Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment

Mihael Arcan mihael.arcan@insight-centre.com
Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland

Marco Turchi turchi@fbk.eu
Sara Tonelli satonelli@fbk.eu
FBK - Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy

Paul Buitelaar paul.buitelaar@insight-centre.com
Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland

Abstract

In this paper, we address the problem of extracting and integrating bilingual terminology into a Statistical Machine Translation (SMT) system for a Computer Aided Translation (CAT) tool scenario. We develop a framework that, taking as input a small amount of parallel in-domain data, gathers domain-specific bilingual terms and injects them in an SMT system to enhance the translation productivity. Therefore, we investigate several strategies to extract and align bilingual terminology, and to embed it into the SMT. We compare two embedding methods that can be easily used at run-time without altering the normal activity of an SMT system: XML markup and the cache-based model. We tested our framework on two different domains showing improvements up to 15% BLEU score points.

1 Introduction

Recent studies (Federico et al., 2012; Läubli et al., 2013; Green et al., 2013) have shown significant productivity gains when human translators post-edit machine translation output rather than translating documents from scratch. This evidence has raised interest in the integration of machine translation systems within CAT software. In this context, an important open issue is how to support translators with domain-specific information when dealing with highly specific texts, i.e. manuals coming from different domains (information technology (IT), legal, agriculture, etc.). Translation tools such as Google Translate,[1] Bing Translator[2] or open source SMT systems such as Moses (Koehn et al., 2007) trained on generic data are the most common solutions, but they often result in unsatisfactory translations. A valuable alternative to support professional translators is represented by online terminology resources, e.g. IATE,[3] which are continuously updated and can be easily queried. However, the manual use of these services can be very time demanding when working with a CAT tool. For these reasons, the automatic identification and integration of bilingual domain-specific terms into an SMT system is a crucial step towards increasing translation quality of high-specific texts in a CAT environment. This also reduces translators' initial overload when dealing with different domains, because terminological lists are managed directly by the SMT system and no additional human intervention for retrieving domain-specific terminology is required.

[1] `http://translate.google.com/` [2] `http://www.bing.com/translator`
[3] Inter-Active Terminology for Europe, `http://iate.europa.eu/`

In this paper, we propose a framework for extracting bilingual terms from parallel data and using them to enhance the performance of an SMT system embedded in a CAT tool. We focus on a real scenario, where a large translation project is split across different translators and each translator post-edits daily a limited amount of sentences provided by the SMT system. Our approach takes advantage of such post-edited data to gather bilingual domain-specific terms. The parallel data produced every day are then used to continuously improve a generic machine translation system by *(i)* automatically injecting the bilingual terms into the SMT system, and *(ii)* optimising the log-linear weights on this specific data.

Bilingual term extraction is performed in two steps. First, the source and the target sides of the data are processed by a keyword extractor to identify the most relevant terms in each language. Taking advantage of the parallel data, each monolingual term in the source language is paired with a term in the target language. We perform this step by comparing different techniques, showing that simple approaches based on word alignment and term translation are more robust and more efficient than the state-of-the-art method based on supervised classification (Aker et al., 2013).

As regards the integration of the bilingual terms in an SMT system, we cannot apply well-known approaches (Bouamor et al., 2012) adding the terms to training data or at the end of the phrase table, because in our CAT scenario we cannot stop the translation service and let translators wait for a long training time. For this reason, we investigate for the first time the integration of cache-based translation and language models (Bertoldi et al., 2013) in the context of terminology embedding comparing them with the XML markup technique. The cache-based model makes it possible to periodically add bilingual terms into an SMT system in real-time, without the need to stop it. In addition, we compare the cache-based models with a recently developed technique, namely the *Realtime Adaptive Translation Systems with cdec* (Denkowski et al., 2014), that, based on lexicalized synchronous context-free grammars, takes as input the whole source and post-edited sentences and automatically updates the models. The evaluation of our framework on two different domains (IT and medical) suggests that: *(i)* an SMT model enriched with the identified bilingual terms substantially improves translation quality in terms of BLEU score over a generic SMT system; *(ii)* strategies to integrate terminology need to take into consideration also the surrounding context of a translated term; *(iii)* in order to take advantage of the continuous appending of new information inside the SMT system a constant updating of the contribution of each component in the log-linear model is needed.

