Expanding queries with term and phrase translations in patent retrieval

Charles Jochim, Christina Lioma, and Hinrich Schütze

Institute for Natural Language Processing,
Computer Science, Stuttgart University
70174 Stuttgart, Germany
jochimcs@ims.uni-stuttgart.de, liomca@ims.uni-stuttgart.de

Abstract. Patent retrieval is a branch of Information Retrieval (IR) that aims to enable the challenging task of retrieving highly technical and often complicated patents. Typically, patent granting bodies translate patents into several major foreign languages, so that language boundaries do not hinder their accessibility. Given such multilingual patent collections, we posit that the patent translations can be exploited for facilitating patent retrieval.

Specifically, we focus on the translation of patent queries from German and French, the morphology of which poses an extra challenge to retrieval. We compare two translation approaches that expand the query with (i) translated terms and (ii) translated phrases. Experimental evaluation on a standard CLEF-IP European Patent Office dataset reveals a novel finding: phrase translation may be more suited to French, and term translation may be more suited to German. We trace this finding to language morphology, and we conclude that tailoring the query translation per language can lead to improved results in patent retrieval.

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

1 Introduction

Information retrieval (IR) systems used in the domain of patents need to address the difficult task of retrieving relevant yet very technical and highly specific content [20,21]. On one hand, patent content is notoriously difficult to process [1]. This difficulty is often exacerbated by the patent authors themselves who intentionally make their patents difficult to retrieve [2]. On the other hand, patent searchers require an exhaustive knowledge of all related and relevant patents, because overlooking a single valid patent potentially has detrimental and expensive consequences, e.g., infringement and litigation. In practice, this means that on one hand we have a very hard retrieval task, and on the other hand, we have demands for very high retrieval effectiveness. We tackle this difficult problem, by focusing on the multilingual aspect of patents. Since patents are partially translated into one or more languages, a collection of patents can be seen as a multilingual corpus. Given such a multilingual patent collection, we posit that one way to improve retrieval is by turning monolingual queries into multilingual queries, hence potentially improving their coverage.

We create multilingual patent queries by query translation. We present two alternatives for this approach: (i) term-by-term translations and (ii) translations that can also involve phrases. In the latter case, there are four different types of translations that can occur in the query:

- term to term translation
- term to phrase translation
- phrase to term translation
- phrase to phrase translation

For brevity, we simply refer to these two approaches as *term translation* and *phrase translation* – even though phrase translations are in reality a superset of term translations.

Our query translation is realized using a domain-specific translation dictionary of terms and phrases. We extract this dictionary from the patent collection used for retrieval, using parallel translations in the patents. Specifically, we identify such parallel translations, align them, and compute the translation probabilities between terms and phrases in the aligned translations. These translations constitute the entries in our domain-specific patent translation dictionary.

Our approach differs from previous work in that we derive a bilingual term and phrase dictionary from the retrieval collection itself – that is, we do not derive the dictionary from unrelated parallel corpora. This aspect of our work is important because it is difficult to obtain good translation coverage when using a generic dictionary or parallel text from a different corpus.

To evaluate our query translation hypothesis, we compare retrieval performance of our multilingual queries versus monolingual queries, using a competitive retrieval model. We further include runs where translation has been realised with Google's competitive MT system [12], *Google Translate*[1], so that our translation approach can be compared to a state of the art and freely available competitive approach. In addition, because our query translation can also be seen as a form of query expansion (since queries are expanded with their translations), we conduct experiments with pseudo-relevance feedback. Experimental evaluation on a standard CLEF-IP [19] dataset, focusing on the morphologically difficult cases of German and French queries, shows good results: Regarding the choice of term versus phrase translation, phrase translation seems to be more beneficial for French than for German. The improvement of our query translation approach is especially beneficial to queries of very poor baseline recall. We also find that our translation approaches are compatible with relevance feedback, and even enhance its performance (when combined with it).

The remainder of this paper is organized as follows. Section 2 overviews related work on patent IR with a focus on translation. Section 3 presents our

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

methodology for translating patent queries. Section 4 describes the experimental evaluation of our approach. Finally, section 5 summarizes this work.

