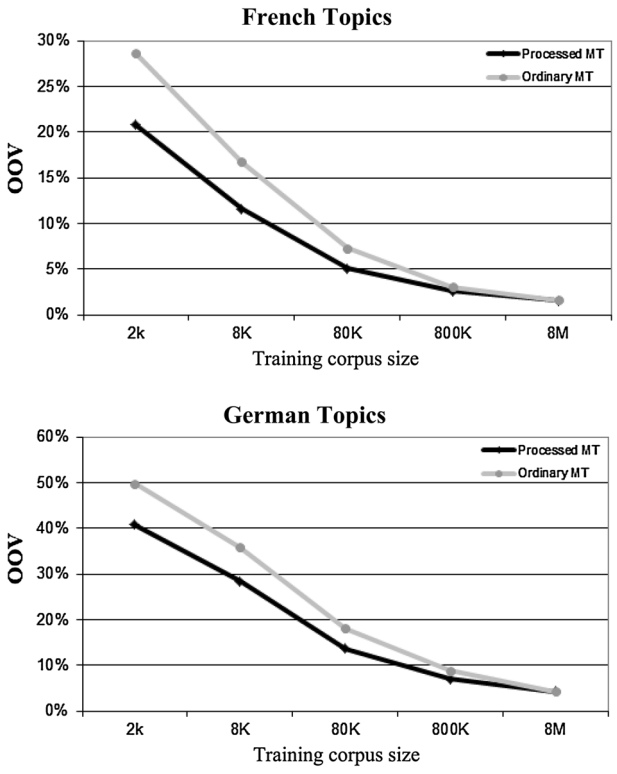

Fig. 4 Out of vocabulary (OOV) rates of French and German topics with different sizes of training data sets compared for processed and ordinary MT