2 Bilingual Domain-Specific Terminology Generation

We propose a two-step approach to extract bilingual terminology for machine translation that requires only small amounts of parallel data (few hundred), as foreseen in a CAT scenario. The first step is the extraction of domain-specific terms from monolingual data (target and source sides of the parallel data), 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 unsupervised term extraction tools. For bilingual alignment, we compare different alignment strategies. The two steps are detailed in the following subsections.

2.1 Monolingual Terminology Extraction

In order to find the best performing approach to identify monolingual terms, we compare three available term extractors: the KX toolkit (Pianta and Tonelli, 2010), TWSC (Pinnis et al., 2012) and AlchemyAPI.[4] Given our experimental scenario, where no or little training data are available, we chose three unsupervised terminology extractors supporting different languages.

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

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 and Italian and was the first-ranked unsupervised system in the Semeval2010 task on keyword extraction (Kim et al., 2010). TWSC follows an approach which is very similar to KX, integrating morpho-syntactic patterns with statistical features. One of the main differences w.r.t. 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. A third system considered is AlchemyAPI. This commercial tool employs sophisticated statistical algorithms and linguistic approaches to analyse textual content and extract topic keywords, but no further implementation details are given.

2.2 Bilingual Terminology Alignment

Once the lists of monolingual terms for the source and target language are automatically gathered, the alignment across languages is created. We propose and compare different strategies.

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 et al., 2011), where the translation system the word aligner are trained on the same data from which the bilingual terminology is extracted. The main idea is that the translation system should know how to translate a source term, since it has seen it in the training data; this reduces the number of untranslated terms. Moreover, this allows us to take advantage of monolingual term extractors and regular phrase extraction method, used to build the phrase table, to generate bilingual terminology.

Given a set of possible translations for each term, the correct translation is retrieved 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 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.

In our experiments, we compare our strategies with Term Aligner, a state-of-the-art bilingual alignment tool, based on the method proposed by Aker et al. (2013). In this method, the authors treat bilingual term alignment as a classification problem. An SVM binary classifier is trained on data derived from the multilingual thesaurus EuroVoc, using language dependent and independent features. The former ones are based on bilingual dictionaries created by the GIZA++ tool, while the latter use cognate-based features, e.g. the longest common subsequence ratio. The cognate features are binarized using a manually defined threshold. Since the original work focuses on term alignment in comparable corpora, we limit the tool to search for terms that appear in the same parallel sentence pair. Moreover, we use the same GIZA++ dictionaries built for identifying term translation.

3 Enhancing Terminology Translation

After the extraction of domain-specific bilingual terms, they need to be integrated into the workflow of the SMT system. We focus on a real scenario, where a large translation project is split into partitions with around 3,000 tokens, 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$_1$ ... partition$_{n-1}$) using the XML markup or the cache-based models. To further improve the translation quality of partition$_n$, the decoder accesses the log-linear weights from the previous partition, which were tuned beforehand with MERT (Bertoldi et al., 2009).

Given the extracted terms and the parallel sentences, we improve the translation capability 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 working days.

3.1 Integration of Bilingual Terms into SMT

Since we place our work a CAT scenario, where an SMT system should continuously provide suggestions to the translator for each source sentence, we cannot integrate bilingual terms by retraining the whole model (Bouamor et al., 2012) or switching off the system and adding the terms at the end of the phrase table (Bouamor et al., 2011). Also the incremental training method introduced by Levenberg et al. (2010), which makes it possible to continuously add data without retraining the model, is not the best solution in our setting, because it tends to penalise terms with ambiguous translations favouring the most frequent and generic translations. For these reasons, we test two methods that can be easily used at run-time without altering the normal work of the SMT system and differentiate domain-specific from 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 specific spans of the source sentence. In 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.

Cache-Based Models In this work, we propose for the first time the use of the cache-based translation and language models (Bertoldi et al., 2013) for embedding bilingual terms into the 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. Differently from XML markup that 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 into the final translation.

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 an ageing of the existing phrase pairs and hence their re-scoring; in 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, whereby we test different rewarding and penalizing functions (hyperbola, power, exponential, cosine) as well as a constant function, where the 'age' is always set to 1. 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.

