Adaptation of Statistical Machine Translation Model for Cross-Lingual Information Retrieval in a Service Context

Nikoulina, V.; Kovachev, B.; Lagos, N.; Monz, C.

Publication date
2012

Document Version
Final published version

Published in
EACL 2012: 13th Conference of the European Chapter of the Association for Computational Linguistics

License
CC BY

Citation for published version (APA):
Abstract

This work proposes to adapt an existing general SMT model for the task of translating queries that are subsequently going to be used to retrieve information from a target language collection. In the scenario that we focus on access to the document collection itself is not available and changes to the IR model are not possible. We propose two ways to achieve the adaptation effect and both of them are aimed at tuning parameter weights on a set of parallel queries. The first approach is via a standard tuning procedure optimizing for BLEU score and the second one is via a reranking approach optimizing for MAP score. We also extend the second approach by using syntax-based features. Our experiments show improvements of 1-2.5 in terms of MAP score over the retrieval with the non-adapted translation. We show that these improvements are due both to the integration of the adaptation and syntax-features for the query translation task.

1 Introduction

Cross Lingual Information Retrieval (CLIR) is an important feature for any digital content provider in today's multilingual environment. However, many of the content providers are not willing to change existing well-established document indexing and search tools, nor to provide access to their document collection by a third-party external service. The work presented in this paper assumes such a context of use, where a query translation service allows translating queries posed to the search engine of a content provider into several target languages, without requiring changes to the underlying IR system used and without accessing, at translation time, the content provider’s document set. Keeping in mind these constraints, we present two approaches on query translation optimisation.

One of the important observations done during the CLEF 2009 campaign (Ferro and Peters, 2009) related to CLIR was that the usage of Statistical Machine Translation (SMT) systems (eg. Google Translate) for query translation led to important improvements in the cross-lingual retrieval performance (the best CLIR performance increased from ~55% of the monolingual baseline in 2008 to more than 90% in 2009 for French and German target languages). However, general-purpose SMT systems are not necessarily adapted for query translation. That is because SMT systems trained on a corpus of standard parallel phrases take into account the phrase structure implicitly. The structure of queries is very different from the standard phrase structure: queries are very short and the word order might be different than the typical full phrase one. This problem can be seen as a problem of genre adaptation for SMT, where the genre is “query”.

To our knowledge, no suitable corpora of parallel queries is available to train an adapted SMT system. Small corpora of parallel queries1 however can be obtained (eg. CLEF tracks) or manually created. We suggest to use such corpora in order to adapt the SMT model parameters for query translation. In our approach the parameters of the SMT models are optimized on the basis of the parallel queries set. This is achieved either directly in the SMT system using the MERT (Minimum Error Rate Training) algorithm and optimiz-

---

1Insufficient for a full SMT system training (~500 entries)
ing according to the BLEU\(^2\) (Papineni et al., 2001) score, or via reranking the Nbest translation candidates generated by a baseline system based on new parameters (and possibly new features) that aim to optimize a retrieval metric.

It is important to note that both of the proposed approaches allow keeping the MT system independent of the document collection and indexing, and thus suitable for a query translation service. These two approaches can also be combined by using the model produced with the first approach as a baseline that produces the Nbest list of translations that is then given to the reranking approach.

The remainder of this paper is organized as follows. We first present related work addressing the problem of query translation. We then describe two approaches towards adapting an SMT system to the query-genre: tuning the SMT system on a parallel set of queries (Section 3.1) and adapting machine translation via the reranking framework (Section 3.2). We then present our experimental settings and results (Section 4) and conclude in section 5.

2 Related work

We may distinguish two main groups of approaches to CLIR: document translation and query translation. We concentrate on the second group which is more relevant to our settings. The standard query translation methods use different translation resources such as bilingual dictionaries, parallel corpora and/or machine translation. The aspect of disambiguation is important for the first two techniques.