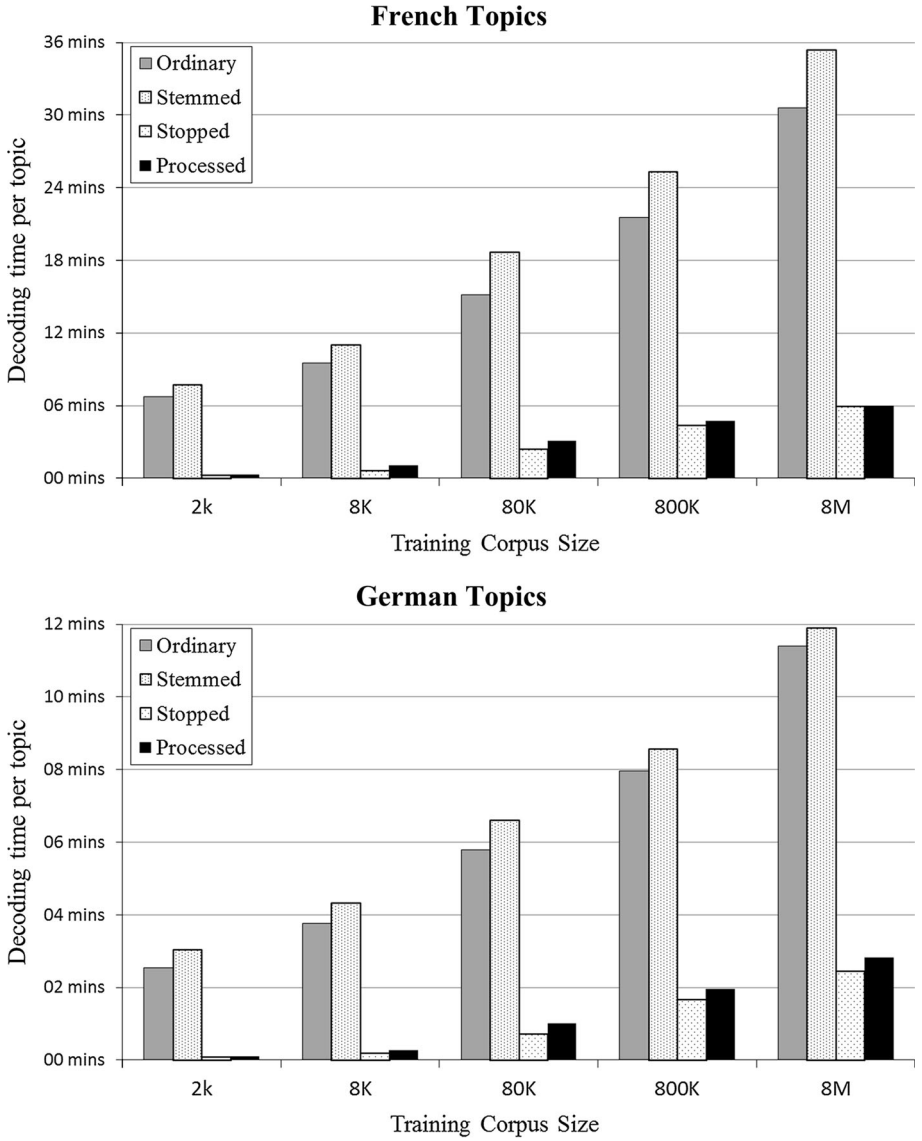

Fig. 5 Average decoding time for translating French and German topics with different sizes of training data sets compared for ordinary MT, stemmed MT, stopped MT, and processed MT

the same training parallel corpus. In addition, with smaller training datasets sizes, the speed of decoding using the MT systems with no stop words reaches up to 23 times faster than the MT systems which include stop words. In fact, the decoding time needed for the processed and stopped MT systems when it is trained with 8 M parallel sentence is less than the decoding time required for the ordinary system when it is trained with only 2k examples. This result demonstrates the strong impact of filtering out stop words before the translation process on the translation speed. Fig. 5 also shows that the fastest MT system is the stopped MT, which is slightly faster than the processed MT; and the slowest MT system is the stemmed MT, which

Table 4 Retrieval effectiveness, OOV, and decoding time for French and German topics compared when using ordinary MT compared to 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

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

is slightly slower than the ordinary MT. This result shows that stemming leads to a slight reduction in the translation speed. However, this is not comparable to the effect of stop word removal which leads to a greater increase in translation time.

Similar results to those for decoding time shown in Fig. 5 were obtained for the training time of the MT systems. The training time for the processed and stopped MT systems was 5–15 times faster than the training time for the ordinary MT system.

All the values of the results reported in Figs. 3, 4, 5 for the ordinary and processed MT systems are presented in Table 4 for a precise comparison between the MT systems. The significant changes in the retrieval effectiveness between the systems and Google Translate are marked in the table.

6 Discussion

The results in this section lead to two main findings as follows:

1. Stemming the text before translation leads to improved retrieval effectiveness when only a limited parallel corpus is available to train the MT system. However, the effect of stemming on retrieval for large training corpora is not significant. In addition, stemming leads to a slight slowing down of the translation speed, which was found to be between 10 and 20 % slower than when no stemming is applied.
2. Stop word removal before translation leads to a large speeding up of the translation system for any size of training data without having a significant effect on the retrieval effectiveness.

These two findings show the importance of applying both stemming and stop word removal together, which is labelled as the "processed MT" approach, and achieves both effectiveness and efficiency in the translation and retrieval processes in CLPR.

The effect of stemming was analysed by checking the OOV rates to understand the reason behind the improved retrieval effectiveness for limited training data. This analysis showed that stemming overcomes a significant proportion of the OOV terms that will not be translated if no stemming is applied. For example, if only the word *played* appeared in the training data as the surface form for the term *play*, any other form of the word will not be translated if it appears in the sentences to be decoded by the MT system, such as: *play*, *plays*, *playing*. When applying stemming, all these terms will be normalized to the term *play*, and will be translated regardless of the surface form that appears in the text to be translated.

Regarding the processing time when applying stemming which appears to be slightly slower than when no stemming is applied. This is explained by the translation tables created for the stemmed terms which are expected to be larger than for words, since the entries of the words: *play*, *played*, *plays*, and *playing* will be combined to only one entry which is *play*. This creates some additional confusion for the MT system to select the proper translation based on the context, which requires additional time. The positive thing is that this additional time was found not to be significant.

For the effect of stop word removal, removing the stop words from the text reduces the amount of text to be translated by nearly half. However, the gain in speed for the translation process is much more than the double (5–23 times). A possible reason for this comes from the special 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, since they are the most

confusing terms to be translated by an MT system. This may arise due to the wide variation in the use and behaviour of pronouns between languages.

