Natural Language Engineering **1** (1): 1–27. Printed in the United Kingdom

Leveraging Bilingual Terminology to Improve Machine Translation in a CAT Environment †

MIHAEL ARCAN[1], MARCO TURCHI[2], SARA TONELLI[2], and PAUL BUITELAAR[1]

[1] Insight Centre for Data Analytics, National University of Ireland, Galway
{firstname.lastname}@insight-centre.org

[2] FBK- Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
{turchi,satonelli}@fbk.eu

(*Received 1 December 2016*)

Abstract

This work focuses on the extraction and integration of automatically aligned bilingual terminology into a Statistical Machine Translation (SMT) system in a Computer Aided Translation (CAT) scenario. We evaluate the proposed framework that, taking as input a small set of parallel documents, gathers domain-specific bilingual terms and injects them into an SMT system to enhance translation quality. Therefore, we investigate several strategies to extract and align terminology across languages and to integrate it in an SMT system. We compare two terminology injection methods that can be easily used at run-time without altering the normal activity of an SMT system: XML markup and cache-based model. We test the cache-based model on two different domains (information technology and medical) in English, Italian and German, showing significant improvements ranging from 2.23 to 6.78 BLEU points over a baseline SMT system and from 0.05 to 3.03 compared to the widely-used XML markup approach.

1 Introduction

In a typical CAT scenario, professional translators carry out domain-specific projects and work on assigned documents with the help of software modules, which suggest translations by looking at past translated sentences (*i.e.* translation memories). Such tools include modules for terminology management and support collaborative work by several translators on different partitions of the same project. Overall, the translation is performed manually, while the CAT modules support translators in taking informed and correct translation decisions.

Several attempts have been made to automate parts of this process (Heyn, 1996;

† This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

Federico, Cattelan, and Trombetti, 2012; Läubli, Fishel, Massey, Ehrensberger-Dow, and Volk, 2013; Green, Heer, and Manning, 2013), in particular, to reduce human intervention in terms of time and effort without affecting translation quality. Recently, a solution implemented within the MATECAT project[1] (Federico, Bertoldi, Cettolo, Negri, Turchi, Trombetti, Cattelan, Farina, Lupinetti, Martines, Massidda, Schwenk, Barrault, Blain, Koehn, Buck, and Germann, 2014) integrates an SMT system in a CAT scenario, so that the SMT system not only translates but also learns to adapt to translator's preferences. This approach proved to yield significant efficiency gains over an unadapted approach (Bentivogli, Bertoldi, Cettolo, Federico, Negri, and Turchi, 2016). Although promising, this novel configuration leaves an important question open on how to automatically leverage and integrate domain-specific terminology in this scenario. Since high-quality terminology is crucial to produce an accurate and coherent translation, in this work, we analyse the impact of bilingual terminology extraction and integration in a CAT scenario including an SMT system. In particular, the impact is investigated at different stages of the translation process, trying to address the following questions:

1. What strategy should be chosen to extract monolingual terminology in this translation scenario?
2. What are the best ways to align bilingual terminology from a set of monolingual terms?
3. What are the best strategies to inject bilingual terms into an SMT system in a CAT environment?

All experiments in this work are carried out taking into account the typical working environment of professional translators (Figure 1): When a customer gives a translation company a large translation project, a document or a set of documents related to the same topic are split into partitions according to the daily workload of different translators. Each partition is first automatically translated with the SMT system and then post-edited by a professional translator with a CAT tool. Our approach takes advantage of such post-edited data, where the source and the post-edited target documents are used to automatically extract bilingual terminology (Section 3). The aligned terminology is then injected into the SMT system to improve its performance when applied to the other partitions of the same document (Sections 4.1 and 4.2).

Such environment has clearly some constraints that affect the experimental setup presented in this work. First, supervised approaches for terminology extraction should be avoided, because translators who start working on a new project in a new domain often have to produce new terminology from scratch, and enrich it incrementally. We show that approaches based on word alignment and term translation applied to the daily extracted data are more robust and more efficient than the state-of-the-art method based on pre-trained classifiers (Aker, Paramita, and Gaizauskas, 2013). Second, we cannot choose an SMT solution that requires the translation service to be regularly stopped to re-train the model, as shown in Bouamor, Semmar,

[1] http://www.matecat.com/

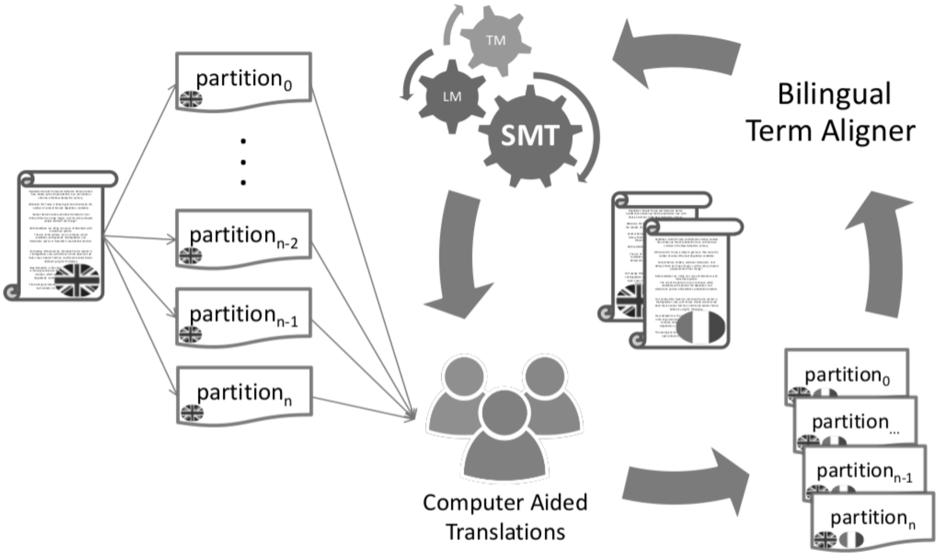

Fig. 1: System architecture of bilingual terminology injection into an SMT system.

and Zweigenbaum (2012) or Pinnis and Skadins (2012). Instead, we investigate the integration of cache-based translation and language models (Bertoldi, Cettolo, and Federico, 2013) in the context of terminology integration. This approach makes it possible to periodically add bilingual terms to an SMT system in real-time, without the need to stop it, as required in the considered scenario.

The evaluation of our framework on two different domains (IT and medical) and two language pairs (English-Italian and English-German) shows significant improvements in terms of BLEU score over a generic SMT system as well as over an integration method based on XML markup, suggesting that the proposed strategy is portable across different domains and language pairs.

This paper is an extension of our previous work (Arcan, Turchi, Tonelli, and Buitelaar, 2014b) on extracting bilingual terminology from a post-edited corpus to enhance the performance of an SMT system in a CAT environment. In addition to the experiments based on the English → Italian translation direction, we extend this work by evaluating results for Italian → English. In order to address issues related to richer morphology and noun compounds, we also examine the performance of our framework on the more challenging German-English language pair. All experiments in term extraction, alignment and translation are supported by a manual evaluation. Additionally, we test the robustness of the bilingual term integration strategies when artificially misaligned terminology is injected into the SMT system.

The remainder of the paper is organised as follows: in Section 2 we give an overview of past works related to bilingual terminology extraction from parallel texts and to the integration of domain-specific terms in SMT systems. In Section 3 we first describe monolingual terminology extraction, followed by the bilingual alignment. In Section 4 we present the integration of bilingual terms into SMT through the XML markup approach and cache-based models. Section 5 introduces the overall framework of our experiment as well as the different datasets used in the experiment. In Section 6 we present our results, and in Section 7 the impact of misaligned terminology on translation quality is evaluated. We finally summarise our findings and give an outlook on future research in Section 8.

