Preliminary Study into Query Translation for Patent Retrieval

Charles Jochim, Christina Lioma, and Hinrich Schütze
Institute for Natural Language Processing
Universität Stuttgart
70174 Stuttgart, Germany
{jochimcs,liomaca}@ims.uni-stuttgart.de

Steffen Koch and Thomas Ertl
Institute for Visualization and Interactive Systems
Universität Stuttgart
70569 Stuttgart, Germany
{Steffen.Koch,Thomas.Ertl}@vis.uni-stuttgart.de

ABSTRACT

Patent retrieval is a branch of Information Retrieval (IR) aiming to support patent professionals in retrieving patents that satisfy their information needs. Often, patent granting bodies require patents to be partially translated into one or more major foreign languages, so that language boundaries do not hinder their accessibility. This multilinguality of patent collections offers opportunities for improving patent retrieval. In this work we exploit these opportunities by applying query translation to patent retrieval. We expand monolingual patent queries with their translations, using both a domain-specific patent dictionary that we extract from the patent collection, and a general domain-free dictionary. Experimental evaluation on a standard CLEF-IP dataset shows that using either translation dictionary fetches similar results: query translation can help patent retrieval, but not always, and without great improvement compared to standard statistical monolingual query expansion (Rocchio). The improvement is greater when the source language is English, as opposed to French or German, a finding partly due to the effect of the complex French and German morphology upon translation accuracy, but also partly due to the prevalence of English in the collection. A thorough per-query analysis reveals that cases where standard query expansion fails (e.g. zero recall) can benefit from query translation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*dictionaries*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation, relevance feedback, selection process*; J.5 [**Computer Applications**]: Arts and Humanities—*language translation*

General Terms

Experimentation, Performance

Keywords

patent retrieval, cross-language information retrieval, query translation, statistical machine translation, relevance feedback, query expansion

PaIR'10, October 26, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

The task of an information retrieval (IR) system is to retrieve documents in response to a user information need from a previously indexed collection of documents [39]. Patent IR, also referred to as *patent retrieval* or *patent search*, is a specialized branch of IR that aims to support patent professionals in retrieving patents that satisfy their information needs and search criteria [34]. Patent retrieval is generally considered to be a difficult task [34, 35]. One difficulty is the vocabulary used in patents (*'patentese'*) [2], because it often contains highly specialized or technical words not found in everyday language. Another difficulty is the structure of patents. Patents are structured documents that contain several different fields, such as *description, claims,* or *prior-art.* The text in these fields is built over time, may not necessarily be in logical sequence [2], and can be partially translated into one or more different languages (e.g. English, French, and German, in the case of the European Patent Office (EPO)). A further difficulty in patent retrieval stems from the frequently intentional obfuscation of content by patent writers who wish to make their patents difficult to retrieve. This exacerbates the retrieval problem and can throw off robust standard IR approaches and systems [3].

A common scenario in patent retrieval is *prior-art retrieval*, which is performed by patent searchers to determine the novelty of a new invention [37]. One difficulty in this scenario is that patent searchers require an exhaustive knowledge of all related and relevant patents. Overlooking a single valid patent could lead to detrimental and very expensive implications, such as infringement and litigation. In

practice, this means that *recall* is very important for prior-art retrieval. In addition, the increasingly large amounts of patent data available for retrieval, combined with the frequent and deliberate obfuscation of patent content, create an equally important need for increased *precision* in retrieval.

In this work, we ask whether we can improve the precision and recall of patent retrieval, and more specifically of prior-art retrieval, by query translation. We reason that, since patents are partially translated into one or more languages, a collection of patents can be seen as a multilingual corpus, which contains multiple languages across documents (e.g. some patents are written in French, others in English, and still others in German), but also within documents (e.g. a patent originally written in English can contain sections which are translated into French and German). Given such a multilingual patent collection, we propose to expand queries using translations of the original query terms. Our goal is to create multilingual queries, in line with the multilingual patents available for retrieval. Our intuition is that within a multilingual collection, queries in more than one language may be useful to retrieval. This is the reason why we choose to expand queries with translated terms, as opposed to replacing the original query terms with their respective translations. This type of query translation can also be seen as a form of query expansion, because the queries are expanded with their respective translations.

We tackle query translation using a dictionary-based approach, where query term translations are fetched from a translation dictionary. We expect that the more accurate the translation, the better the retrieval performance. Our hypothesis is that a domain-specific translation dictionary on patents will give more accurate translations and hence better retrieval performance, than a general domain-free translation dictionary, because the former will have better coverage of patent domains than the latter. However, maintaining a domain-specific patent translation dictionary is neither trivial nor always feasible: dictionary coverage is affected by the various different and dynamically changing patent subdomains, where even coining entirely novel concepts is not unusual. An additional drawback to static dictionaries is their weakness to deal with the ambiguous language often used by patent writers to deliberately obfuscate details of their patents. To address these points, we propose extracting a domain-specific translation dictionary from the patent collection used for retrieval. We do so by taking advantage of the parallel translations existing between parts of patents in the collection. Specifically, we identify such parallel translations, we align them, and we compute the translation probabilities between terms in the aligned translations. These translations constitute the entries in our domain-specific patent translation dictionary.

To evaluate our query translation hypothesis, we conduct experiments separately with (i) a general domain-free translation dictionary, and (ii) the domain-specific translation dictionary that we extract from the patent collection used for retrieval. In addition, because our query translation can also be seen as a form of query expansion, we conduct experiments with a standard statistical query expansion technique (Rocchio [27]). Experimental evaluation on a standard CLEF-IP [28] dataset indicates that using either translation dictionary fetches similar results.