One of the observations from the results reported in Fig. 5 is the difference in the average translation time for a French patent compared to that of a German patent. This arises from the length of the patents, where the French patents are nearly double the length of the German patents on average because of the compounds in German, which also leads to a higher percentage of OOV terms in the German-English translation process that speeds up the translation since no translation is examined for OOV words.

This section can be concluded by noting the overall positive impact of using processed MT on both efficiency and effectiveness of translation and retrieval. In the next section, we use this high quality and fast MT to translate the patent documents, which enables us to explore an approach that has always been considered impractical with ordinary MT systems due to the translation time required to translate large amounts of text.

6.1 Document translation for CLPR

As described in Sect. 4.1, 31 % of the evaluation patents in the collection are German and French patents. In the experiments reported so far, only the sections of these patents that have manual English translations have been indexed. Thus the sections of the patents which are only available in the original languages of French and German were not indexed for search. This was the standard approach adopted by many participants in the CLEF-IP tracks (Piroi 2010; Piroi et al. 2012; Roda et al. 2009). However, this approach has a significant drawback since for a large proportion of these patents, only the title field has an English translation. This means that these documents are very short since (only the small number of words in the title are available for search) and hence the documents have a very low chance of being retrieved reliably. Some of the submitted runs used a multilingual index formed by indexing the English, French, and German text into a single index without translation and then searched the collection with patent topics in their original languages in an attempt to exploit the non-English content to improve search effectiveness, but the results of these runs were lower than those submitted using only the English language text (Piroi 2010). Other later trials attempted to improve the results by using multilingual queries through translating patent queries into the three languages (Jochim et al. 2010), while indexing the patent documents in their original languages. However, this approach also showed lower results than those reported in this paper and those reported at CLEF-IP 2010, where the best achieved MAP scores for the German and French topics were 0.04 and 0.056 respectively (Jochim et al. 2010) (the PRES score was not reported in this research). This low result may stem from the multilingual query approach itself and also from using translation dictionaries which fail to utilize context in the translation process.

Previous studies in different IR applications using document translation in CLIR have shown that it can improve retrieval effectiveness, particularly when combined with query translation (Chen and Gey 2003; Parton et al. 2008). The main hindrance to continued research into document translation for CLIR has been the impractical translation time required to translate the documents. In Chen and Gey (2003), an "approximate fast translation" for documents was applied. This method was based on using an MT system to translate only the unique terms in the document collection without taking account of their context. The top translation for each term was used to replace the original term in the documents. However, ignoring the context was found to lead to low quality translation. In Parton et al. (2008), a sophisticated SMT system (DARPA Gale MT) was tested to translate Chinese and Arabic document collections into English in a translingual IR task. However,

in this work it was estimated that the time needed to translate the full retrieval collection would exceed 30 years. This excessive translation time led them to drop many of the steps in the SMT process in order to speed up translation. However, they comment that dropping these steps led to poor translation.

In this section, we use the processed MT method to translate the claims and abstract sections of the French and German patents into English where such translations do not exist in the original documents in order to enrich the documents with this information. This resulted in all patents having a comparable document length and amount of information regarding the patent content. Our main objective in this experiment was to explore whether the increased speed of our new approach to MT enables practical document translation for CLIR. As shown in the results in Figs. 3, and 5, an acceptable quality of translation can be achieved for the purpose of CLPR while using the faster translation of the processed MT approach. Hence, we apply the processed MT method using the 2k and 8k translation models, which were found to be very fast, to translate the missing French and German parts of the patents to enable them to be added to the index.

We checked the number of non-English patents in the collection which do not contain either an English abstract or claims section. Our analysis showed that nearly 44 % of the French and German patents do not contain either the abstract or the claims sections. This resulted in translation of non-English sections from 137k German patents and 47k French patents. In total 1.15 M German sentences and 390k French sentences were translated. The sizes of the plain text to be translated for the German and French were 343 and 128 MB respectively, and 211 and 70 MB after stemming and stop word removal respectively. After translation, the size of the English text for the German patents in the collection increased from 353 to 545 MB when using the 2k model and to 541 MB when using the 8k model. For the French patents, the size increased from 127 to 192 MB when using the 2k model and to 190 MB when using the 8k model. The difference in the sizes of the translated texts generated using the two models arises due to the differences in the percentage of untranslated OOV terms for each translation model. Table 5 presents the values of the amount of French and German text to be translated.