Different methods were proposed to deal with disambiguation issues, often relying on the document collection or embedding the translation step directly into the retrieval model (Hiemstra and Jong, 1999; Berger et al., 1999; Kraaij et al., 2003). Other methods rely on external resources like query logs (Gao et al., 2010), Wikipedia (Jadidinejad and Mahmoudi, 2009) or the web (Nie and Chen, 2002; Hu et al., 2008). (Gao et al., 2006) proposes syntax-based translation models to deal with the disambiguation issues (NP-based, dependency-based). The candidate translations proposed by these models are then reranked with the model learned to minimize the translation error on the training data.

To our knowledge, existing work that use MT-based techniques for query translation use an out-of-the-box MT system, without adapting it for query translation in particular (Jones et al., 1999; Wu et al., 2008) (although some query expansion techniques might be applied to the produced translation afterwards (Wu and He, 2010)).

There is a number of works done for domain adaptation in Statistical Machine Translation. However, we want to distinguish between genre and domain adaptation in this work. Generally, genre can be seen as a sub-problem of domain. Thus, we consider genre to be the general style of the text e.g. conversation, news, blog, query (responsible mostly for the text structure) while the domain reflects more what the text is about – eg. social science, healthcare, history, so domain adaptation involves lexical disambiguation and extra lexical coverage problems. To our knowledge, there is not much work addressing explicitly the problem of genre adaptation for SMT. Some work done on domain adaptation could be applied to genre adaptation, such as incorporating available in-domain corpora in the SMT model: either monolingual (Bertoldi and Federico, 2009; Wu et al., 2008; Zhao et al., 2004; Koehn and Schroeder, 2007), or small parallel data used for tuning the SMT parameters (Zheng et al., 2010; Pecina et al., 2011).

3 Our approach

This work is based on the hypothesis that the general-purpose SMT system needs to be adapted for query translation. Although in (Ferro and Peters, 2009) it has been mentioned that using Google translate (general-purpose MT) for query translation allowed to CLEF participants to obtain the best CLIR performance, there is still 10% gap between monolingual and cross-lingual IR. We believe that, as in (Clinchant and Renders, 2007), more adapted query translation, possibly further combined with query expansion techniques, can lead to improved retrieval.

The problem of the SMT adaptation for query-genre translation has different quality aspects. On the one hand, we want our model to produce a “good” translation (well-formed and transmitting the information contained in the source query) of an input query. On the other hand, we want to obtain good retrieval performance using
the proposed translation. These two aspects are not necessarily correlated: a bag-of-word translation can lead to good retrieval performance, even though it won’t be syntactically well-formed; at the same time a well-formed translation can lead to worse retrieval if the wrong lexical choice is done. Moreover, often the retrieval demands some linguistic preprocessing (e.g., lemmatisation, PoS tagging) which in interaction with badly-formed translations might bring some noise.

A couple of works studied the correlation between the standard MT evaluation metrics and the retrieval precision. Thus, (Fujii et al., 2009) showed a good correlation of the BLEU scores with the MAP scores for Cross-Lingual Patent Retrieval. However, the topics in patent search (long and well structured) are very different from standard queries. (Kettunen, 2009) also found a pretty high correlation ($0.8 \sim 0.9$) between standard MT evaluation metrics (METEOR (Banerjee and Lavie, 2005), BLEU, NIST (Doddington, 2002)) and retrieval precision for long queries. However, the same work shows that the correlation decreases ($0.6 \sim 0.7$) for short queries.

In this paper we propose two approaches to SMT adaptation for queries. The first one optimizes BLEU, while the second one optimizes Mean Average Precision (MAP), a standard metric in information retrieval. We’ll address the issue of the correlation between BLEU and MAP in Section 4.

Both of the proposed approaches rely on the phrase-based SMT (PBMT) model (Koehn et al., 2003) implemented in the Open Source SMT toolkit MOSES (Koehn et al., 2007).