2 Related Work

Our work is based on a framework that includes monolingual extraction of domain-specific terms from a small parallel corpus, followed by bilingual term alignment, and the integration of the identified bilingual terminology into an SMT system. In past years, a number of techniques have been applied to the task of bilingual terminology extraction from parallel or comparable corpora. Most of the works rely on the identification of monolingual candidates using linguistic knowledge, statistical methods, or a combination of the two. Kim, Baldwin, and Kan (2009) propose an unsupervised method to extract monolingual domain-specific terms from a document collection using term frequency and inverse document frequency (*tf-idf*) (Salton, Wong, and Yang, 1975; Sparck Jones, 1972). Although their method does not extract a large number of domain-specific terms, the quality of terms is generally high and well distributed over all domains. Daille, Gaussier, and Langé (1994) make use of linguistic knowledge to identify certain noun phrases which are likely to be domain-specific terms. They compare statistical scores, such as frequency count, association criteria or bilingual count, to discriminate domain-specific terms among the candidates across languages. Similarly, Wu and Chang (2003) propose an algorithm that uses syntactic and statistical analysis to extract bilingual collocations from a Chinese-English parallel corpus. Phrases matching the syntactic patterns in a sentence-aligned corpus are identified via cross-linguistic statistical association.

Due to the small amount of sentences stored in each examined partition, the use of pure statistical methods is not suitable for our scenario. For this reason, we rely on three monolingual term extraction tools, which use linguistic annotations (POS) in combination with statistical methods (*tf-idf*). The tools considered are the KX toolkit (Pianta and Tonelli, 2010), TWSC (Pinnis, Ljubešić, Ştefănescu, Skadiņa, Tadić, and Gornostay, 2012) and AlchemyAPI,[2] which support term extraction on all targeted languages, i.e. English, Italian and German.

For bilingual term alignment, Aker et al. (2013) cast this task as a classification problem and use the EuroVoc thesaurus as training data. Their work mainly focuses on the quality of the extracted alignments, where the performance often reaches 100% precision. The alignment algorithm proposed by Bouamor et al. (2012) is based on a vector space model. The entries in the vectors are co-occurrence statistics between the terms computed over the entire corpus. Furthermore, their term integration methods focus on concatenating the newly obtained bilingual data to

[2] `http://www.alchemyapi.com/products/features/keyword-extraction/`

the existing corpus or adding entries directly into the phrase table. The necessity of dealing with several domains implies the need to keep a large static translation model separate from specific parallel data, such as a bilingual terminology. Both mentioned methods for term alignment rely on models, which have to be trained in advance. Thurmair and Aleksić (2012) extract terms and lexicon entries from SMT phrase tables. In their approach, they apply linguistic, lexical and frequency filters to obtain good lexicon entries. Vintar and Fišer (2008) present an approach to extend the automatically created Slovene WordNet with nominal multi-word expressions. First, they translate the multi-word expressions from Princeton WordNet into Slovene based on the word alignment models and lexico-syntactic patterns. Then, new terms for the Slovene WordNet are extracted from a monolingual corpus using 'keywordness' ranking and contextual patterns. Arcan, Giuliano, Turchi, and Buitelaar (2014a) address the problem of automatically identifying and disambiguating terms in a document and propose an approach to translate them using cross-lingual links in Wikipedia. All these works rely on the presence of external resources (i.e. annotated data to train a classifier, a phrase table already containing domain-specific terms or a translation system able to correctly translate specific terms) that are not available in our scenario. Our approach only takes advantage of the small quantity of parallel data provided by each translator.

As for the integration of domain-specific parallel data such as dictionaries or bilingual terminology into an SMT system, three main strands of research have been explored in the past: incorporating existing terminology within word alignment training (Okita and Way, 2010), retraining additional in-domain parallel resources (Arcan, Federmann, and Buitelaar, 2012; Haddow and Koehn, 2012) or adding new entries to the phrase table (Ren, Lü, Cao, Liu, and Huang, 2009). These approaches all allow the integration of domain-specific terms, but they require either switching-off the SMT system, which is unsuitable for our scenario or accessing prior knowledge to translate specific expressions. Several approaches (Xiong, Meng, and Liu, 2016; Weller, Fraser, and Heid, 2014; Pinnis, 2015) investigate the use of XML markup (see Section 4.1 for more details) to inject bilingual terms for SMT at decoding phase. Being simple to apply and not requiring the interruption of the normal activity of an SMT system, the XML markup is the most widely-used technique, but it has shown some limitations in efficiently considering the surrounding information of the span to translate.

As a post-processing step, Itagaki and Aikawa (2008) propose a way to identify terminology translations from SMT output and automatically swap them with user-defined translations. Although showing large improvements, this work requires manual linguistic templates (e.g. case markers, predicates) for each language pair, which makes it unsuitable in our CAT scenario.

3 Bilingual Domain-Specific Terminology Generation

In the light of the framework presented in Figure 1, we propose a two-step approach to extract bilingual terminology, requiring only small amounts of parallel data (few hundred sentences). The *first* step is the extraction of domain-specific terms from monolingual data (target and source side of a document partition), while the *second* is the creation of bilingual terminology starting from the monolingual ones. In order to obtain the best possible performance, we compare different approaches in both steps. At the monolingual level, we test the extraction using three term extraction tools. For bilingual alignment, we compare different alignment strategies. The two steps are detailed in the following subsections.

3.1 Monolingual Terminology Extraction

To identify and extract the most appropriate set of monolingual terms in our CAT scenario, we compare three term extraction tools: the KX toolkit (Pianta and Tonelli, 2010), TWSC (Pinnis et al., 2012) and AlchemyAPI.[3]

KX is a terminology extractor, which combines frequency information and part-of-speech patterns of n-grams to identify the most relevant terms in a corpus. It is freely available for English, Italian and German and was the best performing rule-based system in the Semeval2010 task on keyword extraction (Kim, Medelyan, Kan, and Baldwin, 2010). TWSC follows an approach which is very similar to KX, integrating morpho-syntactic patterns with statistical features. One of the main differences with respect to KX is the implementation of different co-occurrence statistics to rank term candidates and the treatment of nested terms. Nevertheless, we expect the performance of these two tools to be very similar. For both extraction tools, we limit the length of a term to 5-grams. A third system considered is AlchemyAPI. This commercial tool employs sophisticated statistical algorithms and linguistic knowledge to analyse textual content and extract topic keywords, but no further implementation details are given.

3.2 Bilingual Terminology Alignment

Once we obtain the lists of automatically extracted monolingual terms for the source and the target language, we perform different strategies for terminology alignment across languages (Figure 2).

Given a source term and the parallel sentence pair in which it appears, a set of possible translations is found by either *translating* the term or by applying a *word aligner*. In both cases, we use a technique similar to the methodology proposed by Ehrmann, Turchi, and Steinberger (2011), where the translation system is trained on the same data it needs to translate. This approach reduces the number of untranslated terms, since the translation system should know how to translate a source term seen in the training data. In our case, we train the SMT system and word aligner on the same data from which the bilingual terminology is extracted. Specifically, for the evaluation of the bilingual terminology alignment (Section 6.2), we train the system on the gold standard dataset in the IT domain, while for the translation evaluation (Section 6.3), the translation models were trained on the targeted document partition. For each term, the word aligner produces only one

[3] `http://www.alchemyapi.com/products/features/keyword-extraction/`

Approach / Method	Term lookup	Sentence lookup
Word alignment (WA) - aligning t_s with t_t based on word alignment information	given t_s, a link to t_t is built, if t_t has also been extracted by an extraction tool in s_t	given t_s, a link to t_t is built, if t_t appears in s_t
SMT - translating t_s into t_t		
Term Aligner - given t_s, a link to t_t is built based on SVM binary classifier		
Phrasetable2Glossary - bilingual terminology of t_s and t_t is filtered based on frequency, direct phrase translation probability and POS information		

Fig. 2: Summary of Word Alignment, SMT methods, Term Aligner and Phrasetable2Glossary for Bilingual Terminology Alignment (t_s = source term; t_t = target term; s_t = target sentence)

possible candidate translation, while an n-best list of possible translation candidates is obtained by the SMT system.

Given a set of possible translations for each term, the correct one is retrieved by taking advantage of the parallelism between source and target sentences, whereby two methods are investigated: *sentence lookup* or *term lookup*. With the first, a target translation from the candidate list is accepted as correct if it matches a span of words in the target sentence. With the second, a translation is accepted if it has also been identified as a term in the target sentence by the monolingual term extractor. The *term lookup* method reduces the number of extracted bilingual terms but guarantees a better quality of the alignments.