Table 6 shows the time taken to translate the German and French documents into English using the processed MT with the 2 and 8k translation models on the server machine described earlier. In addition, the estimated translation time if the ordinary MT were used is reported too based on experiments reported in Fig. 5. As shown in Table 6, it would be unrealistic to use the ordinary MT system to translate this amount of text, since the time required for translation even when using our very small 2k translation model is more than 1 month. This becomes even more unrealistic when considering the slightly larger 8k translation models where the estimated translation time will exceed 2 months. The degree of reduction in the processing time is highlighted clearly in this context where the amount of data to be translated is very large. Using the 2k model the translation time is reduced to only 2 days, and for the 8k model this is reduced to less than 1 week. In addition, it is expected that the retrieval effectiveness would be better or at least the same when compared to using ordinary MT, based on results in Fig. 3. This significant reduction in the translation time makes document translation feasible for CLPR.

Figure 6a, b show the retrieval effectiveness evaluated using PRES and MAP, for two new sets of results when searching the patent collection after adding the translated parts to the German and French patents using the 2k and 8k translation models. All sets of translated queries were used to search the two new collections, including Google Translate and processed MT system with different sizes of translation models. Translation of the

Table 5 Amount of French and German content to be translated and added to the patent collection index

	French	German
Number of patent enriched with translations	47k	137k
Number of sentences translated	390k	1.15 M
Number of words translated	9.89 M	22.74 M
Size of text to be translated	128 MB	343 MB
Size of text translated after processing	70 MB	211 MB
Increase in patents content after adding translated text	51 %	54 %

Table 6 French and German documents translation time with 2k and 8k translation models using processed MT vsrsus ordinary MT (estimated)

Training model		Processed MT	Ordinary MT (*estimated*)
2k	German	1 day 4 h 46 min	27 days 20 h 1 min
	French	22 h 47 min	9 days 17 h 49 min
	Total	2 days 3 h 33 min	37 days 13 h 50 min
8k	German	3 days 7 h 26 min	44 days 19 h 6 min
	French	3 days 4 h 22 min	24 days 22 h 15 min
	Total	6 days 11 h 48 min	69 days 17 h 21 min

query set using ordinary MT is not included here since its performance has already been shown to be similar or lower to that of the processed MT.

Figure 6 compares the new results to the earlier ones in Fig. 3 when only the English text of the patents was indexed. Comparing the results using MAP shows that there is some statistically significant improvement in results for the French query sets, but very limited improvement or no improvement at all for the German ones. However, and as mentioned earlier, we are focusing more on the PRES score since it better reflects the objective in a patent search task. When comparing PRES values, a statistical significant improvement in the retrieval effectiveness can be seen for most of the query sets for both the French and German topics. Considering the French topics, adding the translated parts to the non-English documents showed a large improvement in all the query sets (including those translated by Google Translate) except when both the queries and documents are translated with the 2k model. For the German topics, significant improvements only occurred when documents are translated with the 8k model and queries are translated with 80k or higher models (including Google Translate). The logical explanation of this observation is the impact of the high percentage of OOV terms in both the translated queries and documents for German resulting from the smaller training set. It can be seen that a high OOV rate in the translation model in both documents and queries leads to some matching of untranslated terms, which is like using mixed language queries as reported in Jochim et al. (2010), Piroi (2010), which showed that using mixed language queries and documents leads to an unstable effect on the retrieval effectiveness, where it can sometimes improve the results, but often degrades it, which is consistent with the results reported here.

One of the conclusions of this study is that poor and fast translation models, such as 2 and 8 k, can be used with the processed MT approach, for translating multilingual document collections given that a better translation model will be used to translate the queries. The terms