### 3.1 Tuning for genre adaptation

First, we propose to adapt the PBMT model by tuning the model’s weights on a parallel set of queries. This approach addresses the first aspect of the problem, which is producing a “good” translation. The PBMT model combines different types of features via a log-linear model. The standard features include (Koehn, 2010, Chapter 5): language model, word penalty, distortion, different translation models, etc. The weights of these features are learned during the tuning step with the MERT (Och, 2003) algorithm. Roughly the MERT algorithm tunes feature weights one by one and optimizes them according to the BLEU score obtained.

Our hypothesis is that the impact of different features should be different depending on whether we translate a full sentence, or a query-genre entry. Thus, one would expect that in the case of query-genre the language model or the distortion features should get less importance than in the case of the full-sentence translation. MERT tuning on a genre-adapted parallel corpus should leverage this information from the data, adapting the SMT model to the query-genre. We would also like to note that the tuning approach (proposed for domain adaptation by (Zheng et al., 2010)) seems to be more appropriate for genre adaptation than for domain adaptation where the problem of lexical ambiguity is encoded in the translation model and re-weighting the main features might not be sufficient.

We use the MERT implementation provided with the Moses toolkit with default settings. Our assumption is that this procedure although not explicitly aimed at improving retrieval performance will nevertheless lead to “better” query translations when compared to the baseline. The results of this approach allow us also to observe whether and to what extent changes in BLEU scores are correlated to changes in MAP scores.

### 3.2 Reranking framework for query translation

The second approach addresses the retrieval quality problem. An SMT system is usually trained to optimize the quality of the translation (e.g., BLEU score for SMT), which is not necessarily correlated with the retrieval quality (especially for the short queries). Thus, for example, the word order which is crucial for translation quality (and is taken into account by most MT evaluation metrics) is often ignored by IR models. Our second approach follows (Nie, 2010, pp.106) argument that “the translation problem is an integral part of the whole CLIR problem, and unified CLIR models integrating translation should be defined”. We propose integrating the IR metric (MAP) into the translation model optimisation step via the reranking framework.

Previous attempts to apply the reranking approach to SMT did not show significant improvements in terms of MT evaluation metrics (Och et al., 2003; Nikoulina and Dymetman, 2008). One of the reasons being the poor diversity of the Nbest list of the translations. However, we be-
lieve that this approach has more potential in the context of query translation.

First of all the average query length is $\sim 5$ words, which means that the Nbest list of the translations is more diverse than in the case of general phrase translation (average length 25-30 words).

Moreover, the retrieval precision is more naturally integrated into the reranking framework than standard MT evaluation metrics such as BLEU. The main reason is that the notion of Average Retrieval Precision is well defined for a single query translation, while BLEU is defined on the corpus level and correlates poorly with human quality judgements for the individual translations (Specia et al., 2009; Callison-Burch et al., 2009).

Finally, the reranking framework allows a lot of flexibility. Thus, it allows enriching the baseline translation model with new complex features which might be difficult to introduce into the translation model directly.

Other works applied the reranking framework to different NLP tasks such as Named Entities Extraction (Collins, 2001), parsing (Collins and Roark, 2004), and language modelling (Roark et al., 2004). Most of these works used the reranking framework to combine generative and discriminative methods when both approaches aim at solving the same problem: the generative model produces a set of hypotheses, and the best hypothesis is chosen afterwards via the discriminative reranking model, which allows to enrich the baseline model with the new complex and heterogeneous features. We suggest using the reranking framework to combine two different tasks: Machine Translation and Cross-lingual Information Retrieval. In this context the reranking framework doesn’t only allow enriching the baseline translation model but also performing training using a more appropriate evaluation metric.