The remainder of this paper is organized as follows. Section 2 describes related work on patent retrieval, and the use of query translation and query expansion in IR. Section 3 presents the methodology of our proposed query translation approach. Section 4 describes and discusses the experimental evaluation of our approach. Section 5 contextualizes our approach and outlines future research. Finally, section 6 summarizes this work.

2. RELATED WORK

There is increasing scientific interest in patent retrieval [34, 35], the difficulty of which has long attracted natural language processing (NLP) approaches [17, 24]. In this work, we use NLP, and specifically statistical word alignment, to translate patent queries, i.e. we focus on patent multilinguality. There are several initiatives focusing on multilingual patent retrieval. For instance, the Cross Language Evaluation Forum (CLEF)[1] sponsored an Intellectual Property (IP) track [28] in 2009 with three subtasks dedicated to crosslingual IR (CLIR). Similarly, NTCIR[2] has had separate workshops for both CLIR and patent retrieval since 2002, and the two most recent meetings included a patent translation task [12, 20].

More generally, translating queries is an idea that has been studied for some time [9] and it is typically realized using translation dictionaries, machine translation (MT) systems, parallel corpora or combinations of these (see [16, 22] for an overview). Mainstream approaches to CLIR aim to maximize translation accuracy in order to improve retrieval performance [16], however more recent approaches have also focused on improving retrieval performance using "approximate" rather than accurate translations [13, 40]. Specifically, in [13], Gao et al. present a system for cross-lingual query suggestion reliant on web query logs. Queries are not literally translated, but instead a multilingual web query log is used to find target queries similar to the original source queries. Their system relies on word translations derived from the Europarl corpus, as well as co-occurrence statistics, and click-through information from the web query log to estimate the similarity between queries crosslingually. This method outperforms MT-based and dictionary-based query translation. However, it would be difficult to use a similar method with patent retrieval because of the lack for query log data for patent retrieval. Although, in principle, query logs and click-through data are available from the web, in practice, collecting this information from patent searchers might prove difficult. Even releasing what and how one is searching can possibly be a liability for patent professionals.

The second approach to depart from exact query translation, presented by Wang and Oard [40], considers translation as a problem of *meaning matching*. Bidirectional term alignments are extracted from Europarl (for English-French) and English-Chinese parallel news corpora, the terms of which are then augmented with WordNet synset information. This method performs well, but it would be difficult to apply to the patent domain. There exists no resource like WordNet for patents, so *meaning matching* would have to be done in some other way. Like our current system, Wang and Oard use only term translations and state that they might benefit from also using phrases.

In fact, the idea of using phrase translations in IR is not recent. Ballesteros and Croft [4] have illustrated the possible

[1]http://www.clef-campaign.org/
[2]http://research.nii.ac.jp/ntcir/

advantages of using phrase translations over term translations. Their conclusion is that, to improve retrieval, phrase translations must consistently be of high quality. Their work has also reported positive results using translation and query expansion (local feedback and local context analysis). However, their study uses much older and hence much lower baselines (from 1997); the same approach might produce different findings with current baselines.

The type of query translation we apply in this work can also be seen as a form of query expansion because the queries are expanded with their respective translations. There is extended literature on query expansion, from the early work of Rocchio [27] and Salton and Buckley [30], to more recent studies [7, 36], including approaches to expand queries using terms from dictionaries or Wikipedia entries [41]. These approaches however are largely monolingual. Efforts have been made to use such query expansion approaches across languages [4], however the well-known dataset sensitivity of query expansion often leads to instability.

Finally, several studies use word alignment algorithms from statistical MT to extract dictionaries from corpora [18, 33], or analyze multilingual collections with the goal of improving retrieval [8, 11, 16, 21, 43]. However, most of these approaches use statistical word alignment to extract a multilingual dictionary not directly from the retrieval collection, but on some external collection. Our approach differs because we extract the translation dictionary directly from the patent retrieval collection, which is something not done before to our best knowledge. We believe that this is a promising approach because dictionaries are highly domain-dependent and the better the correspondence between the dictionary's domain and the collection's domain, the more improvement in retrieval performance we would expect.

3. METHODOLOGY

In this work, we use patents granted by the EPO. When a patent is granted, the EPO provides manual translations of their claims, so that they appear in English, French and German. We use these parallel translations of the claims to extract bilingual dictionaries for each language pair. Section 3.1 describes how we extract a domain-specific patent translation dictionary from the patent claims, and section 3.2 describes how we translate queries using a translation dictionary.

3.1 Extracting a Domain-Specific Patent Translation Dictionary

In order to extract a bilingual translation dictionary from the patent claims, we need to align the parallel translations of the claims, and then estimate translation probabilities for pairs of terms in the source and target language.

Aligning the parallel translations of the patent claims is not straightforward. Patent claims are very particular in that they are usually composed of a single sentence; however this single sentence can often be 100-200 words long, with some upwards of 600 words in length. Since alignment is typically done on a sentence basis, these very long sentences create a problem. To address this, we split the sentences into smaller clauses and align these. Different heuristics may be used to automatically divide large sentences into clauses, for instance to split sentences by punctuation. However, punctuation may vary between the three languages, and so we split sentences by XML markup. Since we now have clauses instead of actual sentences, we need a sentence aligner that performs well with clauses as well as sentences as input. We use the freely-available *gargantua* sentence aligner[3], which has a reported F_1 measure of 98% in sentence alignment [5].

Source-Target Language	dict.cc	PatDict
English-French	2,950	521,387
English-German	109,961	532,042
French-English	2,913	467,176
French-German	7,338	466,435
German-English	124,596	1,461,929
German-French	8,743	1,794,897
$\sum$	256,501	5,243,866
avrg. translations per entry	2.00	9.31
pct overlapping terms	22.37%	1.09%

Table 1: Statistics of our translation dictionaries: the number of terms in each of the six source-target language pairs, the sum of those six numbers ($\sum$), the number of translations per entry (averaged over all pairs) and the percent of overlapping terms (i.e. 22.37% of terms in dict.cc are also found in PatDict).