2 Related Work

Thorough patent retrieval includes searching over multiple patent databases and potentially over multiple languages. Due to the multilinguality of the task, advances in cross-lingual IR (CLIR) have been explored to improve patent retrieval, for instance the NTCIR initiative on patent IR and machine translation (MT) [9], and the intellectual property (IP) track of the Cross-Language Evaluation Forum (CLEF) [19].

More generally, in the area of CLIR, translation can be broadly realized using a combination of bilingual dictionaries and/or parallel corpora and/or MT (see [17] for an overview). All three of these resources are covered in this work: we present a way of extracting a bilingual dictionary from a parallel corpus, and we also include experiments where translation is realised using Google's competitive MT system, *Google Translate*.

An early, well-cited phrase-based CLIR study was by Ballesteros and Croft [3], who expanded bilingual dictionaries with phrases and used them effectively in IR. Their definition of a phrase differs from the definition of phrase we use in this paper: They defined phrases grammatically as sequences of nouns and adjective-noun pairs, meaning that their approach required some sort of part-of-speech preprocessing. In this work, we define phrases statistically as any string of words, meaning that no grammatical preprocessing is required. Further studies followed in the 1990s and early 2000s, aiming to improve CLIR performance by first improving translation accuracy. However, more recent studies have shown that even though translation accuracy clearly affects CLIR [12], good IR performance may still be obtainable with suboptimal translation accuracy. For example, Gao et al. [10] obtain better results by using cross-lingual query suggestion than by traditional query translation. Combining different monolingual and bilingual resources, they develop a discriminative model for learning cross-lingual query similarity. With this model, they find target queries in a collection of query logs, which are similar to, but not direct translations of, the source queries. Another example to depart from exact query translation, is the work of Wang and Oard [22], who tackle translation for IR as a case of *meaning matching*. Specifically, they align candidate term translations from parallel corpora, which they then augment with WordNet synset (i.e. meaning matching) information. Motivated by these more recent advances into CLIR, in this work we also adopt a translation approach that does not aim to translate the query as accurately as possible, but rather to "gist" the query and translate its most salient parts – to our knowledge, this is novel for patent IR.

Our approach of translating queries consists in expanding them with their respective translations, hence it may be seen as a form of query expansion. Query expansion in general has been shown to be effective in CLIR. For instance, Chinnakotla et. al. [7] studied Multilingual Pseudo-Relevance Feedback (*MultiPRF*):

They first translated the query into a target language and then ran retrieval with both source and target language queries to obtain feedback models for both languages. The target language feedback model was then translated back to the source language with a bilingual dictionary, and the resulting model was combined with the original query model and the source feedback model to rank results. Their MultiPRF method was found to be beneficial to IR, and could even improve monolingual retrieval results. The potential usefulness of PRF specifically for patent IR has also been studied [4,5], however solely for monolingual patents.

3 Methodology

The aim of our approach is to turn monolingual queries into multilingual for patent IR. To this end, we extract a translation dictionary of terms and phrases from a parallel patent corpus (Section 3.1). This patent corpus is also the retrieval collection used in this work (see Section 4.1 for its description). We use the extracted dictionary to translate the original monolingual queries (Section 3.2). Section 3.3 describes how we integrate this translation process into retrieval.

3.1 Extracting a Translation Dictionary of Terms and Phrases

Our retrieval collection contains European Patent Office (EPO) patents, which comprise text fields, e.g., *title*, *abstract*, *description*, and *claims*; metadata fields, e.g., *applicant*, *inventor*, *International Patent Classification (IPC)*, and *date published*; and figures and illustrations. The claims field is very important for patent IR, because it contains the legally-binding portion of the patent that may be used later to determine the patent's validity or defend it against infringement [1,2]. In our collection, the claims are manually translated into English, French, and German. Therefore, the claims of our patent collection may be seen as a parallel corpus which can be used to extract translation dictionaries specific to the patent domain.

First, we extract the claims field from documents in our collection that contain claims in all three languages. Then, we align these translated claims. We use a word alignment tool for building our dictionary. However, the sentences in the claims field are often very long [1], and this may cause a problem to the word alignment tool because generally these tools do not handle very long sentences very well; therefore it is necessary to limit the length of sentences. We divide claims into shorter subsentences by splitting sentences using the XML markup found in the patents. This approach is chosen over splitting by non-terminating punctuation (i.e. colon, comma, semicolon, etc.) because punctuation use seems to vary more between languages than the XML markup.