We compare our strategies with Term Aligner (Aker et al., 2013) and the approach called 'Phrasetable2Glossary' (Thurmair and Aleksić, 2012). Term Aligner treats bilingual alignment as a classification problem. An SVM binary classifier is trained on data derived from the multilingual thesaurus EuroVoc,[4] using language dependent and independent features. The former are based on bilingual dictionaries created by the GIZA++ tool, while the latter use cognate-based features, such as the longest common subsequence/substring ratio, Dice similarity or the Levenshtein distance between a source and target term. Since our setting is different from the one presented in their original work, focusing on term alignment in comparable corpora, we limit the tool to search for terms that appear in the same parallel sentence pair. Moreover, to make the comparison fairer, Term Aligner accesses the required GIZA++ dictionaries, which were used for the word alignment projection strategy of our proposed framework.

The Phrasetable2Glossary approach uses pre-trained phrase tables to extract bilingual terminology. Each entry of the phrase table is analysed checking its frequency in the training data and its direct translation probability. If both frequency and probability values satisfy some pre-defined thresholds, the last step consists in

[4] `http://eurovoc.europa.eu/`

annotating the source and target phrases with POS information and filtering out entries that do not match syntactic patterns of terms. In our experiments, we tested frequency threshold values equal to 1, 3 and 5 and direct translation probability intervals, where p(e|f) is larger than 0.6, larger than 0.8, between 0.2 and 0.6 and between 0.4 and 0.8. Differently to other approaches, the Phrasetable2Glossary approach does not require any reference to the monolingual data and, in our setting, it leverages the same domain-specific phrase tables used by our SMT alignment approach.

4 Enhancing Terminology Translation

Once the domain-specific bilingual terms are automatically aligned, we integrate them into the workflow of the SMT system. In a typical translation scenario, a large project is usually split into partitions of around 3,000 words, which represent the average workload of a professional translator in the post-editing task per day. Translating partition$_n$, the decoder is supported by the extracted and aligned bilingual terminology from previous partitions (partition$_0$ ... partition$_{n-1}$) using either the XML markup or the cache-based models (Section 4.1). To further improve the translation quality of partition$_n$, the decoder takes advantage of log-linear weights obtained by running MERT (Minimum Error Rate Training) (Bertoldi, Haddow, and Fouet, 2009) over the previous partition (see Section 4.2).

To summarise, given the extracted bilingual terminology from the parallel sentences, we improve the translation quality of the SMT system by (*i*) using the bilingual terms during the translation process and (*ii*) running an incremental tuning on different sets of parallel sentences coming from different partitions.

4.1 Integration of Bilingual Terms into SMT

Focusing on a CAT scenario, where an SMT system should provide suggestions to the translator for each source sentence, we cannot retrain the whole model with additional domain-specific terms (Bouamor et al., 2012). Adding bilingual terms directly into the phrase table is not suitable either since it would require switching off the system (Bouamor, Semmar, and Zweigenbaum, 2011). Additionally, the results obtained in our preliminary experiments (Arcan et al., 2014b) showed that also the incremental training methods introduced by Levenberg, Callison-Burch, and Osborne (2010) and Denkowski, Dyer, and Lavie (2014), which make it possible to continuously add sentences without retraining the model, are not the best solution in our setting. We observed that incremental training does not perform well when short expressions, such as bilingual terminology, are continuously added. This depended on the re-tuned features that led the SMT system to generate short and inconsistent translations. For these reasons, we compare two methods that can be easily used at run-time without altering the normal work of the SMT system and differentiate well between domain-specific and general translations: the widely-used XML markup and the cache-based model (Bertoldi et al., 2013).

XML Markup With the XML markup approach, external knowledge is directly passed to the decoder by specifying the translation of a specific span of the source sentence. In the case of multiple translations of the same source span, a score can be used to indicate the level of association between the source and target phrases. We compared three different XML settings, i.e., *exclusive*, *inclusive* and *constraint.* In the *exclusive* setting, only the proposed translations are used for the input phrase. Translation candidates stored in the phrase table and overlapping with that span are ignored. Differently, the proposed translations compete with the translation candidates in the phrase table, if the *inclusive* setting is selected. In the *constraint* setting, the proposed translations compete with phrase table choices that contain the specified translation.

For instance, in the following example: "*the <n translation="tipo di dati||tipo dati||tipo dei dati" prob="1||0.8||0.3">data type</n> of the column*" the XML markup approach allows a user to provide the decoder with three Italian translations (*tipo di dati*, *tipo dati* and *tipo dei dati*) of the English term "data type", along with their relative translation probabilities (*1*, *0.8* and *0.3*). The decoder will then use the proposed translations and probabilities to produce the final translation.

Cache-Based Models We analyse the use of cache-based translation and language models (Bertoldi et al., 2013) for integrating bilingual terms into an SMT system. The main idea behind these models is to combine a large static global model with a small, but dynamic local model. This allows users to define and dynamically adapt domain-specific models that are combined during decoding with the global SMT models built on the training data.

The cache-based model relies on a local *translation model* (CBTM) and *language model* (CBLM). The first is implemented as an additional phrase table providing one score. All entries are associated with an 'age' (initially set to 1), corresponding to the time when they were actually inserted. Each new insertion causes the ageing of the existing phrase pairs and hence their re-scoring. In the case of re-insertion of a phrase pair, the old value is set to the initial value. Phrase pairs in the model are scored based on the decaying function. In our experiments, we test different rewarding and penalising functions.[5] Similarly to the CBTM, the local *language model* is built to give preference to target terms found by the extraction tool. Each target term stored in CBLM is associated with a decaying function of the age of insertion into the model. Both models are used as additional features of the log-linear model in the SMT system. While the XML markup only substitutes the annotated source strings with a given translation without considering the surrounding context for proper lexical choice, the cache-based model offers a better integration of the terms in the final translation. In particular, when using the CB models, the decoder receives the source sentence without any extra information, but when translating it has access to extra translation (CBTM) and language (CBLM) models.

If we consider the previous example of "data type", the CBTM and CBLM will

[5] Hyperbola, power, exponential, cosine.

CBTM	CBLM
. . .	. . .
1 ‖ *data type* ‖ *tipo di dati*	*1* ‖ *tipo dei dati*
1 ‖ *data* ‖ *dati*	*1* ‖ *dati*
1 ‖ *type* ‖ *tipo*	*1* ‖ *tipo*
2 ‖ *data type* ‖ *tipo dati*	*2* ‖ *tipo dati*
4 ‖ *data type* ‖ *tipo dei dati*	*4* ‖ *tipo di dati*
. . .	. . .

Table 1: Entries in CBTM and CBLM related to "data type"

contain the entries reported in Table 1. Each entry in the CBTM contains age, source and target phrase, while only age and target phrase are present in the CBLM. In our example, age = 4 for the entry "*data type* ‖ *tipo dei dati*" means that it has been extracted from a sentence belonging to four partitions ago. It is interesting to note that, if single tokens are extracted from a long term (e.g. "*data* ‖ *dati*" and "*type* ‖ *tipo*"), the CB models can handle and provide them to the decoder. On the contrary, when using the XML markup methodology, the user needs to take a decision of what span to mark: "*data type*" or "*data*" and "*type*", because nested annotations are not allowed. This decision can affect the quality of the final translation. In our example, annotating "*data*" and "*type*" separately would probably produce "*dati tipo*", that is less accurate than "*tipi di dati*".

4.2 Incremental Tuning

The continuous extraction and collection of bilingual terms improves the domain-specific knowledge of the SMT system. This dynamically changes the capability of the SMT to correctly translate new sentences and the contribution of each component in the log-linear model. For this reason, when a new partition of parallel sentences is available ($partition_n$ in Figure 3), bilingual terms are first extracted. Then, before using them in the cache-based or XML markup module, the tuning step is performed using $partition_{n-1}$ as development set and taking advantage of all terms extracted from $partition_0$ to $partition_{n-2}$. When the new weights are computed, the bilingual terms extracted from $partition_{n-1}$ are added to the terms obtained from all the previous partitions, and the new configuration of the SMT system is used to translate $partition_n$. The aim of this procedure is to update the weights of each feature, taking into consideration the new translation capability of the model. The starting weights used by MERT at time $n - 1$ are obtained optimising the system at time $n - 2$. Once the new weights are computed, the old weights need to be overwritten. This is done by passing the new weights to Moses (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst, 2007) through XML tags for each incoming sentence, which required the extension of Moses with this new option.