In addition, we also conduct our own in-house manual evaluation of *gargantua*'s accuracy on the patent clauses. The manual evaluation is done by two researchers (including an expert in sentence alignment) by taking 2898 sentences from randomly chosen patents in the German-English parallel patent claims. The sentence alignment returned from *gargantua* is manually edited to create a small *gold standard* for patent clause alignment. In two different evaluations, testing *gargantua* against this gold standard has given $F_1 =$ 98% and 99% respectively.

We compute the translation probabilities between pairs of source-target language terms in the aligned patent claims using the freely-available GIZA++ toolkit [23]. For each language pair, we run GIZA++ twice, using each of the languages once as the source language. Our GIZA++ training consists of four HMM iterations, five IBM Model 1 iterations, and ends with four iterations of IBM Model 4. The output of this process is a table of translation candidate terms and their probabilities, which makes up our domain-specific translation dictionary for patents (*PatDict* henceforth). Even though patents encompass a number of sub-domains, we consider PatDict domain-specific to patents, in the sense that it covers solely the patent domain.

3.2 Translating Queries with a Bilingual Dictionary

This section describes our methodology for translating queries using a bilingual translation dictionary. Specifically we use two dictionaries: we compare the PatDict dictionary described above, to a publicly available, domain-free dictionary, *dict.cc*[4]. dict.cc is a collection of bilingual dictionaries that contains all three of the language pairs that are found in PatDict. A summary of each of the translation dictionaries is given in Table 1.

Given a query q in its original language, our aim is to expand it with translations of the original query terms. For

[3]http://sourceforge.net/projects/gargantua/
[4]http://www.dict.cc/

Query terms not covered in the translation dictionary				
Source-Target Language	dict.cc		PatDict	
	Total	Per Query	Total	Per Query
Eng-Fre	11,154	56.9	140	0.7
Eng-Ger	2317	11.8	187	1.0
Fre-Eng	715	47.7	106	7.1
Fre-Ger	626	41.7	106	7.1
Ger-Eng	3595	40.4	247	2.8
Ger-Fre	4787	53.8	186	2.1
All languages	23,194	38.7	972	1.6

Table 2: Query terms (from the whole query set used in this work, described in section 4.1.2) not covered in dict.cc and PatDict.

each term $t \in q$ we select a single translation t' from the bilingual dictionary, and we expand the original query with it. We select the single best translation t' from the dictionary, where we define as single best the translation with the highest probability. If t is not covered in the dictionary or if t' is a stopword[5], no translation takes place. We repeat this for all language combinations. At the end of this process, our new translated and expanded query q' is the union of the original query terms and their single best available translations.

We select translations according to the translation probabilities stored in the dictionary. Our own PatDict contains the translation probabilities estimated by GIZA++, but this is not the case for dict.cc. So, we augment dict.cc with translation probabilities, which we generate using word frequencies from the English, French, and German Wikipedias. We use Wikipedia because it is domain-independent, just like dict.cc.

3.3 Dictionary Coverage and Translation Selection

Both the term coverage and the translation probabilities may be different between our domain-specific PatDict and the domain-free dict.cc. By definition, PatDict has better coverage, which we expect to give more complete translations. In fact, the better coverage using PatDict is shown in Table 2 with many fewer query terms without translations. However, better coverage alone does not necessarily mean more accurate translation; working with a large number of low probability translations can lower translation accuracy (hence retrieval effectiveness) and increase computational costs [40]. This is why translation probabilities are also needed.

Our decision to select the single best translation is not the only possible option. Another option would be to set a translation probability threshold for selecting only terms that have a good enough translation probability. The threshold value could then be decided on the basis of either translation accuracy or retrieval performance. A further alternative would be to consolidate translation probabilities from various resources (e.g. to combine our two translation dictionaries by renormalising their respective translation probabilities), a method which has been shown to improve overall translation accuracy [6].

Moreover, tokenizing the queries and conducting term-based translation is not the only possible option. An interesting alternative would be to do phrase-based translation, in order to capture any non-compositional semantics in the queries that may be lost in term-based approaches. This might be of use in patent retrieval, as phrase-based approaches have shown promising results (in monolingual scenarios) [17, 24].

4. EXPERIMENTS

4.1 Experimental Settings

4.1.1 Retrieval dataset

The experiments are conducted using the CLEF-IP 2010 test collection (84GB), a subset of the Matrixware Research Collection, provided by the IRF[6]. The collection contains 2.7 million EPO patent documents from 1985-2002, covering 1.3 million separate patents in English (69%), French (7%) and German (24%). Patents are roughly comprised of textual data, bibliographic metadata, and drawings. In this paper, we ignore metadata fields like inventor, applicant, publication date, and International Patent Classification (IPC), and we use only the text fields for retrieval: title, abstract, description, and claims. Of these four text fields, we draw attention to the abstracts, which we use for queries (described below), and to the claims, which we use for creating the patent translation dictionary (described in section 3.1).

The CLEF-IP 2010 test collection also contains 300 queries (or *topics*), with their respective relevance assessments, from the prior-art CLEF-IP task. However, these 300 queries are not provided in the form of predefined keywords and/or phrases, like in other standard test collections, but instead as pointers to a patent file. Hence, an extra processing step is needed to generate queries from the patent documents (described in section 4.1.2). Overall, out of the 300 queries, 196 are English, 89 are German, and 15 are French. Table 3 displays the statistics of this collection per language. Overall we have much more data available in English than German, and even more so French. This means that when analyzing differences in retrieval performance between languages we need to look at several possible factors: both different linguistic properties and the different per-language query and document statistics could be the cause.