As a result, even though initially the patent claims were perfectly aligned, the subsentences that we get after splitting are not necessarily perfectly aligned, due mainly to variation in the XML markup. To correct this, we align subsentences using the sentence aligner gargantua[2]. We choose gargantua because of its high accuracy in sentence alignment (F_1 scores of 98.47 for aligning German-English sentences and 98.60 for French-English [6]).

Having aligned subsentences, we can now start training a statistical machine translation system which will produce a translation dictionary. In this work, we use the state of the art Moses statistical machine translation system [13]. First, we run the GIZA++ word aligner [18] inputting the subsentences we previously aligned with gargantua. GIZA++ produces a word to word alignment, which can also be used as a translation dictionary for terms [11,15,22]. To get the phrase translations, we use the default GDFA (*grow diagonally final AND*) alignment [14] from GIZA++ to train Moses and produce a bidirectional phrase table. The phrase table includes phrase translations with a source phrase of length m and a target phrase of length n, where m and n are between 1-7 inclusive. This means that the phrase table contains term to term translations, term to phrase translations, phrase to term translations, and phrase to phrase translations. The phrase table is also bidirectional, so the size of the German-English dictionary will be the same as the size of the English-German dictionary. The translation probabilities are not symmetric though; e.g., the German word *gefäß* translates to *vessel* with probability 0.61, but *vessel* translates to *gefäß* with probability 0.26.

Languages	# entries	*term* → *term*	*term* → *phr.*	*phr.* → *term*	*phr.* → *phr.*
French-English	162,840,175	922,952	1,715,491	3,437,236	156,764,496
French-German	157,977,915	1,645,245	3,570,673	10,419,547	142,342,450
German-English	116,611,676	1,290,111	4,642,276	2,863,824	107,815,465

Table 1. Patent dictionary: Note that the dictionaries are bidirectional. So there are 1,715,491 French terms that can be translated to an English phrase and 1,715,491 English phrases that can be translated to a French term.

Table 1 shows the total number of entries for each language pair as well as a breakdown of the total according to type of translation equivalence. Generally, there are many more equivalences involving phrases simply because there are many more phrases than terms in any given language. The highest number of equivalences between a term and a phrase (in either direction) occur for German-English (4,642,276) and French-German (10,419,547) because compounds (which are terms) are frequent in German and they are most often translated as phrases.

3.2 Translating Queries

Given the term and phrase translation dictionary described above, we use two different methods to translate queries; the first is term to term translation (TR_{term});

[2] http://sourceforge.net/projects/gargantua/

the second includes phrase translations (TR_{phrase}). In the TR_{term} method, for every term t in a query q, we identify the single best translation t' of the term, and extend q with t'. We define the single best translation to be the most probable one according to the bilingual dictionary extracted by Moses. We do this for all possible source-language combinations, and end up with a multilingual query. An example of term translation can be seen in Table 2, where the first row is the original German query, and the second and third rows are the French and English term translations. This example shows some of the typical problems that occur in automatic term-by-term translations: some terms are poorly translated (German "aufzeichnungsmaterial" 'recording material' is translated as "support"by the German-French dictionary) and some terms can only be adequately translated as a phrase (German "aufzeichnungsmaterial" can only be adequately translated as a phrase, "recording material", in English).

original German query	ein tintenstrahl aufzeichnungsmaterial mit einem träger und mindestens einer unteren pigment
French term translation	*un jet support avec un support et moins une inférieure pigment*
English term translation	*a inkjet recording with a carrier and least a lower pigment*

Table 2. Example of term and phrase translation from German to French and English.

In the TR_{phrase} method, we extract only those phrases in the original query for which we have a translation. Our definition of phrase is the longest *n-gram* (i.e. string of n words) in the dictionary for n=[1-7]. Terms that are not present in any phrase are translated as terms in the multilingual query. Since stopwords will be removed by the retrieval system we do not need to translate phrases of only stopwords. So although "of a" would be considered a phrase in our approach, we do not translate it. Taking this one step further, we remove any stopwords at the beginning or end of a phrase, but preserve the stopwords within phrases. So "of the ink jet" simply becomes "ink jet" while "coated with aluminum" remains the same.