An issue with incremental tuning is the risk of over-fitting the model on a small development set and then performing poorly on a test set, if it is very different from

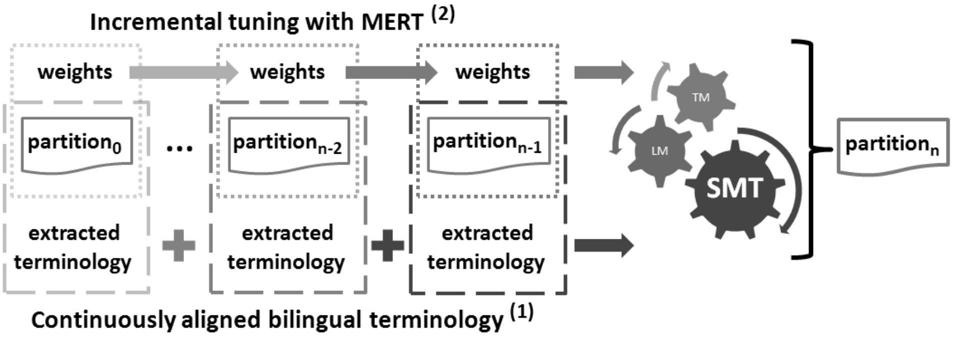

Fig. 3: Translating partition$_n$ with the bilingual terminology and optimized log-linear weights

the development one. In our scenario, this is prevented by the fact that all the sets come from the same document, or from different documents on a similar topic in the same project. Although it is important to tune an SMT system on a sufficiently large development set, reasonably good weights can be obtained even if such data are very few, as shown in Bertoldi and Federico (2009). In our framework, it is not possible to concatenate all the previous partitions to enlarge the development set, because the presence of already extracted bilingual terms in the cache-based models would artificially favour the cache-based components during tuning.

5 Experimental Setting

In this section, we perform a set of experiments demonstrating the capability of our framework to extract high quality domain-specific bilingual terms from a small amount of parallel data and to integrate them in the translation task. The language pairs considered are English-Italian and English-German, performing translations in both directions. To identify the best monolingual term extraction tool as well as the most suitable bilingual alignment, we developed a gold standard based on freely available domain-specific data. Two datasets belonging to the IT domain, namely a portion of GNOME project data (4,3K tokens)[6] and KDE Data (9,5K),[7] are used for domain-specific term extraction for both language pairs.

Our proposed framework for integrating SMT systems with automatically extracted bilingual terminology is tested on a subset of the EMEA corpus (Tiedemann, 2009) for the medical domain (18K tokens) and an IT corpus (18K), extracted from a software user manual. Each corpus is split into partitions of around 3,000 tokens, i.e. the daily workload of a professional translator in post-editing, resulting in six partitions each.

For each domain, we perform the evaluation of the extracted monolingual and bilingual terms against the manually annotated KDE and GNOME datasets by

[6] `https://l10n.gnome.org/`
[7] `http://i18n.kde.org/`

calculating Precision, Recall and F-measure (F_1). The BLEU metric (Papineni, Roukos, Ward, and Zhu, 2002) is used to automatically evaluate the translation quality of the EMEA and the IT manual datasets. BLEU is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. Those scores, between 0 and 100 (perfect translation), are then averaged over the whole *test dataset* to reach an estimate of the translation overall quality.

For each translation task, we use the statistical translation toolkit Moses, where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The SRILM toolkit (Stolcke, 2002) was used to build the 5-gram language model.

For a broader domain coverage of the generic SMT system, we merged parts of JRC-Acquis (Steinberger, Pouliquen, Widiger, Ignat, Erjavec, Tufis, and Varga, 2006), Europarl (Koehn, 2005) and OpenSubtitles2013 (Tiedemann, 2009), obtaining a training corpus of ∼35M tokens for the English-Italian language pair and ∼36M for the English-German. The generic SMT system used in all our experiments is trained on this merged general resource. The difference in size between the domain-specific and the generic data is evident, i.e., approximately few thousands vs. more than 30 million tokens. This reflects a real CAT scenario, where only a small amount of domain-specific data is available.

Gold standard creation In order to evaluate the quality of monolingual and bilingual terms, we created a terminological gold standard for the IT domain. Two annotators with linguistic background in English, Italian and German were asked to mark all domain-specific terms inside the GNOME and KDE datasets. Domain specificity was defined as all (multi-)words that are typically used in the IT domain and that may have different translations in other domains. The intersection between the monolingual term lists provided by the two annotators was then considered as the monolingual gold standard.

In a second step, given two monolingual gold standards, the annotators had to manually create a bilingual pair if two domain-specific terms were found, one being the translation of the other. If a term in one language was the translation of part of a term in the other (e.g. *"Dateien"* and *"hidden files"*), only the intersection was included in the bilingual gold standard (in our example *"Dateien - files"*). The Dice coefficient (Dice, 1945), computed at token level on GNOME and KDE monolingual data as a measure of inter-annotator agreement, was 0.87 on English data, 0.68 on Italian and 0.82 on German. This shows a substantial agreement, even if more partial matches (i.e. terms whose boundaries are uncertain) were annotated on Italian data. The statistics on the annotated monolingual and bilingual data are shown in Table 2.[8]

[8] The annotated data are made freely available to the research community under: `http://hlt-mt.fbk.eu/technologies/bittercorpus`

Monolingual Terms

	English		Italian		English		German	
	GNOME	KDE	GNOME	KDE	GNOME	KDE	GNOME	KDE
Single-tokens	162	311	185	301	201	457	339	717
MWE	120	321	128	326	186	499	68	210
Total	282	632	313	627	387	926	407	868

Bilingual Terms

	English - Italian		English - German	
	GNOME	KDE	GNOME	KDE
Total	237	637	338	447

Table 2: Statistics of monolingual and bilingual terminology within the BitterCorpus.

6 Evaluation

In this section, we evaluate the performance of monolingual term extraction and bilingual terminology alignment on the IT domain. Moreover, we evaluate for the IT and medical domain the translation quality obtained by applying different injection approaches of bilingual terms into an SMT system.

6.1 Monolingual Term Extraction

Our first evaluation concerns monolingual term extraction from English, Italian and German documents provided by KX, AlchemyAPI and TWSC extraction tools. The tool performance is evaluated considering only term exact matches. As shown in Table 3, none of the three tools performs best for all considered languages. KX performance remains stable in all settings, Alchemy API is strongly affected by the language considered and TWSC performs poorly on German. Therefore, we select AlchemyAPI for the extraction of monolingual terms in English, TWSC for Italian and KX for German, to be used in the next phase.

We performed a manual analysis of the monolingual data extracted from the three tools, to understand their different behaviours. On English, the three lists of extracted terms do not show striking differences. In fact, Alchemy API scores the best F-measure, but highest precision is achieved by TWCS and best recall is obtained with KX. The only general difference is that Alchemy extracts multi-word terminology up to 5-6 tokens, while the other two systems prefer shorter multi-words. This favours the performance of Alchemy API in the technical domains considered, where longer, very specific multi-words are quite common in English (e.g. *"ip address of the server"*, *"setting system wide proxy information"*). As for Italian, Alchemy API and KX are affected by different issues. The former often extracts generic multi-word expressions (e.g. *"possibili modi / possible ways"*, *"effetto collaterale / side effect"*), which are not included in our gold standard terminol-

GNOME - KDE (English)	# of Terms	Precision	Recall	F_1
Alchemy API	665	0.393	0.571	**0.466**
KX	1115	0.293	0.596	0.393
TWSC	496	0.413	0.372	0.391
GNOME - KDE (Italian)	# of Terms	Precision	Recall	F_1
Alchemy API	304	0.309	0.167	0.213
KX	950	0.271	0.452	0.339
TWSC	765	0.362	0.481	**0.412**
GNOME - KDE (German)	# of Terms	Precision	Recall	F_1
Alchemy API	492	0.080	0.048	0.044
KX	1969	0.261	0.644	**0.369**
TWSC	529	0.306	0.213	0.251

Table 3: Evaluation of monolingual term extraction for English, Italian and German.

ogy. The latter, instead, includes in the extracted list also verbal forms, which are only marginally present in the gold standard. This is probably due to POS-tagging errors. For German, Alchemy API performs worst due to the lack of information in its background knowledge. KX is in this case the best performing tool, because it extracts with good accuracy shorter terms, that are more frequent in German due to the wide presence of compounds (e.g. *"Endbenutzerzertifikate / end user certificates"*, *"Himmelskartenkontrolle / Sky map control"*).