### 3.2.1 Reranking training

Generally, the reranking framework can be resumed in the following steps:

1. The baseline (generic-purpose) MT system generates a list of candidate translations $GEN(q)$ for each query $q$;
2. A vector of features $F(t)$ is assigned to each translation $t \in GEN(q)$;
3. The best translation $\hat{t}$ is chosen as the one maximizing the translation score, which is defined as a weighted linear combination of features: 
   \[
   \hat{t}(\lambda) = \arg \max_{t \in GEN(q)} \lambda \cdot F(t)
   \]

As shown above the best translation is selected according to features’ weights $\lambda$. In order to learn the weights $\lambda$ maximizing the retrieval performance, an appropriate annotated training set has to be created. We use the CLEF tracks to create the training set. The retrieval scores annotations are based on the document relevance annotations performed by human annotators during the CLEF campaign.

The annotated training set is created out of queries $\{q_1, \ldots, q_K\}$ with an Nbest list of translations $GEN(q_i)$ of each query $q_i$, $i \in \{1..K\}$ as follows:

- A list of $N$ (we take $N = 1000$) translations $GEN(q_i)$ is produced by the baseline MT model for each query $q_i$, $i = 1..K$.
- Each translation $t \in GEN(q_i)$ is used to perform a retrieval from a target document collection, and an Average Precision score ($AP(t)$) is computed for each $t \in GEN(q_i)$ by comparing its retrieval to the relevance annotations done during the CLEF campaign.

The weights $\lambda$ are learned with the objective of maximizing MAP for all the queries of the training set, and, therefore, are optimized for retrieval quality.

The weights optimization is done with the Margin Infused Relaxed Algorithm (MIRA) (Crammer and Singer, 2003), which was applied to SMT by (Watanabe et al., 2007; Chiang et al., 2008). MIRA is an online learning algorithm where each weights update is done to keep the new weights as close as possible to the old weights (first term), and score oracle translation (the translation giving the best retrieval score: $t^*_i = \arg \max_t AP(t)$) higher than each non-oracle translation ($t_{ij}$) by a margin at least as wide as the loss $l_{ij}$ (second term):

\[
\lambda = \min_{\lambda'} \frac{1}{2} \|\lambda' - \lambda\|^2 + C \sum_{i=1}^{K} \max_{j=1..N} \left( l_{ij} - \lambda' \cdot (F(t^*_i) - F(t_{ij})) \right)
\]

The loss $l_{ij}$ is defined as the difference in the retrieval average precision between the oracle and non-oracle translations: $l_{ij} = AP(t^*_i) - AP(t_{ij})$. $C$ is the regularization parameter which is chosen via 5-fold cross-validation.
3.2.2 Features

One of the advantages of the reranking framework is that new complex features can be easily integrated. We suggest to enrich the reranking model with different syntax-based features, such as:

- features relying on dependency structures: called therein coupling features (proposed by Nikoulina and Dymetman, 2008);
- features relying on Part of Speech Tagging: called therein PoS mapping features.

By integrating the syntax-based features we have a double goal: showing the potential of the reranking framework with more complex features, and examining whether the integration of syntactic information could be useful for query translation.

Coupling features. The goal of the coupling features is to measure the similarity between source and target dependency structures. The initial hypothesis is that a better translation should have a dependency structure closer to the one of the source query.

In this work we experiment with two different coupling variants proposed in (Nikoulina and Dymetman, 2008), namely, Lexicalised and Label-specific coupling features.

The generic coupling features are based on the notion of “rectangles” that are of the following type: ((s1, d, s2), (t1, d, t2)), where d is an edge between source words s1 and s2, d is an edge between target words t1 and t2. Lexicalised features take into account the quality of lexical alignment, by weighting each rectangle (s1, s2, t1, t2) by a probability of aligning s1 to t1 and s2 to t2 (eg. p(s1|t1)p(s2|t2) or p(t1|s1)p(t2|s2)).

The Label-Specific features take into account the nature of the aligned dependencies. Thus, the rectangles of the form ((s1, subj, s2), (t1, subj, t2)) will get more weight than a rectangle ((s1, subj, s2), (t1, nmod, t2)). The importance of each “rectangle” is learned on the parallel annotated corpus by introducing a collection of Label-Specific coupling features, each for a specific pair of source label and target label.