3.3 Translating Salient Terms

In the previous section, we described our query translation method. We do not apply this to all query terms, but to a selection of the most salient terms in the query. The motivation for this is two-fold: on one hand, previous work on query translation for patent IR has showed the limitation of fully translating whole patent queries ([11]); on the other hand, recent work on CLIR has showed that satisfactory retrieval performance can be achieved with approximate translations of queries (e.g. [10]).

We select the terms that are to be translated as follows. Using our baseline retrieval model (see Section 4.1), we look at the term weights assigned to the individual query terms. Figure 1 displays this distribution over our whole query-set (described in Section 4.1). By defining a threshold θ of term weights, we can assume that we reasonably separate the most salient query terms from the rest. Hence, we translate only the terms whose weight $> \theta$. The higher the threshold, the fewer terms are translated. For example, for $\theta = 0.02$, 9.1% of query terms are selected for translation.

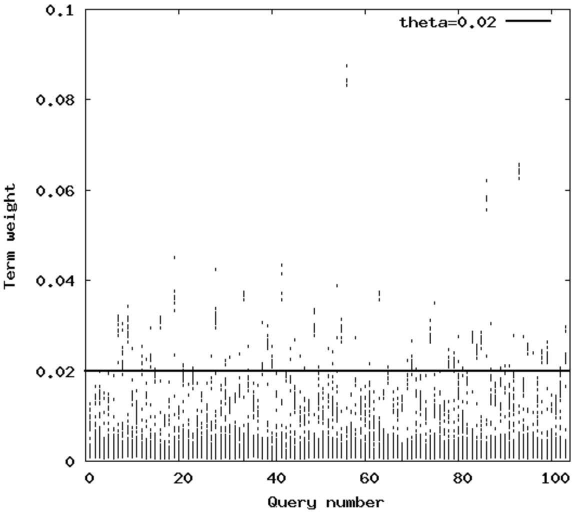

Fig. 1. Term weights of the query set in the baseline run

4 Experiments

4.1 Experimental Settings

For our experiments we focus on the translation of patent queries from German to English & French and from French to English & German, i.e. we transform an originally monolingual query into a trilingual query. We focus only on French and German as source languages because they have a more complex morphology than English, and hence pose a bigger problem for CLIR than translation from English. We use the patent collection from CLEF-IP 2010 [19] (see Table 3), which has 104 topics in German and French with relevance assessments. These topics are not TREC-style queries, but full patent documents, hence an open problem is how to generate queries out of them [5,23]. We do this by taking the abstract field of the topic patent as the query, following [23]. Our initial query is the set of unique terms from the abstract (average query length= 57.6 terms).

Size	84 GB		French	German
# documents	2,680,604	pct queries	5.0%	29.7%
# tokens	9,840,411,560	pct documents	7.1%	24.0%
# unique terms	20,132,873	pct relevance assessments	5.0%	21.8%

Table 3. CLEF-IP 2010 collection: total size (left) and % by original language (right).

We index the collection using Indri[3] without removing stopwords or stemming. For retrieval we use the Kullback-Leibler language model with Dirichlet smoothing [8]. We tune Dirichlet's μ parameter within $\mu = \{5000, 7500, 10000, 12500, 15000, 17500, 20000\}$. For retrieval, we use standard stoplists for German and French[4]. Our translation approach includes a term weight threshold θ (described in Section 3.3), which we tune: $\theta = \{0.016, 0.02, 0.025\}$. Because our query translation approach expands queries with their (partial) translations, we also do runs with pseudo-relevance feedback (PRF), using Indri's default PRF implementation, which is an adaptation of Lavrenko's relevance model [16]. PRF uses these parameters: the number of documents ($fbDocs$) and the number of terms ($fbTerms$). We set $fbDocs = 1$ and $fbTerms = 40$, following [11]. Finally, we include a run that uses Google Translate for query translation. This type of translation differs from ours (we submit the whole query for translation to Google Translate, whereas our approach translates only salient terms/phrases; also, Google Translate is domain-free, whereas our approach uses a patent translation dictionary extracted from the retrieval collection). We include the Google Translate runs simply to contextualize the results from our approach. We use standard TREC evaluation measures: mean average precision (MAP), precision at 10 (P10), and recall. We tune separately for each evaluation measure.