6.2 Bilingual Terminology Alignment

In this step, we evaluate the creation of bilingual terminology using word alignment and SMT methods, and compare them against the performance of Term Aligner and Phrasetable2Glossary (see Figure 2 in Section 3.2). Given the list of automatically extracted terms in the source and the target language, we test the *term* and *sentence lookup* strategy to obtain a high-quality terminological list in the target language. As a comparison, we also evaluate bilingual terminology extraction starting from gold monolingual terms, in order to better investigate the contribution of the monolingual and bilingual steps.

When using automatically extracted monolingual terms, the SMT *sentence lookup* method mostly outperforms other approaches in terms of F-measure (Table 4 and 5). The advantage of this method lies in the fact that identifying translated terms (out of an n-best list) in the target sentence enables a larger search space compared to the stricter *term lookup*, which aligns terms only if they were extracted independently by the term extraction tools. However, when aligning terms for English to German translation, the quality of the *sentence lookup* approach drops. This is related to the difficulty of SMT to translate into German. In our experiment the stricter word alignment strategy shows to be the best option for a target language with a complex morphology (Table 5).

As for the Term Aligner tool, we run experiments with different cognate sim-

English → Italian

	Auto. Extr. Terms			Gold Standard Terms		
Translation Projection	Precision	Recall	F1	Precision	Recall	F1
Word A., sent. lookup	0.251	0.273	0.259	0.463	0.399	0.425
Word A., term lookup	0.430	0.076	0.129	0.768	0.285	0.415
SMT n-best, sent. lookup	0.220	0.374	**0.271**	0.426	0.577	0.483
SMT n-best, term lookup	0.451	0.150	0.221	0.779	0.517	**0.616**
Term Aligner	Precision	Recall	F1	Precision	Recall	F1
cognate threshold, 0.3	0.308	0.136	0.188	0.776	0.466	0.582
cognate threshold, 0.5	0.375	0.131	0.191	0.854	0.467	0.603
cognate threshold, 0.7	0.401	0.128	0.189	0.900	0.467	0.613
cognate threshold, 0.9	0.396	0.122	0.182	0.904	0.449	0.597
Phrasetable2Glossary	Precision	Recall	F1	Precision	Recall	F1
freq., prob., POS filtering	0.372	0.107	0.167	/	/	/

Italian → English

	Auto. Extr. Terms			Gold Standard Terms		
Translation Projection	Precision	Recall	F1	Precision	Recall	F1
Word A., sent. lookup	0.222	0.242	0.229	0.430	0.367	0.392
Word A., term lookup	0.412	0.086	0.140	0.753	0.314	0.442
SMT n-best, sent. lookup	0.224	0.382	**0.276**	0.450	0.605	0.508
SMT n-best, term lookup	0.412	0.153	0.219	0.766	0.500	**0.602**
Term Aligner	Precision	Recall	F1	Precision	Recall	F1
cognate threshold, 0.3	0.322	0.142	0.195	0.755	0.432	0.548
cognate threshold, 0.5	0.390	0.136	0.198	0.851	0.445	0.583
cognate threshold, 0.7	0.412	0.136	0.199	0.894	0.455	0.601
cognate threshold, 0.9	0.397	0.125	0.184	0.895	0.438	0.585
Phrasetable2Glossary	Precision	Recall	F1	Precision	Recall	F1
freq., prob., POS filtering	0.270	0.065	0.105	/	/	/

Table 4: Bilingual term alignment using the automatically extracted monolingual terms (left) and gold monolingual terms from the gold standard (right).

ilarity thresholds (from 0.1 to 1.0 with steps of 0.1) and a classifier trained on the EuroVoc data, as reported in the original paper by Aker et al. (2013). The best performance on the English-Italian language pair alignment is achieved with a threshold of 0.5 and 0.7. For both English-German alignment directions, best performance is obtained if a cognate threshold score of 0.3 is selected. The performance of Term Aligner for this language pair decreases constantly as we increase the cognate score. Nevertheless, the alignment quality of Term Aligner is substan-

English → German

	Auto. Extr. Terms			Gold Standard Terms		
Translation Projection	Precision	Recall	F1	Precision	Recall	F1
Word A., sent. lookup	0.247	0.199	**0.217**	0.514	0.289	0.370
Word A., term lookup	0.444	0.052	0.089	0.893	0.141	0.243
SMT n-best, sent. lookup	0.309	0.134	0.176	0.580	0.260	0.357
SMT n-best, term lookup	0.479	0.073	0.119	0.832	0.233	0.363
Term Aligner	Precision	Recall	F1	Precision	Recall	F1
cognate threshold, 0.1	0.303	0.099	0.158	0.307	0.247	0.270
cognate threshold, 0.3	0.387	0.090	0.132	0.803	0.086	0.167
cognate threshold, 0.5	0.563	0.046	0.123	0.275	0.321	0.281
cognate threshold, 0.7	0.421	0.059	0.098	0.717	0.268	**0.375**
Phrasetable2Glossary	Precision	Recall	F1	Precision	Recall	F1
freq., prob., POS filtering	0.272	0.039	0.069	/	/	/

German → English

	Auto. Extr. Terms			Gold Standard Terms		
Translation Projection	Precision	Recall	F1	Precision	Recall	F1
Word A., sent. lookup	0.227	0.176	0.195	0.422	0.237	0.304
Word A., term lookup	0.537	0.066	0.110	0.857	0.162	0.272
SMT n-best, sent. lookup	0.226	0.183	**0.202**	0.530	0.239	0.328
SMT n-best, term lookup	0.367	0.074	0.117	0.819	0.237	0.367
Term Aligner	Precision	Recall	F1	Precision	Recall	F1
cognate threshold, 0.1	0.329	0.103	0.166	0.240	0.201	0.217
cognate threshold, 0.3	0.508	0.104	0.156	0.758	0.114	0.206
cognate threshold, 0.5	0.416	0.065	0.146	0.267	0.318	0.277
cognate threshold, 0.7	0.354	0.060	0.098	0.701	0.271	**0.378**
Phrasetable2Glossary	Precision	Recall	F1	Precision	Recall	F1
freq., prob., POS filtering	0.361	0.028	0.052	/	/	/

Table 5: Bilingual term alignment using the automatically extracted monolingual terms (left) and gold monolingual terms from the gold standard (right)

tially lower compared to the translation-based approaches. This can depend on the difference between the bilingual terms used to train the classifier (JRC-Acquis and EuroVoc dataset) and the domain-specific terms in our gold standard (GNOME and KDE dataset).

At last, we compare our methodology to the Phrasetable2Glossary approach. Due to the direct usage of the phrase tables, the approach does not require any source or target sentences. Therefore, an evaluation on bilingual terminology built out of the monolingual terms of the gold standard is not feasible. The best results were obtained by using the minimal occurrence frequency (larger or equal 1) and a translation probability larger than 0.6. Since this approach relies on the phrase table and does not consider monolingual terms in the source and target sentences, it generates bilingual alignments with a substantial lower quality in terms of the F-measure.

We also compare the alignment obtained from automatically extracted terms in each language with the performance using terms from the monolingual gold standards. As expected, when using gold monolingual terms, we obtain significantly higher results compared to the real scenario described before. For the English-Italian language pair, the SMT *term lookup* method performs best. This shows that *term lookup* is more sensitive to the heterogeneity in automatically extracted data than the approach based on *sentence lookup* (right part of Table 4). For the English-German pair in both directions, Term Aligner slightly outperforms our proposed approach, showing that it tends to handle better the compound and rich morphological features of German but only if monolingual terms of high quality are available.

We manually evaluated the bilingual pairs extracted by the tools in the different settings, and we observed that the word alignment approach produces significantly less MWE alignments compared to the SMT approach. In detail, for the English-Italian language pair using the *sentence lookup* method, only 337 source and target MWE were aligned, compared to 553 within the SMT approach. Due to the larger possibility of finding an SMT generated target term in the sentence, alignments like *window focus mode* → *la modalità di focus*, *sorted by size* → *disposti per dimensione* or *contenuti ingranditi* → *magnified contents* (for Italian → English) are entirely missing in other alignment approaches.