PoS mapping features. The goal of the PoS mapping features is to control the correspondence of Part Of Speech Tags between an input query and its translation. As the coupling features, the PoS mapping features rely on the word alignments between the source sentence and its translation\(^3\). A vector of sparse features is introduced where each component corresponds to a pair of PoS tags aligned in the training data. We introduce a generic PoS map variant, which counts a number of occurrences of a specific pair of PoS tags, and lexical PoS map variant, which weights down these pairs by a lexical alignment score (p(s|t) or p(t|s)).

4 Experiments

4.1 Experimental basis

4.1.1 Data

To simulate parallel query data we used translation equivalent CLEF topics. The data set used for the first approach consists of the CLEF topic data from the following years and tasks: AdHoc-main track from 2000 to 2008; CLEF AdHoc-TEL track 2008; Domain Specific tracks from 2000 to 2008; CLEF robust tracks 2007 and 2008; GeoCLEF tracks 2005-2007. To avoid the issue of overlapping topics we removed duplicates. The created parallel queries set contained 500 – 700 parallel entries (depending on the language pair, Table 1) and was used for Moses parameters tuning.

In order to create the training set for the reranking approach, we need to have access to the relevance judgements. We didn’t have access to all relevance judgements of the previously described tracks. Thus we used only a subset of the previously extracted parallel set, which includes CLEF 2000-2008 topics from the AdHoc-main, AdHoc-TEL and GeoCLEF tracks.

The number of queries obtained altogether is shown in (Table 1).

4.1.2 Baseline

We tested our approaches on the CLEF AdHoc-TEL 2009 task (50 topics). This task dealt with monolingual and cross-lingual search in a library catalog. The monolingual retrieval is

\(^3\)This alignment can be either produced by a toolkit like GIZA++(Och and Ney, 2003) or obtained directly by a system that produced the Nbest list of the translations (Moses).
performed with the lemur⁴ toolkit (Ogilvie and Callan, 2001). The preprocessing includes lemmatisation (with the Xerox Incremental Parser-XIP (Aït-Mokhtar et al., 2002)) and filtering out the function words (based on XIP PoS tagging).

Table 2 shows the performance of the monolingual retrieval model for each collection. The monolingual retrieval results are comparable to the CLEF AdHoc-TEL 2009 participants (Ferro and Peters, 2009). Let us note here that it is not the case for our CLIR results since we didn’t exploit the fact that each of the collections could actually contain the entries in a language other than the official language of the collection.

The cross-lingual retrieval is performed as follows:

- the input query (eg. in English) is first translated into the language of the collection (eg. German);
- this translation is used to search the target collection (eg. Austrian National Library for German).

The baseline translation is produced with Moses trained on Europarl. Table 2 reports the baseline performance both in terms of MT evaluation metrics (BLEU) and Information Retrieval evaluation metric MAP (Mean Average Precision).

The 1best MAP score corresponds to the case when the single translation is proposed for the retrieval by the query translation model. 5best MAP score corresponds to the case when the 5 top translations proposed by the translation service are concatenated and used for the retrieval.

The 5best retrieval can be seen as a sort of query expansion, without accessing the document collection or any external resources.

Given that the query length is shorter than for a standard sentence, the 4-gramm BLEU (used for standart MT evaluation) might not be able to capture the difference between the translations (eg. English-German 4-gramm BLEU is equal to 0 for our task). For that reason we report both 3- and 4-gramm BLEU scores.

Note, that the French-English baseline retrieval quality is much better than the German-English. This is probably due to the fact that our German-English translation system doesn’t use any decoumpounding, which results into many non-translated words.

4.2 Results

We performed the query-genre adaptation experiments for English-French, French-English, German-English and English-German language pairs.

Ideally, we would have liked to combine the two approaches we proposed: use the query-genre-tuned model to produce the Nbest list which is then reranked to optimize the MAP score. However, it was not possible in our experimental settings due to the small amount of training data available. We thus simply compare these two approaches to a baseline approach and comment on their respective performance.