Our experiments are set up as follows:

1. **baseline:** original monolingual query;
2. $\boldsymbol{TR}_{\text{term}}$**:** the original query is expanded with term translations of its most salient terms;
3. $\boldsymbol{TR}_{\text{phrase}}$**:** the original query is expanded with phrase translations of its most salient terms;
4. **PRF:** same as baseline but with PRF;
5. $\boldsymbol{TR}_{\text{term}}$**+PRF:** same as $\boldsymbol{TR}_{\text{term}}$ but with PRF;
6. $\boldsymbol{TR}_{\text{phrase}}$**+PRF:** same as TR_{phrase} but with PRF;
7. **Google translate:** the original query is expanded with its full translation using Google translate.

[3] http://www.lemurproject.org

[4] accessible from http://members.unine.ch/jacques.savoy/clef/

4.2 Experimental Results

Table 4 summarizes our experimental results. For German, term translation seems better than phrase translation. We originally expected that term-to-phrase translations would handle German compounds better than term-to-term translation. Indeed, in query 237, *korngrößenverteilung* translates to *particle size distribution* with phrase translation, and only to *size* using term translation. For this query, term translation improves MAP by 4% over the baseline, and phrase translation improves MAP by 36% over the baseline. However, there are also cases where German is equally well translated with phrases or terms, e.g., in query 201, *tintenstrahl* is translated as *inkjet* or *ink jet*, respectively. For French, phrase translation is clearly better than term translation: MAP and P10 improve substantially over the baseline, compared to term translation. Recall however decreases. For the PRF runs, MAP and P10 are about the same for term and phrase translation, but recall of phrase translation shows a clear improvement compared to term translation. A reason for French benefiting more from phrase translation than German may be that it does not have as many compounds: concepts that are expressed as phrases in French are often translated as compounds in German. E.g., the phrase “flux de matière” ‘flux of material’ in query 242 gets translated into *materialströme* in phrase translation, and into *material*, *von*, and *strom* in term translation.

On several occasions PRF outperforms the baseline. This is not unexpected, as PRF has been shown to help patent retrieval [4,5]. However, the best result overall for each combination of language and evaluation measure fuses PRF with translation (although in the case of P10 for German, PRF with and without translation are tied). This shows that PRF and translation can contribute different improvements to retrieval performance, and that these two very different approaches are not incompatible. Furthermore, looking at the Google Translate run, it is not surprising that it does best on recall, but underperforms in the other measures: the queries translated by Google contain the full patent abstract and its full translation, meaning that they are very lengthy queries with probably better coverage at the expense of precision.

	MAP		**P10**		**Recall**	
	German	**French**	**German**	**French**	**German**	**French**
baseline	0.0581	0.0527	0.0864	0.0667	0.2456	0.2772
TR_{term}	**0.0598**	0.0556	**0.0875**	0.0867	0.2622	**0.2871**
TR_{phrase}	0.0577	**0.0614**	**0.0875**	***0.1000**	**0.2671**	0.2772
PRF	0.0664	0.0730	**0.0875**	**0.1067**	0.2661	0.2822
TR_{term}+PRF	0.0667	0.0719	0.0841	0.1000	**0.2749**	0.2871
TR_{phrase}+PRF	**0.0672**	**0.0744**	0.0864	0.1000	0.2739	**0.3218**
Google translate	0.0473	0.0652	0.0659	*0.1200	0.3168	0.3614

Table 4. Best scores in bold. * marks statistical significance $p<0.05$ using the two-tailed t-test. TR_{term} is term translation and TR_{phrase} is phrase translation

4.2.1 Analysis by query difficulty In order to further understand our findings we look more closely at performance on a per-query basis. Specifically, we group queries on the basis of their baseline recall, on the assumption that queries of very low baseline recall will be much more difficult to improve (using either PRF or translation), than queries with higher baseline recall.