Examining the results on English → German terminology alignment, where word alignment outperformed the generally better SMT approach, we observed that the former provided only one accurate bilingual alignment per source term, e.g. *display* - *anzeigen* or *activated* - *aktiviert*. Differently, the SMT n-best approach aligned a larger amount of English-German terminology, which also included wrongly aligned terms. In detail, *display* was not only aligned to the German *anzeigen*, but also to *duplizieren* (en. to duplicate). Similarly, *activated* was correctly aligned to *aktiviert*, while it was also wrongly aligned with *legt* (en. puts/lays) and *detailiert* (en. detailed). This behaviour can be explained by the morphological complexity of German, in combination with the small amount of training data.

Although expressions like *mouse*, *keyboard* or *mouse button* were correctly aligned with Term Aligner, it failed to generate alignments of more specific terminology, like *uuid property / proprietà uuid* or *contenuti ingranditi / magnified contents.* This was observed on the English-Italian as well as on English-German language pair (*space-separated string / kommata getrennten zeichenkette*). Term Aligner behaviour in this setting is affected by the domain of the test data, which is different from that of the training data.

As for Phrasetable2Glossary, the low F-measure depends on the fact that this approach can hardly distinguish between domain-specific and generic expressions within the phrase table. Therefore, it extracts generic expressions, such as *things-cose* or *ways-modi*, which are not annotated in the English-Italian gold standard as domain-specific terms. Similarly, English-German generic expressions, such as *examples-Beispiele* or *case-Fall*, were extracted. Furthermore, mistakes in POS tags caused in some cases wrongly annotated table entries, which in turn led to the extraction of non-terminological pairs. For example, the verbs *decouple* or *compute* were annotated as nouns, therefore the entries and their translations were wrongly added to the bilingual term list.

Although Table 4 for English-Italian and Table 5 for English-German show low F-measure scores on bilingual terminology alignment, we show in the next section that the SMT system can benefit from the extracted terms and can compete also with wrongly aligned terminology.

6.3 Translation Evaluation

After we identified the best monolingual term extraction tool for each language in Section 6.1 (AlchemyAPI for English, TWSC for Italian, KX for German) and the best performing approach for bilingual terminology alignment in Section 6.2 (SMT n-best sentence lookup), we carry out the final translation evaluation on the EMEA and IT manual datasets with the help of aforementioned tools and approaches.

As described in Section 4, we split our data into several partitions and each of them is translated by:

- a baseline SMT system that was built with the general resource, without integrating terminology,
- XML markup approach to include the terminology paired with the baseline SMT system,
- cache-based model, where bilingual terminology was used to generate CBTM and CBLM in support of the general SMT system.

The probability passed to the XML markup for each aligned bilingual term is set according to the translation probability obtained by the SMT system to project the source term into the target language. Since a source term may have different translation candidates, the different translation probabilities give preference to more probable translations. For each set of partitions, incremental tuning was run to update the log-linear weights. For the sake of comparison, we also run MERT on each partition starting with flat weights (non-incremental tuning). As shown in Figure 4, incremental tuning outperforms, in general, non-incremental tuning at the partition level. At the document level (partition 6) the incremental tuning approach always generates better translations compared to the non-incremental approach.

In Figure 5 we report BLEU scores for the entire document. The approximate randomization approach (Clark, Dyer, Lavie, and Smith, 2011) is used to test whether differences among system performances are statistically significant. Results in the figure marked with * are statistically significantly better than the baseline with a p-value < 0.05.

Among different decay functions in the cache-based models, we report only the

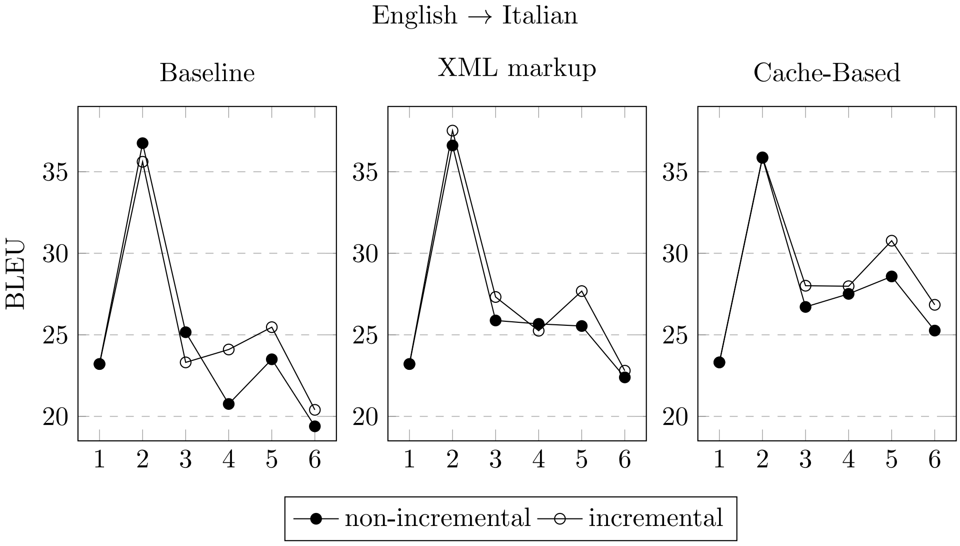

Fig. 4: Comparison between non-incremental and incremental tuning for each document partition in the IT domain.

negative power decay function of the age, which achieves the best overall performance. This confirms the results described in Bertoldi et al. (2013) also when the approach is applied to a different context. For the XML markup approach, we compare three different settings, e.i., *exclusive*, *inclusive* and *constraint*. Based on our evaluation, the *inclusive* setting performs best, due to the possibility that wrongly aligned terminology can still be corrected by entries in the phrase table.

Comparing the three methods (baseline, XML markup, cache-based models) for the target language pairs, we notice that the translation performance of cache-based models always outperforms all the other methods in both domains and translation directions.

In the translation from English into German, both approaches, i.e., XML markup and cache-based models, significantly outperform the baseline system, while the cache-based approach provides better translations than the XML markup (IT: 27.50 vs. 24.69; EMEA: 25.49 vs. 22.46). When translating from German into English, the XML markup performs better than the baseline in both domains (IT: 29.96 vs. 29.34; EMEA: 25.51 vs. 25.21), but statistical significance can be observed only for the IT domain. Examining the results for the English-Italian language pair, the XML markup always significantly outperforms the baseline approach, but in terms of BLEU it generates worse results compared to the cache-based approach.

Comparing the improvements over the language pairs, we observe a large improvement in terms of BLEU for the English-German language pair. While for the English-Italian language pair, the averaged improvement of the cache-based approach over the baseline is 10.36% (1.86% to XML), the improvement raises to 24.16% (10.42% to XML) for the English-German language pair.