4.2.1 Query-genre tuning approach

For the CLEF-tuning experiments we used the same translation model and language model as for the baseline (Europarl-based). The weights were then tuned on the CLEF topics described in section 4.1.1. We then tested the system obtained on 50 parallel queries from the CLEF AdHoc-TEL 2009 task.

Table 3 describes the results of the evaluation. We observe consistent 1-best MAP improvements, but unstable BLEU (3-gramm) (improvements for English-German, and degradation for other language pairs), although one would have expected BLEU to be improved in this experimental setting given that BLEU was the objective function for MERT. These results, on one side, confirm the remark of (Kettunen, 2009) that there is a correlation (although low) between BLEU and MAP scores. The unstable BLEU scores

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Number of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total queries</td>
<td></td>
</tr>
<tr>
<td>En - Fr, Fr - En</td>
<td>470</td>
</tr>
<tr>
<td>En - De, De - En</td>
<td>714</td>
</tr>
<tr>
<td>Annotated queries</td>
<td></td>
</tr>
<tr>
<td>En - Fr, Fr - En</td>
<td>400</td>
</tr>
<tr>
<td>En - De, De - En</td>
<td>350</td>
</tr>
</tbody>
</table>

Table 1: Top: total number of parallel queries gathered from all the CLEF tasks (size of the tuning set). Bottom: number of queries extracted from the tasks for which the human relevance judgements were available (size of the reranking training set).
Table 2: Baseline MAP scores for monolingual and bilingual CLEF AdHoc TEL 2009 task.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>MAP 1-best</th>
<th>MAP 5-best</th>
<th>BLEU 4-gramm</th>
<th>BLEU 3-gramm</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.3159</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French-English</td>
<td>0.1828</td>
<td>0.2186</td>
<td>0.1199</td>
<td>0.1568</td>
</tr>
<tr>
<td>German-English</td>
<td>0.0941</td>
<td>0.0942</td>
<td>0.2351</td>
<td>0.2923</td>
</tr>
</tbody>
</table>

Table 3: BLEU and MAP performance on CLEF AdHoc TEL 2009 task for the genre-tuned model.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>MAP 1-best</th>
<th>MAP 5-best</th>
<th>BLEU 4-gramm</th>
<th>BLEU 3-gramm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr-En</td>
<td>0.1954</td>
<td>0.2229</td>
<td>0.1062</td>
<td>0.1489</td>
</tr>
<tr>
<td>De-En</td>
<td>0.1018</td>
<td>0.1078</td>
<td>0.2240</td>
<td>0.2486</td>
</tr>
<tr>
<td>En-Fr</td>
<td>0.1611</td>
<td>0.1516</td>
<td>0.2072</td>
<td>0.2908</td>
</tr>
<tr>
<td>En-De</td>
<td>0.1062</td>
<td>0.1132</td>
<td>0.0000</td>
<td>0.1924</td>
</tr>
</tbody>
</table>

might also be explained by the small size of the test set (compared to a standard test set of 1000 full-sentences).

Secondly, we looked at the weights of the features both in the baseline model (Europarl-tuned) and in the adapted model (CLEF-tuned), shown in Table 4. We are unsure how suitable the sizes of the CLEF tuning sets are, especially for the pairs involving English and French. Nevertheless, we do observe and comment on some patterns.

For the pairs involving English and German, the distortion weight is much higher when tuning with CLEF data compared to tuning with Europarl data. The picture is reversed when looking at the two pairs involving English and French. This is to be expected if we interpret a high distortion weight as follows: “it is not encouraged to place source words that are near to each other far away from each other in the translation”. Indeed, the local reorderings are much more frequent between English and French (e.g., white house = maison blanche), while the long-distance reorderings are more typical between English and German.