3.2 Incremental Tuning

The continuous extraction and collection of bilingual terms 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 ($partition_n$), 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 $partition_1$ 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 initial configuration of the log-linear weights used by MERT at time $n-1$ is that obtained optimizing 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 through XML tags for each incoming sentence, which required to extend Moses with this new option.

An issue with incremental tuning is the risk of over-fitting of the model on a small development set, when it differs from the test set. In our scenario, this is prevented by the fact that all the sets come from the same document, or from different documents on 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 the tuning.

4 Experimental Setting

In this Section, we propose a set of experiments aimed at showing 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 translation direction considered is from English to Italian. To identify the best monolingual term extraction tool as well as the most suitable bilingual alignment approach, we use freely available data, which were manually annotated to better evaluate all the intermediate steps of the experiment. Two datasets belonging to the IT domain, namely a portion of GNOME project data (4,3K tokens)[5] and KDE Data (9,5K),[6] are used for domain-specific term extraction.

The whole framework, including the machine translation part, 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 (Federico et al., 2014). Each corpus is split in partitions of around 3,000 tokens, i.e. the daily workload of a professional translator in post-editing, resulting in 6 partitions each.

For each translation task, we use the statistical translation toolkit Moses (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The IRSTLM toolkit (Federico et al., 2008) 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 et al., 2006), Europarl (Koehn, 2005) and OpenSubtitles2013 (Tiedemann, 2009), obtaining a training corpus of 37M tokens and a development set of $\sim$25K tokens. The generic SMT system used in all our experiments is trained on this merged general resource. The difference in size between the specific and the generic data is evident, i.e. approximately few thousands vs. more than 30 million tokens. For both domains, this reflects a real CAT scenario, where only a small quantity of domain-specific data is available.

Manual Terminology Annotation In order to evaluate the quality of the bilingual terms, we create a terminological gold standard for the IT domain. Two annotators with linguistic background were asked to mark all domain-specific terms in the monolingual GNOME and KDE corpora. 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. Then, the annotators had to manually create a bilingual pair if two domain-specific terms in a source and target sentence

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

	GNOME - KDE (English)			GNOME - KDE (Italian)		
	KX	AlchemyAPI	TWSC	KX	AlchemyAPI	TWSC
# of Terms	1115	665	496	950	304	765
Precision	0.293	0.393	0.413	0.271	0.309	0.362
Recall	0.596	0.571	0.372	0.452	0.167	0.481
F1	0.393	**0.466**	0.391	0.339	0.213	**0.412**

Table 1: Evaluation of monolingual term extraction for English and Italian

were found, one being the translation of the other. The average Cohen's Kappa on GNOME and KDE data computed at token level was 0.66 for English and 0.53 for Italian, which corresponds to a substantial and moderate agreement following Landis and Koch (1977). This annotation effort resulted in the identification of 874 domain-specific bilingual terms in the two datasets.[7]

5 Evaluation

In this Section, we report the quality of monolingual term extraction and the bilingual alignment. For each domain we evaluate the performance obtained by applying different approaches to the integration of bilingual terms into an SMT system. Evaluation of the extracted monolingual and bilingual terms is performed on the manually annotated KDE and GNOME datasets by calculating precision, recall and f-measure. The BLEU metric (Papineni et al., 2002) is used to automatically evaluate the translation quality of the EMEA and the IT manual datasets.

5.1 Monolingual Term Extraction

Our first evaluation concerns monolingual term extraction from English and Italian documents provided by the KX, AlchemyAPI and TWSC extraction tools.

As shown in Table 1, KX tends to overgenerate when extracting English terms. It extracts the highest number of expressions, which results in a high recall, but low precision. On the other hand, TWSC extracts the least English terms. Based on F1, we observe that AlchemyAPI is the best performing tool when extracting English terms. On Italian data, TWSC achieves the best F1 score, while AlchemyAPI performs worst due to the lack of Italian resources within the Linked Open Data (LOD) cloud.[8] KX shows a similar behaviour when extracting terms both from English and from Italian data, i.e. low precision and high recall. In summary, we select AlchemyAPI as the best performing term extractor for English and TWSC for Italian, to be used in the next phase.