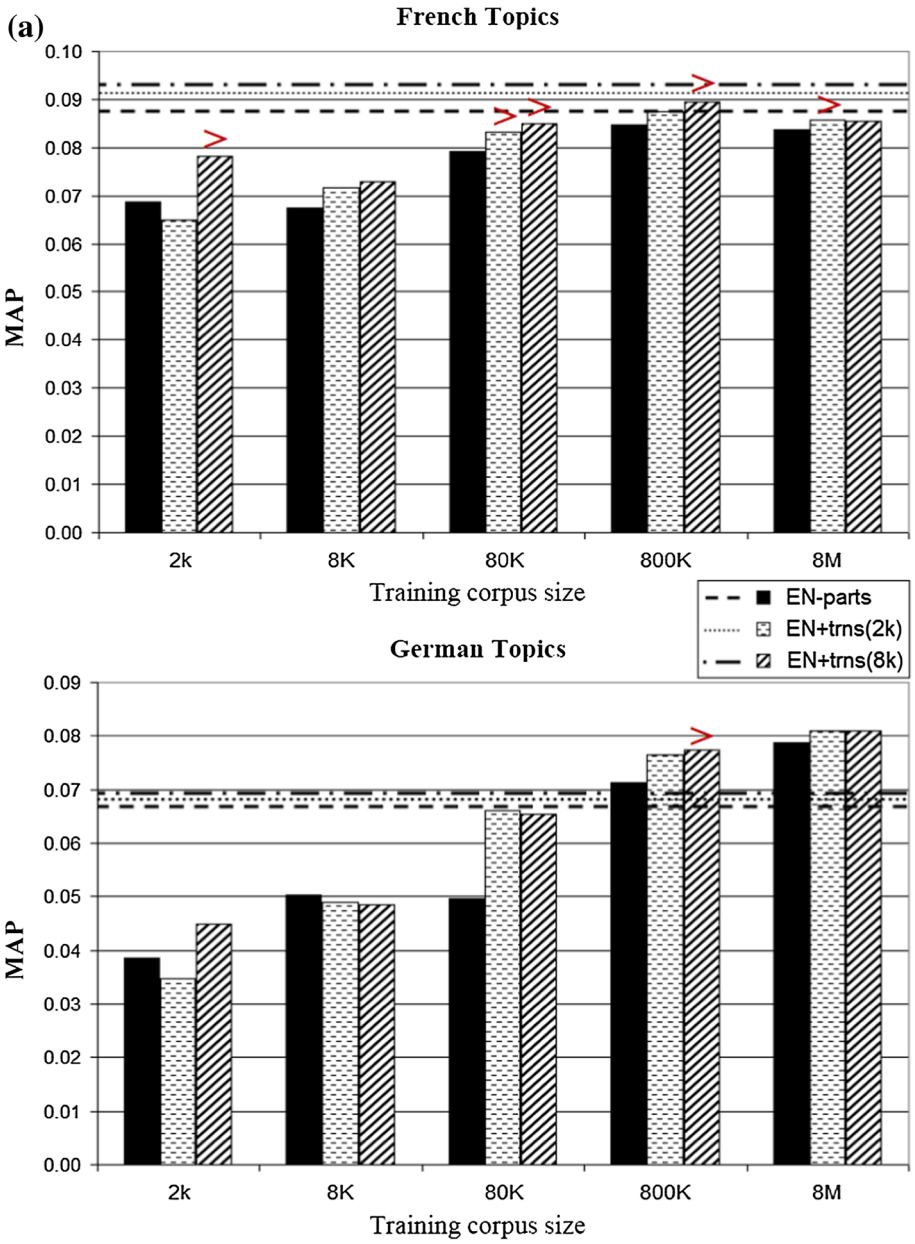

Fig. 6 **a** MAP for French and German topics after adding missing parts to collection. '*Greater than*' refers to statistical significant improvement in results. *Dotted horizontal lines* represents queries translated by Google Translate. **b** PRES for French and German topics after adding missing parts to collection. '*Greater than*' refers to statistical significant improvement in results. *Dotted horizontal lines* represents queries translated by Google Translate

"poor" and "better" depend on the language pair. For example, only the 2k model need be considered poor for French, whereas both the 2k and 8k models are considered poor for German since they have a higher OOV rate because of word compounding.