The word penalty is consistently higher over all pairs when tuning with CLEF data compared to tuning with Europarl data. We could see an explanation for this pattern in the smaller size of the CLEF sentences if we interpret higher word penalty as a preference for shorter translations. This can be explained both with the smaller average size of the queries and with the specific query structure: mostly content words and fewer function words when compared to the full sentence.

The language model weight is consistently though not drastically smaller when tuning with CLEF data. We suppose that this is due to the fact that a Europarl-base language model is not the best choice for translating query data.

### 4.2.2 Reranking approach

The reranking experiments include different features combinations. First, we experiment with the Moses features only in order to make this approach comparable with the first one. Secondly, we compare different syntax-based features combinations, as described in section 3.2.2. Thus, we compare the following reranking models (defined by the feature set): moses, lex (lexical coupling + moses features), lab (label-specific coupling + moses features), posmaplex (lexical PoS mapping + moses features), lab-lex (label-specific coupling + lexical coupling + moses features), lab-lex-posmap (label-specific coupling + lexical coupling features + generic PoS mapping). To reduce the size of feature-functions vectors we take only the 20 most frequent features in the training data for Label-specific coupling and PoS mapping features. The computation of the syntax features is based on the rule-based XIP parser, where some heuristics specific to query processing have been integrated into English and French (but not German) grammars (Brun et al., 2012).

The results of these experiments are illustrated...
in Figure 1. To keep the figure more readable, we report only on 3-gram BLEU scores. When computing the 5-best MAP score, the order in the Nbest list is defined by a corresponding reranking model. Each reranking model is illustrated by a single horizontal red bar. We compare the reranking results to the baseline model (vertical line) and also to the results of the first approach (yellow bar labelled MERT:moses) on the same figure.

First, we remark that the adapted models (query-genre tuning and reranking) outperform the baseline in terms of MAP (1best and 5 best) for French-English and German-English translations for most of the models. The only exception is posmaplex model (based on PoS tagging) for German which can be explained by the fact that the German grammar used for query processing was not adapted for queries as opposed to English and French grammars. However, we do not observe the same tendency for BLEU score, where only a few of the adapted models outperform the baseline, which confirms the hypothesis of the low correlation between BLEU and MAP scores in these settings. Table 5 gives some examples of the queries translations before (T1) and after (T2) reranking. These examples also illustrate different types of disagreement between MAP and 1-gramm BLEU score.

The results for English-German and English-French look more confusing. This can be partly due to the more rich morphology of the target languages which may create more noise in the syntax structure. Reranking however improves over the 1-best MAP baseline for English-German, and 5-best MAP is also improved excluding the models involving PoS tagging for German (posmap, posmaplex, lab-lex-posmap). The results for English-French are more difficult to interpret. To find out the reason of such a behavior, we looked at the translations. We observed the following tokenization problem for French: the apostrophe is systematically separated, e.g. "d ’ aujourd ’ hui". This leads to both noisy pre-retrieval preprocessing (eg. d is tagged as a NOUN) and noisy syntax-based feature values, which might explain the unstable results.

Finally, we can see that the syntax-based features can be beneficial for the final retrieval quality: the models with syntax features can outperform the model basd on the moses features only. The syntax-based features leading to the most sta-

---

Table 4: Feature weights for the query-genre tuned model. Abbreviations: DW - distortion weight, LM - language model weight, PP - phrase penalty, WP - word penalty, \( \phi(f|e) \) - phrase translation probability, \( \text{lex}(f|e) \) - lexical weighting.