5.2 Bilingual Term Alignment

In this step, we evaluate our strategies (i.e. Word Alignment and SMT n-best) to align monolingual terms and compare them against the performance of Term Aligner (see Section 2.2). We consider two different settings: in the first one, we use the two monolingual lists, which are automatically extracted by AlchemyAPI for English and TWSC for Italian. In the second one, instead, parallel terms are built starting from the monolingual terms, which were manually annotated to create the gold monolingual datasets.

Focusing on the translation projections, in the top part of the Table 2 (real situation with automatically extracted terms), we observe that the *term lookup* approaches, where the alignments are generated by *word alignment* and *SMT n-best* method, are too restrictive and output few bilingual terms, resulting in high precision but low recall. The *sentence lookup* strategies

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

[8] `http://linkeddata.org/`

Automat. Ext. Monol. Terms	Translation Projection				Term Aligner			
	Word Alignment		SMT n-best		cognate threshold			
	sent. lookup	term lookup	sent. lookup	term lookup	0.1	0.3	0.5	0.7
Precision	0.207	0.440	0.192	0.413	0.079	0.249	0.333	0.435
Recall	0.270	0.101	0.406	0.178	0.223	0.054	0.079	0.053
F1	0.233	0.164	**0.256**	0.246	0.116	0.085	0.128	0.094

Gold Standard	Translation Projection				Term Aligner				
	Word Alignment		SMT n-best		cognate threshold				
	sent. lookup	term lookup	sent. lookup	term lookup	0.2	0.4	0.6	0.8	1.0
Precision	0.463	0.768	0.426	0.779	0.498	0.782	0.916	0.949	0.970
Recall	0.399	0.285	0.577	0.517	0.573	0.458	0.402	0.389	0.315
F1	0.425	0.415	0.483	**0.616**	0.526	0.577	0.558	0.549	0.474

Table 2: Bilingual term alignment using the automatically extracted monolingual terms and the gold standard

are more tolerant, identifying more bilingual terms and having a better recall. In terms of F1, the *SMT n-best* strategies have better scores compared to *word alignment* methods. This is due to the possibility to select a correct target term from the n-best translations and not only from the single option generated by word alignment. As for the Term Aligner tool, we run experiments with different cognate similarity 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 term alignment is achieved with threshold of 0.5, and, in general, this method tends to align few bilingual terms but with high quality. Nevertheless, the alignment quality is substantially lower compared to the translation projection approaches. This can be deduced from the difference between the bilingual terms used to train the classifier and our test set.

When using monolingual terms provided by human annotators (bottom part of Table 2), we obtain significantly higher results compared to the real scenario described before. In this case, the SMT term-lookup method performs best. This implies that term lookup is more sensitive to the heterogeneity in automatically extracted data than the approach based on sentence lookup.

Term Aligner often obtains a precision close to 1, which is similar to the original results reported by Aker et al. (2013). This indicates that the method performs very good if it operates with high-quality data like our gold standard or the EuroVoc dataset. Nevertheless, it is sensitive to domain specificity and to the homogeneity of the terms to be aligned.

In summary, we compared several alignment approaches, i.e. Translation Projection with Word Alignment and SMT n-best method, both in combination with sentence and term lookup. Our evaluation included also Term Aligner with different thresholds. The SMT n-best approach always outperforms the others, whereby Term Aligner is negatively affected by heterogeneous data, showing the lowest performance with automatically extracted monolingual terms.

5.3 Translation Evaluation

After identifying the best tool for monolingual term extraction and the best approach for bilingual alignment, we carry out the final translation evaluation, based on the EMEA and IT manual datasets.

As described in Section 3, we split our data into several partitions and each of them is translated by: *(i)* a baseline SMT system that was built with the general resource, without embedding terminology; *(ii)* XML markup approach to embed the terminology paired with the