The global results on the document level (Figure 5) are also confirmed at the

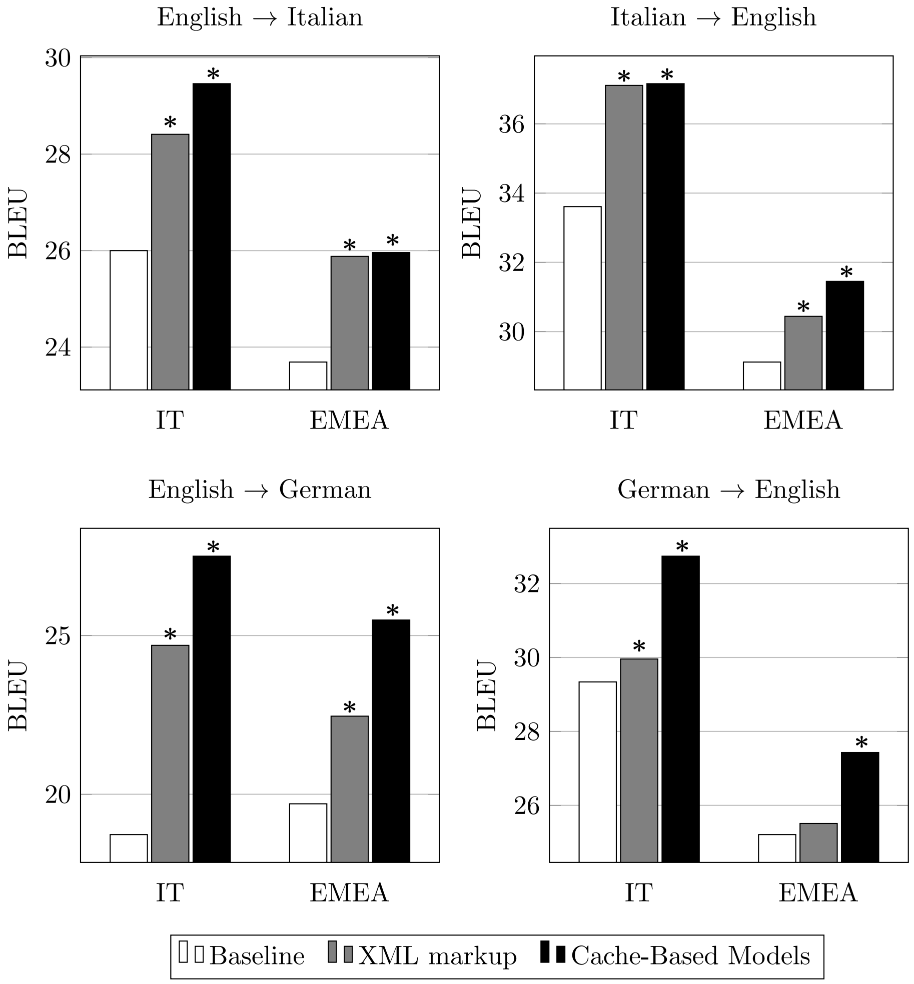

Fig. 5: Automatic evaluation (BLEU) based on the baseline, XML markup and cache-based approach (* statistically significant compared to baseline).

partition level. Figure 6 shows the performance for the English-German language pair for each partition, where the cache-based model always outperforms the XML markup. Compared to the baseline approach, the later shows improved results for the English to German translation direction and comparable, when translating from German into English.

In order to investigate to what extent the approaches differ from a translator's point of view, we manually inspected the translations into Italian and German produced by the XML markup and cache-based approach. The quality of the two translation versions generally reflects the results reported in Figure 5. The XML markup approach takes into account the surrounding context of a translated string only partially, while the cache-based one usually shows a better context-awareness. Specifically, it usually provides a better agreement between adjective and noun

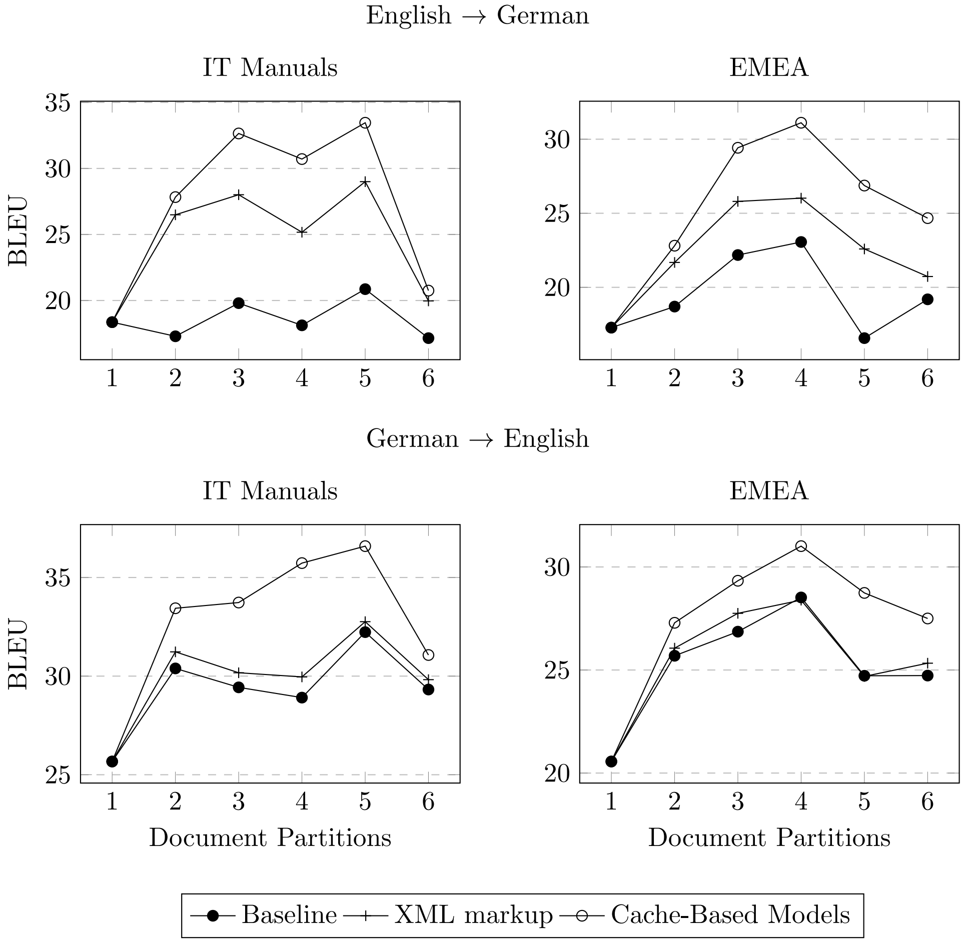

Fig. 6: Automatic evaluation (BLEU) for the English-German language pair based on the XML markup and cache-based models for each document partition.

(which in Italian and German bear gender and number information). It also tends to provide more frequently the correct agreement between noun and verb, and even to translate English verbs in the progressive form as nouns, when appropriate. Instead, sentences translated with XML markup often contain gaps as well as agreement and reordering issues because not all terms are translated.

We report an example where the source sentence is "*Following are the steps for windows operating system*", translated into Italian. The XML markup output is "*seguente sono i passaggi per finestre operanti data del sistema.*", while the cache-based translation "*seguenti sono i passaggi per finestre sistema operativo.*". In the second version, the agreement between "seguenti" ("following") and the verb is correct, while it is missing in the XML markup output. Besides, the cache-based model translated "operating system" as a multi-word ("sistema operativo"), while it is translated word by word in the XML markup version.

These differences are more evident in the medical domain, where the language

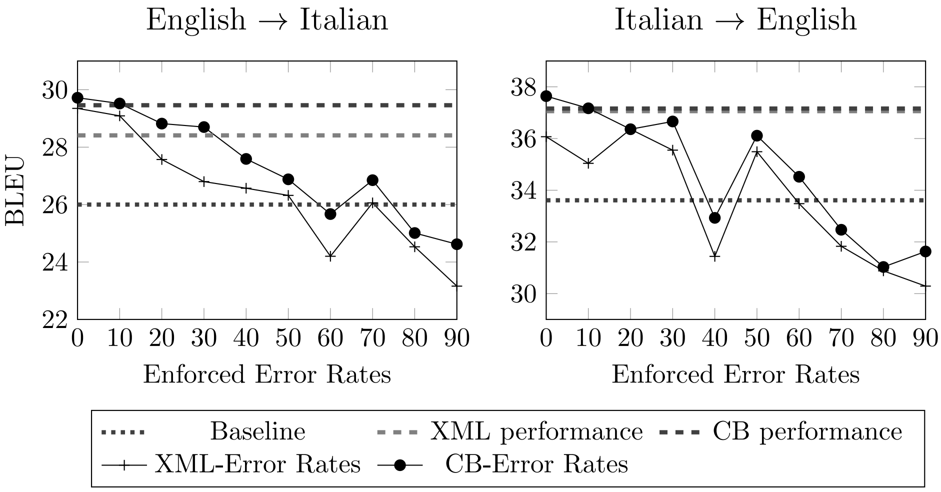

Fig. 7: Impact of misaligned terminology on translation quality.

is highly specific and noun phrases are often composed of complex noun chains (e.g. 'an in vitro mammalian cell assay', 'increased lipid and uric acid values'), with implicit underlying dependencies. This is confirmed also by the results in Figure 5, showing that translation quality is generally lower than for the IT domain.

Similar to the translations into Italian, for the English-German pair the XML approach tends to split MWEs into separate units, although useful MWEs are stored in the provided bilingual terminology. For example, in the sentence *use of the social share plug-in* the XML approach translates *social* and *share* separately, resulting in the German translation *sozialen* and *Anteil*, which is not correct considering the IT domain. The cache-based approach, instead, translates it correctly (*sozialen Netzwerken*). In the medical domain, XML-based approach translated *controlled portions* separately into *kontrollierten Teile* (en. "controlled parts"), while the cache-based approach takes advantage of the contextual information, e.g. *... of clinical trials...* to improve the translation into *kontrollierten Phasen*.