4.1.2 Query creation

There exist several ways for generating queries from patent documents [14, 38, 42]. In this work, we create queries from the patent documents, and specifically from their abstract, following [42], who showed that the abstract is one of the best-performing single fields from which to generate queries. Note that the experiments reported in [42] were not conducted on the same dataset as ours (European patents), but on USPTO patent data which includes fields not always found in European patents. Given the abstract of a patent, we extract queries in two different ways: (i) using the entire abstract, minus stopwords, as the query (*abstract queries* henceforth), and (ii) using the top k weighted terms from the abstract (*weighted queries* henceforth). For these

[5]We use the default stopword lists in Apache Lucene (http://lucene.apache.org) for each of the three languages.

[6]www.ir-facility.org.

experiments, we use *tf-idf* [31] to measure term weight, and we set $k = 20$. As a result, weighted queries are much shorter than abstract queries: the average length for abstract queries is 46.30 terms for English, 40.13 terms for French, and 45.08 terms for German, i.e. roughly double the size of weighted queries. This, combined with the fact that abstract query terms are not weighted (i.e. selected according to their salience), means that we expect the abstract queries to contain more noise (i.e. off-topic terms) than the weighted queries. Note that for the weighted queries, the weights are only used to filter terms and have no effect on the the ranking.

4.1.3 Plan of experiments

We use Apache Lucene[7] to index the collection without omitting stopwords or using any stemming. For retrieval, we use Lucene's standard implementation of the tf-idf retrieval model, and we perform a standard TREC[8] evaluation of the top 1000 returned documents, using the standard measures of mean average precision (MAP), precision at 10 (P10), and recall ($\frac{num\_rel\_ret}{num\_rel}$ from trec_eval).

We organize our experiments as follows.

(i) The **baseline** uses a monolingual query. E.g. an English query is used to search all patents in English, but also the portions (i.e. claims) of German or French patents that have been translated to English.

(ii) This baseline is compared against **query translation**, where we conduct two runs using separately the two different dictionaries, dict.cc and PatDict, to translate queries. We refer to these runs as QT_D, QT_P respectively.

(iii) Because our query translation is also a form of query expansion, in the sense that we expand the original queries with their translations, we also conduct a run with standard statistical **query expansion**. We use Rocchio's query expansion [27], as is implemented in Lucene[9] [29], to expand the queries with the top t most pertinent terms from the top d most relevant documents. We tune t and d as follows: t = [10,30,50,80,100] and d = [1,2,3,4,5,8,10,15,20], separately for MAP, P10, recall, and separately for abstract queries and weighted queries. The best performance is uniformly achieved with $d = 1$, but optimal t varies as follows:

- for MAP, $t = 40$ always;
- for P10, $t = 60$ for abstract queries and $t = 30$ for weighted queries;
- for recall, $t = 40$ always;

The Rocchio formula also includes two weighting parameters α and β, which we keep at default values ($\alpha = 1, \beta = 0.75$) [19]). We refer to this query expansion run as QE.

(iv) Finally, we combine Rocchio's **query expansion with query translation**. Specifically, first we expand the original (monolingual) query using Rocchio's query expansion, and then we translate all the terms in the Rocchio-expanded query using a translation dictionary. Similarly to before, we tune t and d; their optimal values are as in (iii) above. We refer to these runs as QE+QT_D, QE+QT_P respectively.

In total, we conduct six runs: *baseline,* QT_D, QT_P, QE, $QE + QT_D$, $QE + QT_P$.

[7]http://lucene.apache.org
[8]http://trec.nist.gov/
[9]http://lucene-qe.sourceforge.net/

	English	French	German
pct queries	65.3%	5.0%	29.7%
pct documents	69.0%	7.1%	24.0%
pct relevance assessments	72.2%	5.0%	21.8%

Table 3: CLEF-IP collection statistics by original language.

4.2 Results

4.2.1 Analysis by original query language

Table 4 displays the MAP, P10, and recall of our runs, grouped by the language of the original query, and separately for queries consisting of the abstract without stopwords (left hand side (LHS)) and the top weighted terms from the abstract (right hand side (RHS)).

Query translation does not consistently overperform or underperform with respect to the baseline. Comparing the domain-free (dict.cc) and domain-specific (PatDict) dictionaries used for translation, we observe that PatDict leads to higher recall but does not have consistently higher MAP or P10 scores across languages. Since prior-art search heavily relies on recall, the domain-specific dictionary might be a better choice.

Effect of noisy terms.

Comparing the LHS and RHS for Table 4, we observe that the baseline weighted queries outperform the abstract queries (e.g. 0.047 vs. 0.0384 MAP respectively). The abstract queries seem to contain more noise, which hurts overall retrieval performance. This affects query translation, as potentially noisy terms are translated and become translated noise. Often, such potentially noisy terms consist of commonly occurring terms, which are more likely to be covered in the dictionary, than other salient but more technical terms (especially for dict.cc). In this case, such terms may have higher translation probabilities simply because of their increased frequency of (co-)occurrence in the translation resources. We do not see this effect (of introducing more potentially noisy terms) with query expansion, because query expansion chooses weighted terms and effectively ignores less significant terms. In fact, we even see a drop in query expansion score from the LHS of Table 4 to its RHS in some cases, e.g., French QE MAP of 0.05644 (abstract queries) vs. 0.05637 (weighted queries). Overall we see that using query expansion with the abstract queries (LHS) improves results across languages. The addition of translations to query expansion seems to lead to only small improvements, when there is any improvement at all.

Language morphology.