IT manual	Tuning	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Document level
Baseline	non-inc	23.21	36.75	25.16	20.76	23.50	19.39	25.58
Baseline	increm	23.21	35.61	23.31	24.10	25.47	20.40	26.00
XML markup	non-inc	23.21	36.61	25.88	25.67	25.54	22.39	27.40*
XML markup	increm	23.21	37.52	27.32	25.25	27.68	22.80	28.01*
Cache-based TM/LM	non-inc	23.21	35.85	26.71	27.51	28.58	25.26	28.66*
Cache-based TM/LM	increm	23.21	35.88	28.01	27.98	30.77	26.84	**29.46***
EMEA	Tuning	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Document level
Baseline	non-inc	22.19	22.03	24.80	26.50	21.66	23.80	23.53
Baseline	increm	22.19	22.09	24.99	27.04	21.10	24.47	23.69
XML markup	non-inc	22.19	24.72	24.09	24.51	21.54	23.87	23.51
XML markup	increm	22.19	22.58	23.45	26.71	22.25	26.54	24.09
Cache-based TM/LM	non-inc	22.19	23.17	27.09	28.95	25.97	26.71	25.73*
Cache-based TM/LM	increm	22.19	22.34	27.54	28.58	26.72	28.04	**25.96***

Table 3: Automatic evaluation (BLEU) based XML markup and cache-based approach (bold results = best performance ; * statistically significant compared to baseline)

baseline SMT system; *(iii)* cache-based model, where the 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 bilingual term is set according to the translation probability obtained by the SMT system used to project the source term onto the target language. Since a source term may have different translation candidates, the different translation probabilities give preference to more probable translations. Furthermore, XML markup cannot handle overlaps between dictionary entries. In our experiments, we found only 15 cases where the entries overlap, whereby we give preference to longer source terms.

For each set of partitions, the incremental tuning was run to update the log-linear weights. For a comparison, we also run MERT on each partition starting with flat weights (non-incremental tuning).

In Table 3, we report BLEU scores for each partition separately (columns "Part #"), as well as the evaluation on the whole corpus (column "Document level"). The approximate randomization approach Clark et al. (2011) is used to test whether differences among system performances are statistically significant at document level. Results in the table marked with * are statistically significantly better than the baseline with a p-value < 0.05.

Comparing the baseline XML markup and the cache-based methods, we notice that the translation performance of cache-based models always outperforms significantly all the other methods in both domains. This is also confirmed at partition level, with few exceptions for the initial partitions. The XML markup performs better than the baseline in both domains, but statistical significance is obtained only for the IT domain. Among different decay functions in the cache-based models, we report only the 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. To our surprise, the constant function did not outperform the reported decay function.

At document level, the incremental tuning always outperforms the results obtained starting MERT with flat weights. It is interesting to notice that the gap between the performance obtained by the incremental tuning and the standard approach generally increases partition after partition. This behaviour is more evident for the EMEA corpus, suggesting a more coherent distribution of sentences in the dataset. This favourable situation allows the incremental tuning

IT manual	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6
Available bilingual terms per partition	0	162	288	402	485	627
Coverage of bilingual terms in phrase table [%]	0	11.11	14.58	15.67	15.46	14.03
Coverage of source terms in source sentences [%]	0	0.80	7.86	19.79	17.64	20.21
Coverage of target terms in reference sentences [%]	0	77.78	91.75	92.79	82.90	75.80
EMEA	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6
Available bilingual terms per partition	0	259	402	540	662	761
Coverage of bilingual terms in phrase table [%]	0	11.58	11.19	11.30	10.12	9.99
Coverage of source terms in source sentences [%]	0	13.67	17.34	24.44	20.57	24.34
Coverage of target terms in reference sentences [%]	0	83.04	69.37	73.75	69.96	70.90

Table 4: Extracted and used bilingual terms for the IT and medical domain partitions

to better leverage the optimized weights of the previous partitions. Although the IT data show different levels of difficulty in the partitions (e.g. Partition 2 is easier to be translated than Partition 6), the incremental tuning is still able to smooth such differences and computes weights capable to produce better translations. The proposed framework has shown to be a valuable alternative to the well known XML markup method outperforming it in both domains.

Analysis of bilingual terms in test set To better understand the performance of our framework, Table 4 reports additional statistics related to *(i)* the number of extracted terms used by the SMT system to translate the current partition *(ii)* the number of bilingual terms covered by the baseline phrase table, *(iii)* the percentage of unique terms that have the source side in the source part of the test set and *(iv)* the percentage of terms that have the source side in the source part of the test set and the target side in the reference part of the test set.