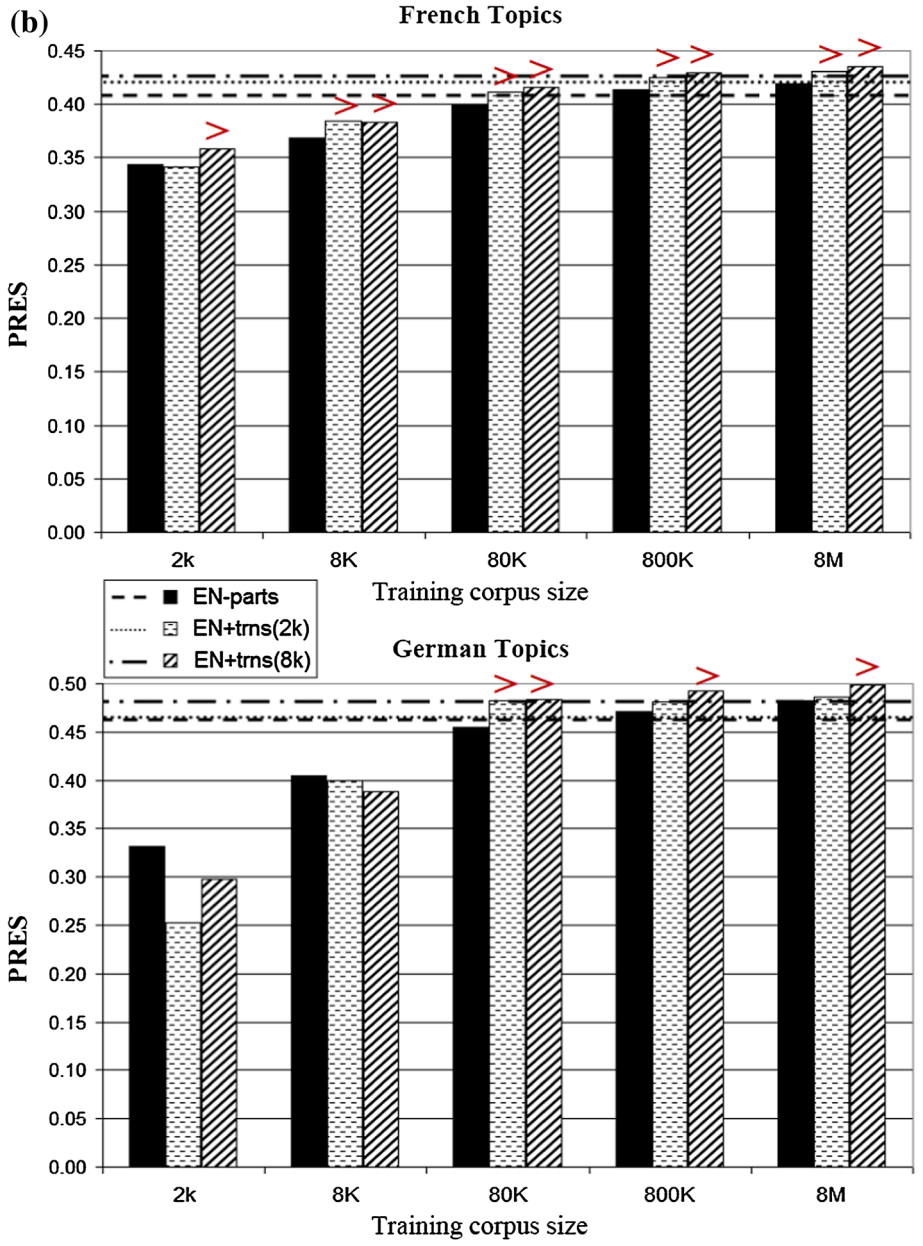

Fig. 6 continued

In overall conclusion, the results of the experiments in this section show that the processed MT approach makes document translation techniques a realistic component in CLPR systems. Depending on the document collection size, and the estimated time for translation, a convenient translation model size can be selected.

Regarding the application of document translation for multilingual patent search, it has been shown that significantly better results can be achieved even using only small size translation models for translating the documents in a reasonable amount of time, but in this case, queries are recommended to be translated with a better translation model to achieve significantly better results. After adding the translated French and German text to the index, our best result for the French and German topics in this paper are PRES 0.436 and 0.499, which are, to the best of our knowledge, the highest achieved scores for these topics for the CLEF-IP 2010 dataset when using IR techniques without inclusion of citation information. Nevertheless, including the translated content in the index file would be expected to lead to further improvements in results obtained in the work previously reported by other task participants, including those which use citation extraction.