Looking at retrieval performance by language, the most consistent result is that German retrieval does worse than English or French. German has the lowest baseline scores,

Results by original query language								
Query:	**Patent abstract without stopwords**				**Weighted terms from abstract**			
	German	**English**	**French**	**all**	**German**	**English**	**French**	**all**
	MAP							
baseline	0.03144	0.04101	0.04588	0.03840	**0.04081**	0.04968	0.04971	**0.04700**
QT_D	0.02902	0.04082	0.04923	0.03770*	0.03615	0.04886	0.04278	0.04480*
QT_P	0.02803	0.04283	0.04389	0.03850	0.02313	**0.05154**	0.04209	0.04260*
QE	0.03829	**0.04750**	0.05644	**0.04520***	0.03777	0.04899	0.05637	0.04600*
QE+QT_D	0.03829	0.04749	0.05644	**0.04520***	0.03777	0.04899	0.05637	0.04600*
QE+QT_P	**0.03831**	**0.04750**	**0.05645**	**0.04520***	0.03783	0.04898	**0.05642**	0.04600*
	P10							
baseline	0.04494	0.05590	**0.07333**	0.05350	**0.06067**	0.06513	**0.08000**	**0.06450**
QT_D	0.04157	0.05641	**0.07333**	0.05280	0.05506	0.06256	**0.08000**	0.06120*
QT_P	0.03708	0.05744	0.06667	0.05180	0.04157	**0.06769**	0.06000	0.05950
QE	**0.05281**	**0.06205**	**0.07333**	**0.05990***	0.05169	0.05487	0.06667	0.05450
QE+QT_D	**0.05281**	**0.06205**	**0.07333**	**0.05990***	0.05169	0.05487	0.06667	0.05450
QE+QT_P	**0.05281**	**0.06205**	**0.07333**	**0.05990***	0.05169	0.05487	0.06667	0.05450
	Recall							
baseline	0.17767	0.32093	0.27723	0.27540	0.23592	0.35588	0.27723	0.31528
QT_D	0.17476	0.31820	0.27228	0.27249*	0.21165	0.35361	0.27228	0.30626*
QT_P	0.23495	0.33137	**0.29703**	0.30044*	**0.24660**	**0.36450**	**0.31683**	**0.32635**
QE	0.23592	**0.34589**	0.28713	0.30946*	0.22913	0.32910	0.30198	0.29753*
QE+QT_D	0.23592	0.34544	0.28713	0.30917*	0.22913	0.32910	0.30198	0.29753*
QE+QT_P	**0.23883**	0.34544	0.28713	**0.31004***	0.22913	0.32910	0.30198	0.29753*

Table 4: Results by original query language separately for queries consisting of the abstract without stopwords (left hand side) and the top weighted terms from the abstract (right hand side). *baseline*: monolingual query. QT_D, QT_P: query translation with dict.cc or PatDict. QE: query expansion. Best scores marked bold. * marks statistical significance with respect to *baseline* at 5% using the Wilcoxon matched-pairs signed-ranks test (on all queries only).

Results by query difficulty										
Query:	**Patent abstract without stopwords**					**Weighted terms from abstract**				
	hard++ 23.7%	**hard** 48.5%	**easy** 25.1%	**easy++** 2.7%	**all** 100%	**hard++** 19.4%	**hard** 48.8%	**easy** 28.4%	**easy++** 3.3%	**all** 100%
	MAP									
baseline	0.00000	0.02361	0.09294	0.13616	0.03840	0.00000	0.02781	**0.09696**	0.17368	**0.04700**
QT_D	0.00000	0.02237	0.09246	0.13789	0.03770*	0.00000	0.02578	0.09222	0.17585	0.04480*
QT_P	**0.00059**	0.02611	0.08710	0.14288	0.03850	**0.00103**	0.02663	0.08358	0.16629	0.04260*
QE	0.00026	0.02696	0.10771	0.18873	**0.04520***	0.00010	0.02845	0.08932	0.19800	0.04600*
QE+QT_D	0.00026	0.02696	0.10770	0.18873	**0.04520***	0.00010	0.02845	0.08933	0.19802	0.04600*
QE+QT_P	0.00027	**0.02698**	**0.10772**	**0.18882**	**0.04520***	0.00010	**0.02847**	0.08934	**0.19805**	0.04600*
	P10									
baseline	0.00000	0.04207	0.12000	0.11250	0.05350	0.00000	0.04762	**0.13059**	0.12000	**0.06450**
QT_D	0.00000	0.04069	0.12000	0.11250	0.05280	0.00000	0.04354	0.12588	0.12000	0.06120*
QT_P	0.00000	**0.04345**	0.11067	0.11250	0.05180	**0.00172**	**0.04898**	0.10706	**0.14000**	0.05950
QE	0.00000	**0.04345**	**0.14000**	**0.13750**	**0.05990***	0.00000	0.04014	0.10588	**0.14000**	0.05450
QE+QT_D	0.00000	**0.04345**	**0.14000**	**0.13750**	**0.05990***	0.00000	0.04014	0.10588	**0.14000**	0.05450
QE+QT_P	0.00000	**0.04345**	**0.14000**	**0.13750**	**0.05990***	0.00000	0.04014	0.10588	**0.14000**	0.05450
	Recall									
baseline	0.00000	0.23252	0.58435	**1.00000**	0.27540	0.00000	0.24642	**0.61784**	**1.00000**	0.31528
QT_D	0.00000	0.22808	0.58191	**1.00000**	0.27249*	0.00000	0.23155	0.61454	**1.00000**	0.30626*
QT_P	**0.05483**	0.26249	0.57335	0.97959	0.30044*	**0.05802**	**0.26075**	0.59692	0.93103	**0.32635**
QE	0.03916	0.27969	**0.58924**	0.95918	0.30946*	0.02560	0.23951	0.56388	0.75862	0.29753*
QE+QT_D	0.03916	0.27913	**0.58924**	0.95918	0.30917*	0.02560	0.23951	0.56388	0.75862	0.29753*
QE+QT_P	0.03916	**0.28080**	**0.58924**	0.95918	**0.31004***	0.02389	0.24004	0.56388	0.75862	0.29753*