As expected, the number of terms increases after each partition giving a larger contribution to the SMT system. Analysing the percentage of extracted terms covered by the phrase table, we noticed that on average around 14% of the terms in the IT domain are known by the baseline system, while only 10% are covered for the medical domain. On one hand, this explains the lower performance on average of the baseline system translating the medical domain in comparison to the IT domain. On the other hand, it does not motivate the larger improvements for the IT compared to the EMEA domain, because less terms can contribute to enhance translation quality. Larger improvements in IT are also not supported by the larger number of source terms covered in the medical domain (ninth versus fourth row in Table 4), which indicates that more EMEA bilingual terms are used to cover source spans than the IT domain.

These quantities do not consider two important aspects. The *first* one is the impact of the target terms on the reference sentences. In the IT, we are able to extract more terms that have the correct translation in the reference test set (on average 85% for IT with peaks larger than 90% against 70% for EMAE). The *second* aspect is the level of repetitiveness of terms in each document. To estimate it, we compute the repetition rate (Tiedemann, 2010) that measures how often n-grams are repeated in the whole document. Although it is not limited to domain-specific terms, since it includes all possible n-grams, we consider it a good approximation. Both documents are quite repetitive (32.49 for IT and 12.94 for EMEA), but in the IT corpus repetitiveness is more than twice as high than in the EMEA one. These two aspects suggest that the IT corpus contains more domain-specific terms and we are able to provide translations that better fit the references. This last aspect is crucial for the XML markup that, by definition, is more sensitive to the quality of bilingual terms than the cache-based approach. This is confirmed in the EMEA experiments where the XML markup is not able to significantly outperform the

baseline, while the cache-based approach can produce more than 2 BLEU points improvements.

Manual evaluation of translated sentences In order to investigate to what extent the approaches differ from a translator's point of view, we manually inspected the translations produced by the XML markup and cache-based approach. The quality of the two translation versions generally reflects the results reported in Table 3. The XML markup approach tends not to take into account the surrounding context of a translated string, while the cache-based one usually shows a better context-awareness. Specifically, it usually provides a better agreement between adjective and noun (which in Italian 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.*". 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 is highly specific and noun phrases are often composed by 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 reported in Table 3, showing that translation quality is generally lower than for the IT domain.

6 Cache-Based Model vs. Online Adaptation Model with cdec

To complete our evaluation, we compare the XML markup and the cache-based approach with the *Realtime Adaptive Translation Systems with cdec*,[9] (henceforth *Realtime cdec*) an online model adaptation system. Differently from the cache-based approach, it automatically extracts new translation rules from the whole source and post-edited sentences and adds them to the translation grammar. This system takes advantage of *cdec* (Dyer et al., 2010), a standalone decoder, aligner, and learning framework for SMT. *cdec* allows us to train word-based and phrase-based models, as well as models based on lexicalized synchronous content-free grammars (SCFG), which was used in our experiment. The adaptation of *cdec* to work in real time requires the use of *Fast Align* (Dyer et al., 2013) to perform on-the-fly word alignment between source and post-edited sentences. This makes possible the incremental addition of information to the translation models after a sentence is translated. Furthermore, *Realtime cdec* adapts the Bayesian language model using the hierarchical Pitman-Yor process approach, whereby MIRA (Chiang, 2012) is used to optimize the discriminative parameters of the decoder.

In our experiments we use the *Realtime cdec* similarly to the scenario described in Section 3.2. Each sentence pair (source, post-edition) from $partition_{n-1}$ is added to the model and used by MIRA to optimise the weights. The initial weights employed by MIRA for the first sentence pair at time $n-1$ are obtained after optimizing the system on the last sentence pair of $partition_{n-2}$. When all the sentence pairs from $partition_{n-1}$ are added, all the source sentences from $partition_n$ are translated. It is worth to notice that this setting favours the *Realtime cdec* compared to the cache-based or XML markup method, because it adds the whole sentence and not only the bilingual terms.[10]

[9] `http://www.cs.cmu.edu/~mdenkows/cdec-realtime.html`

[10] The cache-based model can also take advantage of non-terminological n-grams but it requires alignment between source and post-edited sentences, which is out of the scope of this paper.