7 Conclusion and future work

This paper has studied the usage of MT technologies in large data CLIR tasks as represented in the prior-art patent search task. Initially we demonstrated that MT is a more effective technique for translation in CLPR than DBT methods. We showed that using DBT with different setups, by using one or multiple translations, does not achieve comparable results to MT.

We studied an adapted MT system 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 more efficient in the resource and computational requirements of the MT process. The proposed translation technique for CLIR was extensively tested on the recall-oriented patent search task that requires a large amount of training data for conventional SMT and for which the query translation time that can reach more than 50 times the search time. Experimental results showed that processing the text by case folding, stop word removal, and stemming before MT training and decoding leads to speeding up of the translation process by a factor of up to 23. In addition, this technique proved to be much more effective when only a limited amount of training resources were used for a given language pair. Our analysis shows that stemming is responsible for the improved retrieval results when limited training data is available, and that stop word removal is responsible for speeding up the translation process. This shows the importance of having both techniques applied before translation. The modified MT system is shown to be effective when only two thousand parallel sentences are available for MT training. Furthermore, we showed that the very large reduction in translation time using this approach makes document translation in CLIR a practical proposition with improved retrieval effectiveness.

For future work, a more well founded approach to selecting the sentences used to build the translation model rather than selecting them at random as was done in our experiments could be explored. The reason behind selecting parallel sentences at random in this study was to simulate the situation when limited resources are available for a pair of languages and there is no opportunity to select the set of training sentences. However, an algorithm designed for selection of appropriate sentence pairs could be used to build a more effective translation model for the purpose of fast and accurate document translations in a specific domain. Finally, the approach could be tested on different large-data CLIR tasks, such as cross-language duplicate document detection. In duplicate document detection, the full document is used to search for other similar documents. Applying this across languages requires high computational cost for translation. "Processed MT" may be an ideal solution for such a task. It would be very interesting to examine this practically. Similarly, the

"Processed MT" approach can be very useful for automatically linking news articles across languages, where an article in one language can be fully translated and used to search for the corresponding article in another language at very high speed. Translating a full article using standard MT system would be very slow and inefficient.

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

References

Azzopardi, L., Joho, H., & Vanderbauwhede, W. (2010). A survey on patent users search behavior, search functionality and system requirements. *IRF Report, 1*, 2010.

Chen, A., & Gey, F. (2004). Combining Query Translation and Document Translation in Cross-Language Retrieval. *Proceedings of CLEF-2003*.

Darwish, K., & Oard, D. W. (2003). Probabilistic structured query methods. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval SIGIR'03, Toronto, Canada*.

Franz, M., & McCarley, S. (2002). Arabic information retrieval at IBM. *Proceedings of TREC-2002*.

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR'07, Amsterdam, The Netherlands*.

Gao, J., Nie, J-Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2001). Louisiana, USA*.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval Experiments. *Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR' 93), Pittsburgh, Pennsylvania, USA*.

Iwayama, M., Fujii, A., Kando, N., & Takano, A. (2003). Overview of patent retrieval task at NTCIR-3. *Proceedings of the 3rd NTCIR Workshop*.

Jochim, C., Lioma, C., Schütze, H., Koch, S., & Ertl, T. (2010). Preliminary study into query translation for patent retrieval. *Proceedings of the 3rd international workshop on Patent information retrieval (PaIR '10), Toronto, Canada*.

Jones, G. J. F., Sakai, T., Collier, N. H., Kumano, A., & Sumita, K. (1999). A comparison of query translation methods for English-Japanese cross-language information retrieval. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99), San Francisco, U.S.A*.

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. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic*.

Krier, M., & Zacca, F. (2002). Automatic categorization applications at the European patent office. *World Patent Information, 24*(3), 187–196.

Leong, M.K. (2001). Patent data for IR research and evaluation. *Proceedings of the 2nd NTCIR Workshop*.

Leveling, J., Magdy, W., & Jones, G. J. F. (2011). An investigation of decompounding for cross-language patent search. *Proceedings of the 34th annual international SIGIR conference on Research and Development in Information Retrieval (SIGIR'11). Beijing, China*.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management, 41*(3), 523–547.

Lopez, P., & Romary, L. (2010). Experiments with citation mining and key-term extraction for prior art search. *Proceedings of the CLEF-2010*.

Lupu, M., & Hanbury, A. (2013). Patent retrieval. *Foundations and Trends® in Information Retrieval, 7*(1), 1–97.