| Lng pair | Tune set  | DW       | LM       | \( \phi(f|e) \)  | \( \text{lex}(f|e) \)  | \( \phi(e|f) \)  | \( \text{lex}(e|f) \)  | PP      | WP      |
|----------|-----------|----------|----------|----------------|-----------------|----------------|-----------------|--------|--------|
| Fr-En    | Europarl  | 0.0801   | 0.1397   | 0.0625        | 0.0431          | 0.1463         | 0.0638          | -0.0670 | -0.3975 |
|          | CLEF      | 0.0015   | 0.0795   | -0.0046       | 0.0348          | 0.1977         | 0.0208          | -0.2904 | 0.3707  |
| De-En    | Europarl  | 0.0588   | 0.1341   | 0.0181        | 0.0380          | 0.1382         | 0.0398          | -0.0904 | -0.4822 |
|          | CLEF      | 0.3568   | 0.1151   | 0.0350        | 0.0549          | 0.0932         | 0.0805          | 0.0391  | -0.1434 |
| En-Fr    | Europarl  | 0.0789   | 0.1373   | 0.0002        | 0.0766          | 0.1798         | 0.0293          | -0.0978 | -0.4002 |
|          | CLEF      | 0.0322   | 0.1251   | 0.1023        | 0.0534          | 0.0365         | 0.0365          | -0.3182 | -0.2972 |
| En-De    | Europarl  | 0.0584   | 0.1396   | 0.0092        | 0.0821          | 0.1823         | 0.0437          | -0.1613 | -0.3233 |
|          | CLEF      | 0.3451   | 0.1001   | 0.0248        | 0.0872          | 0.2629         | 0.0153          | -0.0431 | 0.1214  |

Table 5: Some examples of queries translations (T1: baseline, T2: after reranking with lab-lex), MAP and 1-gramm BLEU scores for German-English.

<table>
<thead>
<tr>
<th>Query Example</th>
<th>MAP</th>
<th>bleu1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weibliche Märtyrer</td>
<td>0.07</td>
<td>1</td>
</tr>
<tr>
<td>T1</td>
<td>female martyrs</td>
<td>0.4</td>
</tr>
<tr>
<td>T2</td>
<td>Women martyrs</td>
<td></td>
</tr>
<tr>
<td>Src2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genmanipulation am Menschen</td>
<td>0.044</td>
<td>0.167</td>
</tr>
<tr>
<td>T1</td>
<td>On the genetic manipulation of people</td>
<td>0.069</td>
</tr>
<tr>
<td>T2</td>
<td>genetic manipulation of the human being</td>
<td></td>
</tr>
<tr>
<td>Src3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arbeitsrecht in der Europäischen Union</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Labour law in the European Union</td>
<td>0.015</td>
</tr>
<tr>
<td>T2</td>
<td>labour legislation in the European Union</td>
<td>0.036</td>
</tr>
</tbody>
</table>

The higher order BLEU scores are equal to 0 for most of the individual translations.
ble results seem to be lab-lex (combination of lexical and label-specific coupling): it leads to the best gains over 1-best and 5-best MAP for all language pairs excluding English-French. This is a surprising result given the fact that the underlying IR model doesn’t take syntax into account in any way. In our opinion, this is probably due to the interaction between the pre-retrieval preprocessing (lemmatisation, PoS tagging) done with the linguistic tools which might produce noisy results when applied to the SMT outputs. The reranking with syntax-based features allows to choose a better-formed query for which the PoS tagging and lemmatisation tools produce less noise which leads to a better retrieval.

5 Conclusion

In this work we proposed two methods for query-genre adaptation of an SMT model: the first method addressing the translation quality aspect and the second one the retrieval precision aspect. We have shown that CLIR performance in terms of MAP is improved between 1-2.5 points. We believe that the combination of these two methods would be the most beneficial setting, although we were not able to prove this experimentally (due to the lack of training data). None of these methods require access to the document collection at test time, and can be used in the context of a query translation service. The combination of our adapted SMT model with other state-of-the-art CLIR techniques (eg. query expansion with PRF) will be explored in future work.

Acknowledgements

This research was supported by the European Union’s ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430 (Project GALATEAS).

References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: in-


Paul Ogilvie and James P. Callan. 2001. Experiments using the lemur toolkit. In TREC.


Dan Wu and Daqing He. 2010. A study of query translation using google machine translation system. Computational Intelligence and Software Engineering (CISE).