Table 5: Results by query difficulty separately for queries consisting of the abstract without stopwords (left hand side) and the top weighted terms from the abstract (right hand side). Query difficulty estimated from baseline recall rate. *very hard* and *very easy* are given as "hard++" and "easy++". The percentages on line 4 show the distribution of queries by difficulty. *baseline*: monolingual query. QT_D, QT_P: query translation with dict.cc or PatDict. QE: query expansion. Best scores marked bold. * marks statistical significance with respect to *baseline* at 5% using the Wilcoxon matched-pairs signed-ranks test (on all queries only).

but we can also see a notable drop-off when adding translation. In German, and also French, query translation from PatDict has a negative effect on precision. This contrasts with the PatDict translations from English where the opposite is true. The French and German performance is probably caused by the insufficient leverage that QT_p has available when many potential translations cannot be matched because of morphological and compounding variations. This may be aggravated by the fact that no stemming is done in our current retrieval system and that our dictionary lookup does not account for morphology either. It may be that the prevalence of specific compounds in German (a characteristic of complex texts, especially technical texts like patents) is making the translation task harder, which is a well-known problem [26, 32]. Overall, the more complex morphology of both German and French might account for some of the problems with translation, meaning that more sophisticated handling of morphology might improve translation accuracy, and hence retrieval performance.

Effect of "patentese" and language coverage.

Using queries with weighted terms (RHS), query translation is the best performing method for English (all measures) and for recall (all languages). In fact, English queries have higher recall than both French and German (Table 4). This could be due to the difference between how the abstract and claims are written. While the abstract is written for a more general audience, the claims are written in "patentese". They are very formulaic because they are legally-binding and meant to withstand scrutiny.

The large majority of patent documents and queries are in English (see Table 3). Likewise, 72.2% of the relevance assessments are in English. To find a relevant English patent, a query in German, taken from a German abstract, must rely only on the German text which is translated in the claims. As we just stated, the different language usage between abstract and claims might make retrieval difficult for the German abstract, while English queries will benefit from the English abstract, title, and description, in addition to the claims. The large percentage of relevant documents in English should make it easier for the English queries, leading to the higher recall numbers. One possible way to mitigate this effect would be to change how our query is generated, using terms from the entire document, or this could be an area where improved translation models help increase recall.

Translation selection.

Note that in these experiments we only used the single most probable translation from the dictionary. This can be a problem because many words are ambiguous, and by limiting the translations to only one, other possible correct translations will be missed. In future work we intend to test other translation methods that allow contextualisation, for example returning the top n translations, or using phrase-based translations, which has been shown to obtain better retrieval results than word by word translation[4]. We expect that phrase translations would improve the quality of translations from German in particular. For example, a compound word like *Weinflaschen* would be translated as the phrase "wine bottles" instead of just "wine".

4.2.2 Analysis by query difficulty

In order to further understand our results, we look at retrieval performance per query, and group queries based on the recall of the baseline (Table 5). For this analysis, we assume that the lower the recall of the monolingual baseline, the more difficult it will be to improve retrieval performance using either query translation or query expansion. Based on this assumption, we define four groups of *query difficulty* as shown in Table 6.

query difficulty	baseline recall
very hard (hard++) queries	0%
hard queries	1% – 49%
easy queries	50% – 99%
very easy (easy++) queries	100%

Table 6: Definition of query difficulty based on the recall of the monolingual baseline.

We observe different trends in groups of different query difficulty, which are discussed below.

Very hard queries.

For the *very hard* queries, query translation improves performance, but only when using the domain-specific dictionary. Although this improvement is modest, it does highlight the gain brought by the domain-specific dictionary. This modest gain also highlights one difference between query expansion and query translation. If the original query returns no relevant documents, query expansion cannot add meaningful terms (except by accident); translation has a better chance of improving performance in this case because it can add relevant French or German translated terms.

It could be argued that using query translations in this context provides no new information, and that translations just repeat what was in the original query. However, we expect translations to act as synonyms or like other query expansion methods, and indeed this can be seen for example with recall, where QT_P improves results on harder queries.

Generally, query expansion is more likely to perform better when given 'good' queries, where by 'good' we mean queries containing more topical and fewer noisy terms. We can see that this is also true for patents.

Very easy queries.

MAP for *very easy* queries consistently benefits from using query expansion, while P10 benefits consistently and equally from query expansion and query translation. An exception to this is recall (decrease from 1.0 to 0.95918 for abstract queries and 1.0 to 0.75862 for weighted queries), but we believe that this is partially due to estimation bias: *very easy* queries are defined as those that get 1.0 on the baseline and this number is very likely to decrease when comparing to other runs. The larger drop from 1.0 to 0.75862 which occurs with query expansion for weighted queries could be due to topical drift in the expanded queries, which can drastically reduce precision and recall. This is potentially a big advantage for query translation, as it is not affected by a similar problem.

Query translation & query expansion.

If we focus on just the hard queries (*hard* and *very hard*), we see that either QT_p or $QE{+}QT_p$ always performs best – with the exception of P10 for abstract queries where none

of the methods finds a relevant document in the top 10 for any query. Overall, either QT_p or $QE{+}QT_p$ performs best in all cases with few exceptions, which is a trend we also saw in the analysis by language (Table 4). In general, our collective results from these experiments show that query translation and query expansion can be used as complementary techniques without any detrimental effects to retrieval performance.