IT manual	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Document level
Cache-based TM/LM	23.21	35.88	28.01	27.98	30.77	26.84	**29.46***
Realtime cdec	14.25	14.36	27.56	35.85	31.21	39.04	27.90*

Table 5: Comparison between the cache-based method with Moses and the *Realtime Adaptive Translation Systems with cdec*

Table 5 illustrates the performance translating the IT manuals using our proposed approach and the *Realtime cdec*. Although both systems were tuned on the same development set, we observe that cdec/MIRA needs more parallel data to adjust the pre-tuned parameters when translating a new domain.[11] Only after adding the sentence pairs from the first three partitions, *Realtime cdec* is able to outperform the cache-based approach and taking advantage of the content of the whole sentences from the next partitions to substantially improve over the cache-based model. On the contrary, the cache-based approach tuned with MERT shows to be less affected by the change of a domain and performs better with the first partitions at cost of lower performance in the last partitions resulting in a better BLEU score at document level.

Since *Realtime cdec* enhances its translation capability using the whole source and post-edited sentences, it is difficult to measure the impact of terminology. To overcome this problem, we evaluate only the correctness of the translated terms in the target sentences, and not the whole sentence itself. Therefore, we asked a linguist to manually evaluate the bilingual terms automatically extracted from the IT manuals. 403 out of 627 were marked as correct translation in the domain and we used them to check if our approach and the *Realtime cdec* are able to correctly translate these terms. On the whole document, we counted 1,538 occurrences of extracted terms in the target sentences generated by cdec. Although the cache-based model is not using the alignment information from the source and post-edited sentence, we counted 1,495 occurrences of terms in target sentences. For both methods, around 90% of these occurrences are correct translations in the references. Moreover, we measure an overlap of 85% of bilingual terms (appearing in source, target and reference sentences) between the cache-based method and *Realtime cdec*. These results show that both methods, *Realtime cdec* with the alignment information and our proposed framework embedding the extracted terminology, are able to correctly manage the translations of the domain-specific vocabulary.

7 Related Work

Our work is based on a framework that includes the monolingual extraction of domain-specific terms from a small parallel corpus, bilingual term alignment, and the integration of the bilingual terminology into an SMT system. In the past years, a number of techniques have been applied to the task of bilingual multi-word extraction from parallel or comparable corpora. Most of the work (Daille et al., 1994; Wu and Chang, 2003; Vintar and Fišer, 2008; Kim et al., 2009) focuses on identifying monolingual candidates using linguistic knowledge, statistical methods, or a combination of the two.

As for the bilingual alignment of terms, 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. Our approach, however, shows a better performance due to the domain specificity of our dataset. 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 en-

[11] We performed experiments with different C-values of 0.1, 0.01 (default), 0.001, 0.0001, whereby we obtained best results using a C-value of 0.0001 for initial parameter tuning and 0.01 in the learning approach during translation.

tire corpus. Furthermore, their embedding methods focus on concatenating the newly obtained bilingual data to 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, e.g. bilingual terminology. Thurmair and Aleksić (2012) extract terms and lexicon entries from SMT phrase tables. In their approach they apply linguistic, lexicon and frequency filters to obtain good lexicon entries. Similarly, we also access the phrase table to build our bilingual terminology, whereby our filter relies on the term and sentence lookup approach.

Furthermore, there has been research done on the integration of domain-specific parallel data into SMT, e.g. dictionaries or bilingual terminology, either by retraining new and general parallel resources or adding new entries to the phrase table (Langlais, 2002; Ren et al., 2009; Haddow and Koehn, 2012; Pinnis et al., 2012). Furthermore, Okita and Way (2010) investigate the effect of integrating bilingual terminology in the training step of an SMT system, and analyse in particular the performance of a word aligner sensitive to multi-word expressions and translation smoothing. As opposed to their approach, we do not have prior knowledge about the bilingual terminology, since we extract it on the fly based on the document to be translated. 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. Since the manual development of terminological resources is a time intensive and expensive task, our framework continuously builds bilingual terminology knowledge from the already translated sentences. In order to tackle term translation and the out-of-vocabulary issues, Arcan et al. (2012) used the multilingual web to built a parallel domain-specific corpus based on the vocabulary to be translated. Additionally, Arcan et al. (2014) extend their work focusing on disambiguated term extraction using the rich lexical and semantic knowledge of Wikipedia.