Table 5 presents retrieval performance split between three groups of query difficulty: *very hard* (baseline recall = 0%), *hard* (baseline recall = 1%–49%), and *medium* (baseline recall = 50%–100%). We see that German queries of *medium* difficulty underperform with term or phrase translations. On the other hand, results for these runs improve for the *hard*, and *very hard* queries. It seems few queries account for most of this variation. For example, in the group of *medium* queries, one query (query 213) has better recall with term and phrase translation than the baseline, while for two queries (queries 75, 152) the baseline has better recall. The recall for query 152 drops substantially with 8 of 10 relevant documents being retrieved in the baseline and none being retrieved with either word or phrase translation. This single query accounts for most of the decrease in translation results in the *medium* group. Additionally, all of the relevance assessments for this query are German (hence the settings for this query can be seen as biased). A single query can also account for much of the improvement in the *hard* group's translation results. For query 201 for example, 2 of the 20 relevant patent documents are in German and the baseline query only returns those 2 relevant documents. Adding phrase translations (in particular the addition of the phrase "ink jet") increases the number of relevant documents returned to 15. We also observe that *medium* difficulty queries tend to have more relevance judgements in their original source language, and that *hard* queries tend to have relevance judgements from different languages. To the extent that this is the case, it is understandable that results for *medium* queries worsen with translation: a largely monolingual (say, French) result set has high ranks for relevant documents (which are all French), but for a multilingual result set the ranking of some French relevant documents slips. On the other hand, *hard* queries may improve if the baseline monolingual result set (French) does not match many relevant documents (several English documents with a few French), but with the addition of multilingual documents to the result set, more relevant documents are retrieved. This bias of the percentage of a query's relevant documents that are in the original source language of the query also affects the performance of PRF. PRF chooses terms for expansion from the top ranked documents. These documents are likely to be in the same language as the original query. So an original French query in the baseline will expand the query with French terms. In the cases of TR_{term}+PRF and TR_{phrase}+PRF, the highest ranked result with *TR* is often a multilingual patent document and so the multilingual query will have multilingual expansions. With this multilingual patent collection, multilingual query expansion should be more desirable, and in fact *TR*+PRF outperforms PRF for MAP and recall. In particular, *TR*+PRF does better than PRF for *hard* and *very hard* queries where it appears there is a larger percentage of relevant documents in a language different than the query.

	German				**French**			
	hard++	**hard**	**medium**	**all**	**hard++**	**hard**	**medium**	**all**
	(31.8%)	(48.9%)	(19.3%)	(100%)	(20.0%)	(46.7%)	(33.3%)	(100%)
	MAP							
baseline	0.0000	0.0375	**0.2058**	0.0581	0.0000	0.0676	0.0635	0.0527
TR_{term}	0.0002	**0.0417**	0.2036	**0.0598**	0.0000	0.0694	0.0695	0.0556
TR_{phrase}	**0.0003**	0.0406	0.1953	0.0577	0.0000	**0.0794**	**0.0729**	**0.0614**
PRF baseline	0.0013	0.0525	**0.2091**	0.0664	0.0000	**0.0907**	0.0919	0.0730
TR_{term}+PRF	**0.0018**	0.0543	0.2049	0.0667	0.0000	0.0894	0.0905	0.0719
TR_{phrase}+PRF	0.0014	**0.0558**	0.2045	**0.0672**	0.0000	0.0852	**0.1038**	**0.0744**
	P10							
baseline	0.0000	0.0721	**0.2647**	0.0864	0.0000	**0.0857**	0.0800	0.0667
TR_{term}	0.0000	0.0767	0.2588	**0.0875**	0.0000	**0.1286**	0.0800	0.0867
TR_{phrase}	0.0000	**0.0837**	0.2412	**0.0875**	0.0000	**0.1286**	**0.1200**	***0.1000**
PRF baseline	0.0000	0.0860	**0.2353**	**0.0875**	0.0000	**0.1143**	**0.1600**	**0.1067**
TR_{term}+PRF	0.0000	0.0837	0.2235	0.0841	0.0000	**0.1143**	0.1400	0.1000
TR_{phrase}+PRF	0.0000	**0.0907**	0.2176	0.0864	0.0000	0.1000	**0.1600**	0.1000
	Recall							
baseline	0.0000	0.2532	**0.6328**	0.2456	0.0000	0.2301	**0.5882**	0.2772
TR_{term}	**0.0405**	0.2731	0.5989	0.2622	0.0000	**0.2566**	0.5686	**0.2871**
TR_{phrase}	0.0372	**0.2821**	0.6045	**0.2671**	**0.0263**	0.2389	0.5490	0.2772
PRF baseline	0.0372	0.2712	**0.6328**	0.2661	0.0263	0.2301	0.5882	0.2822
TR_{term}+PRF	**0.0642**	**0.2857**	0.5932	**0.2749**	0.0263	0.2301	**0.6078**	0.2871
TR_{phrase}+PRF	0.0608	0.2821	0.6045	0.2739	**0.0526**	**0.2920**	0.5882	**0.3218**