Finally, a note on recall. The CLEF-IP collection, and the NTCIR test collections before it, use the topic patents' citations as relevance assessments (instead of human relevance assessments). Even though the patent citations do indicate relevant documents, it may be that they do not indicate all relevant documents, or other documents, not cited in a patent, which humans would however assess as relevant. So it could be the case that the system is returning highly relevant documents which do not show up in the list of citations. With true human-generated relevance assessments, the evaluation numbers in Tables 4–5 would very likely be higher.

5. DISCUSSION

5.1 Frame of Current Work

The research presented here is part of a larger project that aims to use NLP and visualization methods in a novel way to improve IR. The general area of patent multilinguality is of particular interest in this context for two reasons:

(i) The amount of translated text available for retrieval is increasing, and so is the number of collections that contain the same documents in multiple languages, such as patent collections or the Wikipedia. For example, Wikipedia documents in different languages are not exact translations of each other, but there is significant overlap in content.

(ii) Today's typical users of IR systems, and more specifically patent retrieval systems, are very likely to be multilingual. However, their level of competence in different languages usually varies considerably. They can speak some of the languages perfectly; they may have good passive knowledge of others, but limited active competence.

In this paper, we address this multilinguality scenario for patent retrieval by computing a statistical word alignment on the retrieval collection to induce a translation dictionary. We then translate patent queries on the assumption that even though patent professionals who speak perfectly all languages involved might be better off manually translating their queries themselves, most patent professionals are not capable of doing this. Thus, we use an NLP method (statistical word alignment) on a multilingual patent collection (hence exploiting point (i) above) to help patent professionals that are partially, but not completely multilingual (e.g., they can read French, but cannot translate into French). The latter functionality addresses point (ii) above.

5.2 Frame of Future Work

Our longer term research goal is to use interactive visual interfaces that let users select a subset of the translations, potentially using a rich representation of the statistical word alignment. We believe that the complex interactions of multilinguality, alignment graphs, queries and relevant documents will require complex interfaces for optimal retrieval, especially in hard retrieval scenarios like patent retrieval, where even early studies have revealed the potential benefit brought in by improved interactive retrieval interfaces [17]. The increased complexity of user interfaces could potentially pay off in the form of enhanced query formulation [1]. Automatic query translation and query expansion are important prerequisites to help users quickly define queries covering multilingual patent documents. However, an interactive approach would provide a much higher level of control to patent specialists, who depend on continually manually fine-tuning each query. Future research will therefore focus on translation and expansion of patent queries, including interactive visual mechanisms for reformulating and changing those translations which do not meet user expectations. Through the mechanisms described in this paper, incorporation of potentially synonymous translated terms can be suggested by the system on a users' request, which should enhance the creation of high quality multilingual queries for professionals with different foreign language skills. Displaying the multilingual queries in parallel, while preserving the Boolean structure (common in patent queries) of the query as well as the translation relations between terms and phrases of the query is one of the goals we are aiming at. In order to realize this parallel display of information we are currently improving the approach for a visual interactive query tool that has been presented in [15] into this direction (see Figure 1). This enhanced query formulation tool would also support our future work which includes using multiple possible word or phrase translations instead of just the single most probable translation. However, in this paper, we only address, in a first exploratory step, automatic non-interactive query expansion by translation based on statistical word alignment.

6. CONCLUSION

In this paper we explored the multilingual aspect of patent retrieval. Starting with a collection of partially translated patents, we studied the effect of query translation on retrieval performance. Specifically, we expanded monolingual patent queries with their translations, using both a domain-specific patent dictionary that we extracted from the patent collection, and a general domain-free dictionary. Experimental evaluation on a standard CLEF-IP dataset showed that using either translation dictionary fetched similar results: query translation could help patent retrieval, but not always, and without great improvement compared to standard statistical monolingual query expansion (Rocchio). This improvement was greater when the source language was English, as opposed to French, and even more so German, a finding partly due to the effect of the complex German and French morphology upon translation accuracy, but also partly due to the prevalence of English in the collection (69% of the original language). A thorough per-query analysis revealed that cases where standard QE fails (e.g. zero recall) could benefit from query translation.

In future work, we plan to address some of the shortcomings of our current system that were discussed above. In particular, we will use phrase translations as they are less likely to introduce spurious senses in the translation. We will

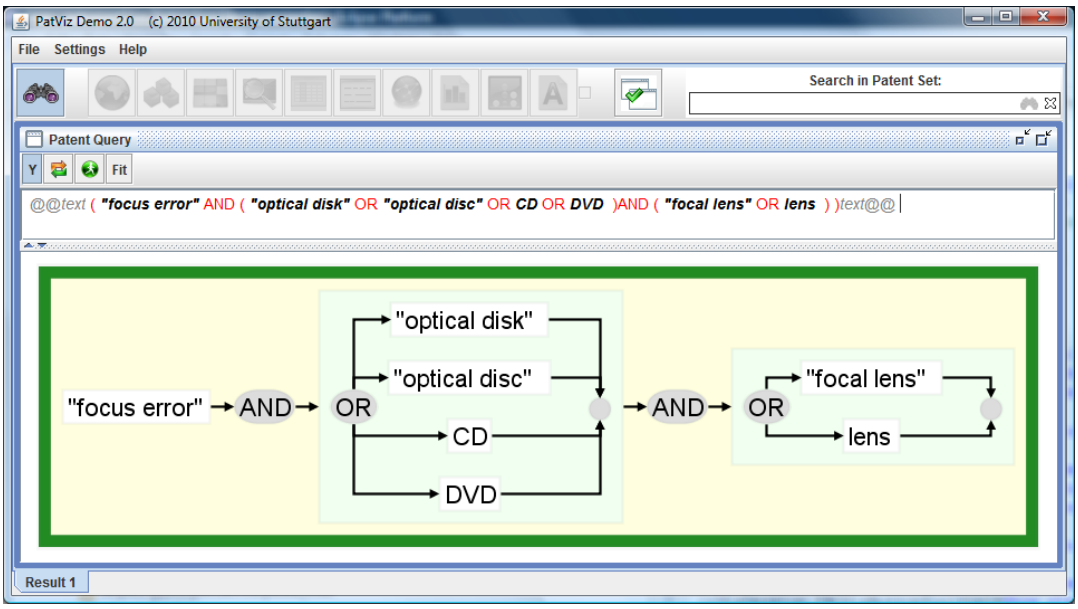

Figure 1: Query formulation tool presented in [15]

also experiment with more sophisticated linguistic analysis techniques such as decompounding for German.