The manual evaluation of translations into Italian and German confirms the observations we made about the automatic evaluation, in particular, the benefit of using the cache-based approach to integrate terminology into an SMT system. The XML approach, with its different settings, often does not entirely explore the contextual information around the provided terminology to be translated. This limitation can lead to issues related to wrong word agreement and as shown also to translations into a wrong domain.

7 Impact of Misaligned Terminology on Translation Quality

Since bilingual terminology is automatically extracted, it is likely that misaligned bilingual terms can be injected into the SMT system. This would clearly affect the final quality of the translated texts. In order to quantify this impact, we investigate which method, the cache-based models or the XML markup, is most robust to the injection of wrongly aligned bilingual terms. For this, we focus on the documents in the IT domain for the English-Italian language pair.

In the first step, we take the automatically generated bilingual terminology and manually discard wrongly aligned entries. By feeding the system with 'gold' bilingual terminology (0 Error Rate in Figure 7), we slightly improve the BLEU score compared to the best performance of the proposed approach using the cache-based method.

Next, we randomly align an increasing number of bilingual terminology, from 10% to 90%. As expected, the more noise we introduce into the SMT system, the more the BLEU score decreases.

In terms of robustness to misaligned bilingual terms, we observe that the cache-based method always outperforms the XML method for all enforced error rates. This depends on the rigid replacement of the provided translations implemented in the XML markup. Furthermore, the cache-based language model (CBLM, see Section 4.1) integrates the extracted domain-specific terminology on the target side, which allows the CB models to better handle noisy bilingual terms.

These additional experiments also confirm the benefits of providing bilingual terminology to the SMT system. Both methods, cache-based and XML, outperform the baseline system (small black dotted line), even if 50% of the extracted terminology is misaligned. This demonstrates that bilingual terminology can be efficiently exploited by the SMT system in a CAT scenario.

8 Conclusions

In this work, we describe a framework to enhance translation quality by exploiting bilingual terms extracted from parallel sentences daily produced by professional translators. This small amount of parallel data is used to continuously improve a generic SMT system by optimising the log-linear weights on these specific data. Furthermore, we investigate the integration of the extracted bilingual terms into the SMT system. We compare the performance of the cache-based model with the widely-used XML markup.

Our proposed framework shows significant improvements for two language pairs, i.e. English-Italian and English-German, in the IT and medical domain. We also evaluate the robustness of the term injection method using artificially misaligned bilingual terminology. This experiment demonstrates that the cache-based model has a better capability of ignoring wrongly aligned terminology compared to the XML markup. Furthermore, we observed that bilingual terminology, automatically extracted from the IT domain, contains aligned terms with good quality since it matches BLEU scores of the bilingual terminology with 10% misaligned entries. This can be observed both for the English-Italian and for the Italian-English translation direction.

This work was designed in order to address three research questions presented in Section 1. As regards the question concerning which strategy should be chosen for monolingual terminology extraction in our scenario, we showed that there is not a general-purpose approach valid for all languages. Instead, each tool considered has language-specific strengths. A possible future research direction may concern the combination of several extraction tools into an ensemble method that is able to leverage the strength of each single tool.

We addressed the second question, regarding the alignment of bilingual terminology starting from monolingual terms, by showing that a wide search space is beneficial to the alignment quality. In particular, combining SMT with the n-best and *sentence lookup* strategy yields best results. However, when the target language is morphologically complex like German, word alignment is a better alternative. Finally, the challenge to find the best strategy for bilingual term injection into an SMT system in a CAT environment was best addressed with the cache-based approach, applying continuous updates of the weights in the log-linear model.

Although several aspects on how to leverage the work of professional post-editors are still under investigation, our work shows that significant gains in translation quality can be obtained by including bilingual terms inferred from human translations. This confirms that information leveraged from post-edits is a valuable resource that cannot be ignored in the future to achieve high-quality machine translation.

References

Ahmet Aker, Monica Paramita, and Robert Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 402–411, Sofia, Bulgaria.

Mihael Arcan, Christian Federmann, and Paul Buitelaar. 2012. Experiments with term translation. In *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 67–82, Mumbai, India.

Mihael Arcan, Claudio Giuliano, Marco Turchi, and Paul Buitelaar. 2014a. Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pp. 22–31, Dublin, Ireland.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2014b. Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment. In *Association for Machine Translation in the Americas (AMTA)*, pp. 54–68, Vancouver, Canada.

Luisa Bentivogli, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):388–399.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 182–9, Athens, Greece.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *Prague Bull. Math. Linguistics*, 91:7–16.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of Machine Translation Summit XIV*, pp. 35–42, Nice, France.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2011. Improved statistical machine translation using multiword expressions. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011)*, pp. 15–20, Barcelona, Spain.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pp. 674–9, Istanbul, Turkey.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 176–181, Portland, Oregon.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 515–521, Kyoto, Japan.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 395–404, Gothenburg, Sweden.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Recent Advances in Natural Language Processing, (RANLP)*, pp. 118–124, Hissar, Bulgaria.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, San Diego, California.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The MateCat Tool. In *Proceedings of 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pp. 129–132, Dublin, Ireland.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 439–448, Paris, France.

Barry Haddow and Philipp Koehn. 2012. Analysing the Effect of Out-of-Domain Data on SMT Systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 422–432, Montréal, Canada.

Matthias Heyn. 1996. Integrating machine translation into translation memory systems. In *Proceedings of the EAMT Machine Translation Workshop, TKE'96*, pp. 113–126.

Masaki Itagaki and Takako Aikawa. 2008. Post-MT Term Swapper: Supplementing a Statistical Machine Translation System with a User Dictionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pp. 1584–8, Marrakech, Morocco.

Su Nam Kim, Tim Baldwin, and Min-Yen Kan. 2009. An unsupervised approach to domain-specific term extraction. In *Australasian Language Technology Workshop*, pp. 94–8, Sydney, Australia.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 21–6, Uppsala, Sweden.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pp. 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180, Prague, Czech Republic.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pp. 83–91, Nice, France.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 394–402, Los Angeles, California.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

T. Okita and A. Way. 2010. Statistical Machine Translation with Terminology. In *Proceedings of the First Symposium on Patent Information Processing (SPIP)*, pp. 1–8, Tokyo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–8, Philadelphia, Pennsylvania.

Emanuele Pianta and Sara Tonelli. 2010. KX: A flexible system for Keyphrase eXtraction. In *Proceedings of SemEval 2010, Task 5: Keyword extraction from Scientific Articles*, pp. 170–3, Uppsala, Sweden.

Marcis Pinnis. 2015. Dynamic terminology integration methods in statistical machine translation. In *Proceedings of the Eighteenth Annual Conference of the*

European Association for Machine Translation (EAMT 2015), pp. 89–96, Antalya, Turkey.

Marcis Pinnis and Raivis Skadins. 2012. MT adaptation for under-resourced domains - what works and what not. In *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, pp. 176–184. doi: 10.3233/978-1-61499-133-5-176. URL `http://dx.doi.org/10.3233/978-1-61499-133-5-176`.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*, pp. 91–6, Jeju Island, Korea.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pp. 47–54, Singapore.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 2142–7, Genoa, Italy.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pp. 901–4, Denver, USA.

Gregor Thurmair and Vera Aleksić. 2012. Creating term and lexicon entries from phrase tables. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pp. 253–260, Trento, Italy.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Proceeding of Recent Advances in Natural Language Processing*, pp. 237–248, Borovets, Bulgaria.

Spela Vintar and Darja Fišer. 2008. Harvesting multi-word expressions from parallel corpora. In *Proceedings of European Language Resources Association*, pp. 1091–6, Marrakech, Morocco.

Marion Weller, Alexander Fraser, and Ulrich Heid. 2014. Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pp. 11–8, Dubrovnik, Croatia.

Chien-Cheng Wu and Jason S. Chang. 2003. Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing*, pp. 1–20, Taiwan.

Deyi Xiong, Fandong Meng, and Qun Liu. 2016. Topic-based term translation models for statistical machine translation. *Artificial Intelligence*, 232:54–75.