Table 5. German and French results by difficulty. * marks statistical significance $p<0.05$ using the two-tailed t-test.

In contrast to the German results, the French MAP and P10 scores improve for term and phrase translation across the *hard* and *medium* groups, and results for *very hard* remain the same. Query 239 is an example of a query which improves for MAP and P10 while recall remains the same. The original French query has recall of 1, returning all four of its relevant documents (two of them include translated claims). Phrase translation still proves to be useful here in improving the relevant documents' ranking (MAP and P10 both improve). For recall, we see the same behavior as with German: recall drops using translations (TR_{term} and TR_{phrase}) in the *medium* group and rises for harder queries. For the *medium* group, only one less document is retrieved using TR_{term} with respect to the baseline. This single document accounts for the 3.6% drop in recall. Note that there are only 15 French queries and an average of 13.5 relevance assessments per query, so one relevant document being added to or dropped from the result set can have a big impact. Table 6 shows that for the majority of queries, recall remains the same.

Furthermore, Table 6 shows how TR_{term} and TR_{phrase} performed against the baseline for individual queries. We counted, for each evaluation measure, the number of queries that performed better than, worse than, or equal to the baseline. All measures for both translation methods, with the exception of recall for TR_{term}, have more queries that exceed the baseline than queries that drop below it. We observe that, for each language/evaluation measure combination, there are more queries improved by phrase translation than queries improved by term translation (with one tie, 2-2, for French P10). However, in three cases (German MAP, German recall, French recall), there are also more queries where phrase translation does worse than term translation; the other three cases (German P10, French MAP, French P10) are tied. Our interpretation of these results is that phrases have significant potential for improving retrieval results, but they have to be carefully selected, otherwise performance will deteriorate. In contrast, term translations are more conservative and less likely to have a negative effect, but at the same time they offer limited improvements.

Eval. meas.	$TR_{\text{term}} >$ *baseline*	$TR_{\text{term}} <$ *baseline*	$TR_{\text{term}} =$ *baseline*	$TR_{\text{phrase}} >$ *baseline*	$TR_{\text{phrase}} <$ *baseline*	$TR_{\text{phrase}} =$ *baseline*
German						
MAP	19	13	56	33	23	32
P10	5	4	79	7	4	77
recall	16	9	63	18	12	58
French						
MAP	5	3	7	9	3	3
P10	2	0	13	2	0	13
recall	2	2	11	4	3	8

Table 6. German and French performance per query. $TR_{\text{term}} >$ *baseline* indicates that the evaluation measure was greater for term translation than the baseline. $<$ and $=$ indicate less than and equal to, respectively. Other notation as in Table 4

4.2.2 Tuning of Dirichlet parameter μ Finally, Figures 2-3 show MAP and P10 scores across the tuning range of μ. The more stable the line of our approach, the less sensitive it is to factors pertaining to variation in document length and collection statistics. For the MAP tuning for both German and French, the results for term and phrase translation are quite stable, while the three runs that use PRF drop (between 10000 and 12500 for German and between 12500 and 15000 for French).

Note that results are less stable for the P10 tuning, although word and phrase translations appear more stable than PRF runs. The German PRF results drop for P10 like they did for MAP.