Acknowledgments

The work presented here is supported by the DFG as part of the Priority Program 1335 'Scalable Visual Analytics'.

We would also like to thank Fabienne Braune and Alexander Fraser for their contribution to the paper.

7. REFERENCES

[1] W. Alink, R. Cornacchia, and A. P. de Vries. Running CLEF-IP experiments using a graphical query builder. In Peters et al. [25].

[2] K. H. Atkinson. Toward a more rational patent search paradigm. In Tait [34], pages 37–40.

[3] L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR*, pages 775–776. ACM, 2010.

[4] L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR*, pages 84–91. ACM, 1997.

[5] F. Braune and A. Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *COLING*, Beijing, China, 2010.

[6] K. Darwish and D. W. Oard. Probabilistic structured query methods. In *SIGIR*, pages 338–344. ACM, 2003.

[7] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In Efthimiadis et al. [10], pages 154–161.

[8] F. Diaz and D. Metzler. Pseudo-aligned multilingual corpora. In M. M. Veloso, editor, *IJCAI*, pages 2727–2732, 2007.

[9] T. Dunning and M. W. Davis. A single language evaluation of a multi-lingual text retrieval system. In *TREC*, pages 193–198, 1992.

[10] E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors. *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*. ACM, 2006.

[11] M. Franz, J. S. McCarley, T. Ward, and W.-J. Zhu. Unsupervised and supervised clustering for topic tracking. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 310–317. ACM, 2001.

[12] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *NTCIR*, 2008.

[13] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, K.-F. Wong, and H.-W. Hon. Exploiting query logs for cross-lingual query suggestions. *ACM Trans. Inf. Syst.*, 28(2), 2010.

[14] E. Graf, L. Azzopardi, and K. van Rijsbergen. Automatically generating queries for prior art search. In Peters et al. [25].

[15] S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *IEEE Trans. Vis. Comput. Graph.*, 2010.

[16] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *CoRR*, cs.CL/0312008, 2003.

[17] L. S. Larkey. A patent search and classification system. In *ACM DL*, pages 179–187. ACM, 1999.

[18] E. Lefever, L. Macken, and V. Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In

Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 496–504, Athens, Greece, March 2009. Association for Computational Linguistics.

[19] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008.

[20] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto. The patent mining task in the seventh ntcir workshop. In Tait [34], pages 25–32.

[21] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR*, pages 74–81. ACM, 1999.

[22] D. W. Oard, D. He, and J. Wang. User-assisted query translation for interactive cross-language information retrieval. *Inf. Process. Manage.*, 44(1):181–211, 2008.

[23] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[24] M. Osborn, T. Strzalkowski, and M. Marinescu. Evaluating document retrieval in patent database: A preliminary report. In F. Golshani and K. Makki, editors, *CIKM*, pages 216–221. ACM, 1997.

[25] C. Peters, G. M. D. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors. *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments - Tenth Workshop of the Cross-Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, Lecture Notes in Computer Science (LNCS). Springer, 2010.

[26] M. Popovic, D. Stein, and H. Ney. Statistical machine translation of German compound words. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, pages 616–624. Springer, 2006.

[27] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System–Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Inc., 1971.

[28] G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In Peters et al. [25].

[29] N. Rubens. The application of fuzzy logic to the construction of the ranking function of information retrieval systems. *Computer Modelling and New Technologies*, 10(1):20–27, 2006.

[30] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297, 1990.

[31] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[32] S. Stymne. German compounds in factored statistical machine translation. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL-08)*, Gothenburg, Sweden, 2008.

[33] L. Sun, Y. Jin, L. Du, and Y. Sun. Word alignment of english-chinese bilingual corpus based on chunks. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 110–116, Hong Kong, China, October 2000. Association for Computational Linguistics.

[34] J. Tait, editor. *Proceedings of the 1st ACM workshop on Patent Information Retrieval, PaIR 2008, Napa Valley, California, USA, October 30, 2008*. ACM, 2008.

[35] J. Tait, editor. *Proceedings of the 2nd ACM workshop on Patent Information Retrieval, PaIR 2009, Hong Kong, China, November 6, 2009*. ACM, 2009.

[36] B. Tan, A. Velivelli, H. Fang, and C. Zhai. Term feedback for information retrieval with language models. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 263–270. ACM, 2007.

[37] S. Tiwana and E. Horowitz. Findcite: automatically finding prior art patents. In *2nd international workshop on Patent IR*, pages 37–40, 2009.

[38] J. C. Toucedo and D. E. Losada. University of Santiago de Compostela at CLEF-IP09. In Peters et al. [25].

[39] C. J. K. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[40] J. Wang and D. W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In Efthimiadis et al. [10], pages 202–209.

[41] Y. Xu, G. J. F. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 59–66. ACM, 2009.

[42] X. Xue and W. B. Croft. Automatic query generation for patent search. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *CIKM*, pages 2037–2040. ACM, 2009.

[43] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: Learning from bilingual corpora. *Artif. Intell.*, 103(1-2):323–345, 1998.