Ma, Y., Nie, J., Wu, H., & Wang, H. (2012). Opening Machine Translation Black Box for Cross-Language Information Retrieval. *Information Retrieval Technology. Lecture Notes in Computer Science, 7675*, 467–476.

Magdy W., & Jones, G. J. F. (2011). Should MT systems be used as black boxes in CLIR?. *Proceeding of the 33rd European Conference on Information Retrieval (ECIR'11). Dublin, Ireland.*

Magdy, W. (2012). Toward higher effectiveness for recall-oriented information retrieval: A patent retrieval case study. *PhD Thesis, Dublin City University.*

Magdy, W., & Jones., G. J. F. (2010). PRES: A score metric for evaluating recall-oriented information retrieval applications. *Proceedings of the 33rd annual international SIGIR conference on Research and Development in Information Retrieval (SIGIR'10). Geneva, Switzerland.*

Magdy, W., & Jones, G. J. F. (2010). Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements. *Iroceedings of the CLEF 2010: Conference on Cross-Language Information Retrieval and Evaluation, Padua, Italy.*

Magdy, W., & Jones, G. J. F. (2010). Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. *Proceedings of CLEF-2010.*

Magdy, W., & Jones, G.J.F. (2011). A Study of Query Expansion Methods for Patent Retrieval. *Proceedings of PaIR worjshop 2011, Glasgow, Scotland.*

Magdy, W., & Jones, G. J. F. (2011). An efficient method for using machine translation technologies in cross-language patent search. *Proceedings of the 20th ACM international conference on Information and Knowledge Management (CIKM'11). Glasgow, Scotland.*

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Nie J.-Y. (2010). Cross-Language Information Retrieval. Morgan & Claypool Publishers.

Oard, D. W. (1998). A comparative study of query and document translation for cross-language information retrieval. *Proceedings of the 3rd conference of the association for machine translation in the Americas on MT and the information soup AMTA.*

Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. In M. Williams (Ed.), *Annual review of information science ARIST*, pp. 223–256.

Oard, D. W., & Gey, F. (2002). The TREC-2002 Arabic/English CLIR track. *Proceedings of TREC-2002.*

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 19*(1), 19–51.

Papineni, K., Roukos, S., Ward, T., & Zhu,W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Technical Report RC22176(W0109-022), IBM Research Report.*

Parton, K., McKeown, K. R., Allan, J., & Henestroza, E. (2008). Simultaneous multilingual search for translingual information retrieval. *Proceedings of ACM 17th Conference on Information and Knowledge Management (CIKM'08), California, US.*

Piroi, F. (2010). CLEF-IP 2010: Retrieval experiments in the intellectual property domain. *Proceedings of CLEF-2010.*

Piroi, F., Lupu, M., Hanbury, A., Magdy, W., Sexton, A. P., & Filippov, I. (2012). CLEF-IP 2012: Retrieval experiments in the intellectual property domain. *Proceedings of CLEF-2012.*

Roda, G., Tait, J., Piroi, F., & Zenz, V. (2009). CLEF-IP 2009: Retrieval experiments in the intellectual property domain. *Proceedings of CLEF-2009.*

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2004). Indri: A language model-based search engine for complex queries. *Proceedings of the International Conference on Intelligence Analysis.*

Stroppa, N., & Way, A. (2006). MaTrEx: DCU machine translation system for IWSLT 2006. *Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan.*

Teodoro, D., Gobeill, J., Pasche, E., Vishnyakova, D., Ruch, P., & Lovis, C. (2010). Automatic prior art searching and patent encoding at CLEF-IP'10. *Proceedings of CLEF-2010.*

Ture, F., Lin, J., & Oard, D.W. (2012). Looking inside the box: Context-sensitive translation for cross-language information retrieval. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR'12). New York, NY, USA.*

Verberne, S., D'hondt, E., & Oostdijk, N. (2010). Quantifying the challenges in parsing patent claims. *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval AsPIRe'10.*

Wang, W., Knight, K., & Marcu, D. (2006). Capitalizing machine translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL), New York, USA.*

Wang, J., & Oard, D. W. (2006). Combining bidirectional translation and synonymy for cross-language informzation retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval, Seattle, Washington, USA.*