8 Conclusion

In this paper, we propose a framework to enhance translation quality by exploiting bilingual terms extracted from the parallel sentences daily produced by professional translators. The results show that an SMT model enriched with the identified bilingual terms substantially improves translation quality in terms of BLEU score over a generic baseline system. Furthermore, we investigate the integration of the extracted bilingual terms into the SMT system. For the first time we report on the usage of the cache-based model in the context of terminology embedding, whereby we compare results with the widely-used XML markup. The ability of the cache-based model to take into consideration the surrounding context of a translated term allows it to outperform the XML markup approach. In addition, we report a better performance in bilingual term alignment compared to the state-of-the-art Term Aligner.

In the future, we plan to integrate the proposed framework into a professional post-editing environment, measuring the translators' productivity gain using automatically extracted terminology. Furthermore we plan to combine the strengths of the cache-based model treating a term as one translation unit and the *Realtime cdec* approach of embedding the incrementally extracted bilingual knowledge from the whole sentence into the translation system.

Acknowledgments

We would like to thank Dr. Ahmet Aker and Mārcis Pinnis for providing us with their newest software and the technical support for it. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and by the European Union supported project MateCat (ICT-2011.4.2-287688).

References

Aker, A., Paramita, M., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of ACL*, Sofia, Bulgaria.

Arcan, M., Federmann, C., and Buitelaar, P. (2012). Experiments with term translation. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 67–82, Mumbai, India. The COLING 2012 Organizing Committee.

Arcan, M., Giuliano, C., Turchi, M., and Buitelaar, P. (2014). Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, Dublin, Ireland.

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of Machine Translation Summit XIV*, Nice, France.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189. Association for Computational Linguistics.

Bertoldi, N., Haddow, B., and Fouet, J.-B. (2009). Improved minimum error rate training in Moses. *Prague Bull. Math. Linguistics*, 91:7–16.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2011). Improved statistical machine translation using multiword expressions. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011)*, pages 15–20.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chiang, D. (2012). Hope and fear for discriminative training of statistical translation models. volume 13.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability . In *Proceedings of the Association for Computational Lingustics*.

Daille, B., Gaussier, É., and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *COLING*, pages 515–524.

Denkowski, M., Dyer, C., and Lavie, A. (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*, pages 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *HLT-NAACL*, pages 644–648. The Association for Computational Linguistics.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*.

Ehrmann, M., Turchi, M., and Steinberger, R. (2011). Building a multilingual named entity-annotated corpus using annotation projection. In *RANLP*, pages 118–124.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Christian, B., and Germann, U. (2014). The MateCat Tool. In *Proceedings of the 25th International Conference on Computational Linguistics - Demo Session*, Dublin, Ireland.

Federico, M., Cattelan, A., and Trombetti, M. (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 (AMTA)*.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Haddow, B. and Koehn, P. (2012). Analysing the Effect of Out-of-Domain Data on SMT Systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada. Association for Computational Linguistics.

Itagaki, M. and Aikawa, T. (2008). Post-mt term swapper: Supplementing a statistical machine translation system with a user dictionary. In *LREC*. European Language Resources Association.

Kim, S. N., Baldwin, T., and Kan, M.-Y. (2009). An unsupervised approach to domain-specific term extraction. In *Australasian Language Technology Workshop*, pages 94–98, Sydney.

Kim, S. N., Medelyan, O., Kan, M.-Y., and Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 21–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Landis, J. and Koch, G. (1977). Measurement of Observer Agreement for Categorical Data. volume 33, pages 159–174.

Langlais, P. (2002). Improving a general-purpose statistical translation engine by terminological lexicons. In *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM) '2002, Taipei, Taiwan*, pages 1–7.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., and Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. *Machine Translation Summit XIV*, page 83.

Levenberg, A., Callison-Burch, C., and Osborne, M. (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, pages 394–402, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Pianta, E. and Tonelli, S. (2010). KX: A flexible system for Keyphrase eXtraction. In *Proceedings of SemEval 2010, Task 5: Keyword extraction from Scientific Articles*, Uppsala, Sweden.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*.

Thurmair, G. and Aleksić, V. (2012). Creating term and lexicon entries from phrase tables. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, EAMT 2012.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 8–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vintar, S. and Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *LREC*. European Language Resources Association.

Wu, C.-C. and Chang, J. S. (2003). Bilingual collocation extraction based on syntactic and statistical analyses. In *ROCLING*. Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taiwan.