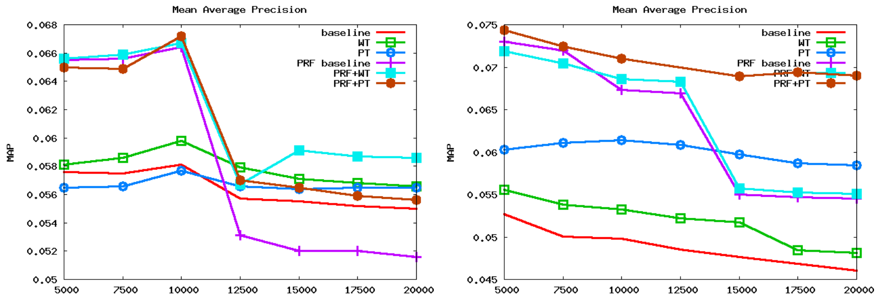

Fig. 2. Dirichlet prior μ tuning for German (left) and French (right) versus MAP (y axis).

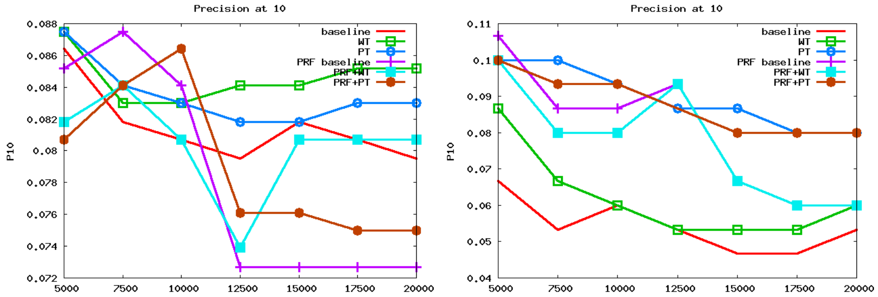

Fig. 3. Dirichlet prior μ tuning for German (left) and French (right) versus P10 (y axis).

5 Conclusions

In an increasingly networked world, problems of multilinguality are gaining importance in information retrieval (IR). Most IR approaches in multilingual settings use some form of translation. In this paper, we adopted an expansion approach to translation for patent IR, where translations of query parts are added as additional terms to the query. We looked at two alternative translation methods, term translation and phrase translation. Our experimental evaluation showed good results for both, especially on hard queries. Phrase translation seems to be more beneficial for French than for German because German often uses single-term compounds instead of phrases, thus limiting the potential benefit of phrase to term and phrase to phrase translations.

References

1. K. H. Atkinson. Toward a more rational patent search paradigm. In *1st ACM workshop on Patent IR*, pages 37–40, 2008.
2. L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In *SIGIR*, pages 775–776, 2010.
3. L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR*, pages 84–91, 1997.
4. S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *CIKM*, pages 1863–1866, 2009.
5. S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, pages 457–470, 2010.
6. F. Braune and A. Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *COLING*, 2010.
7. M. K. Chinnakotla, K. Raman, and P. Bhattacharyya. Multilingual prf: english lends a helping hand. In *SIGIR*, pages 659–666, 2010.
8. W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, 2003.
9. A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *NTCIR*, 2008.
10. W. Gao, C. Niu, J.-Y. Nie, M. Zhou, K.-F. Wong, and H.-W. Hon. Exploiting query logs for cross-lingual query suggestions. *TOIS*, 28(2), 2010.
11. C. Jochim, C. Lioma, H. Schütze, S. Koch, and T. Ertl. Preliminary study into query translation for patent retrieval. In *PaIR*. ACM, 2010.
12. K. Kettunen. Choosing the best mt programs for clir purposes - can mt metrics be helpful? In *ECIR*, pages 706–712, 2009.
13. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *ACL*, pages 177–180, 2007.
14. P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *NAACL*, pages 48–54, 2003.
15. L. S. Larkey and M. E. Connell. Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Inf. Process. Manage.*, 41(3):457–473, 2005.

16. V. Lavrenko and W. B. Croft. Relevance-based language models. In *SIGIR*, pages 120–127, 2001.
17. D. W. Oard and A. R. Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology*, 33:223–256, 1998.
18. F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
19. G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In *CLEF*. Springer, 2010.
20. J. Tait, editor. *1st ACM workshop on Patent IR.* 2008.
21. J. Tait, editor. *2nd ACM workshop on Patent IR.* 2009.
22. J. Wang and D. W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In *SIGIR*, pages 202–209, 2006.
23. X. Xue and W. B. Croft. Automatic query generation for patent search. In *CIKM*, pages 2037–2040, 2009.