Translation project adaptation for MT-enhanced computer assisted translation

Mauro Cettolo · Nicola Bertoldi · Marcello Federico · Holger Schwenk · Loïc Barrault · Christophe Servan

Received: 31 January 2014 / Accepted: 12 August 2014 / Published online: 26 August 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The effective integration of MT technology into computer-assisted translation tools is a challenging topic both for academic research and the translation industry. In particular, professional translators consider the ability of MT systems to adapt to the feedback provided by them to be crucial. In this paper, we propose an adaptation scheme to tune a statistical MT system to a translation project using small amounts of post-edited texts, like those generated by a single user in even just one day of work. The same scheme can be applied on a larger scale in order to focus general purpose models towards the specific domain of interest. We assess our method on two domains, namely information technology and legal, and four translation directions, from English to French, Italian, Spanish and German. The main outcome is that our adaptation

M. Cettolo (✉) · N. Bertoldi · M. Federico
FBK, Fondazione Bruno Kessler, 38123 Povo, Trento, Italy
e-mail: cettolo@fbk.eu

H. Schwenk · L. Barrault · C. Servan
LIUM, University of Le Mans, 72085 Le Mans cedex 9, France
e-mail: holger.schwenk@lium.univ-lemans.fr

L. Barrault
e-mail: loic.barrault@lium.univ-lemans.fr

C. Servan
e-mail: christophe.servan@xrece.xerox.com

Present address:
C. Servan
Xerox Research Centre Europe, 38240 Meylan, France
strategy can be very effective provided that the seed data used for adaptation is ‘close enough’ to the remaining text to be translated; otherwise, MT quality neither improves nor worsens, thus showing the robustness of our method.

**Keywords**  
Statistical machine translation · Self-tuning MT · Domain adaptation · Project adaptation · Computer-assisted translation

1 Introduction

Despite continued significant progress, machine translation (MT) is generally not yet able to provide output that is suitable for publication without human intervention. However, it has been reported several times that post-editing MT output can significantly increase the productivity of human translators (Guerberof 2009; Plitt and Masselot 2010; Federico et al. 2012; Green et al. 2013; Läubli et al. 2013). This application of MT becomes even more effective the better the translation system is integrated with the human translation workflow. In particular, improvements can be obtained by carefully designing the user interface of the translator’s workbench\(^1\) but also by specializing the MT system on the texts the translator is working on. It is well known that a system optimized on the text genre it is to be used to translate performs better than a generic system. This process is usually called *domain adaptation*. In addition, it is reasonable to expect that an MT system should not be static, but should adapt itself over time. From the viewpoint of professional translators, refinements of the MT system in response to their corrections is indeed perceived as crucial in order to improve the usability of MT and to further increase productivity and quality of post-editing. Simply stated, while it is acceptable that MT makes mistakes, it is less acceptable that MT does not learn from user corrections and continues to make the same errors over and over again.

In the translation industry, a typical scenario is where one or more translators works for several days on a given translation project, i.e. a set of homogeneous documents. After one workday, knowledge about the newly translated text and user corrections could be injected into the MT system so that improved translations may be generated the next day. We call this process *project adaptation*. Project adaptation can be repeated daily until the end of the translation project. The corrections performed by the professional translators over a day are in fact a very valuable resource to improve the MT system.

While several works have addressed post-editing to enhance human translation, few works have considered how post-editing can improve MT. This paper presents recent results from the European MateCat project,\(^2\) which is developing a Web-based CAT tool for professional translators that integrates self-tuning MT, i.e. MT with domain- and project-adaptation functionality. We report here the results of our approaches to

---

\(^1\) In computer-assisted translation (CAT), translators work with special text editors, simply called CAT tools, integrating several translation aids, such as translation memories, terminology dictionaries, spell checkers, concordancers, and recently also MT engines.

\(^2\) [http://www.matecat.com](http://www.matecat.com).
domain and project adaptation for four language pairs, English into Italian, French, Spanish and German, and two domains: information technology (IT) and legal documents. Both domains represent relevant sectors in the translation industry and are suitable for exploiting statistical MT (SMT), since the information source is sufficiently homogeneous, the language is sufficiently complex, and there is enough multilingual data available to train and tune MT systems.

MT adaptation methods should be ideally evaluated by measuring productivity gains of human translators working on real translation projects. Hence, we run actual field tests in which professional translators post-edited outputs of a project-adapted MT system. Project data to perform project adaptation was collected from a different portion of the same document during a preceding warm-up session, in which the same translator post-edited outputs of a domain-adapted system.

Obviously, field tests are expensive and time-consuming to organise and cannot be conducted frequently to evaluate and compare many variants of adaptation algorithms. Therefore, we also report here results from our so-called lab tests, in which we simulate human post-edits with manual reference translations. It is important to note that these reference translations were created only once and independently of our MT systems.

In the legal domain, remarkable improvements in terms of automatic MT metrics were observed in the lab test experiments. These results were also confirmed by the field tests, where significant productivity gains were measured. In particular, the speed of translators increased on average by 22.2%, while the post-editing effort improved by 10.7%.

Lab test results on the IT domain were more controversial, due to the mismatch between the adaptation text and the actual test document. Nevertheless, in such a critical set-up our adaptation method proved to be conservative as no degradation of the baseline reference quality was observed.

The remainder of the paper is organized as follows. Section 2 discusses related works. Project adaptation methods are outlined in Sect. 3. In Sect. 4, the data used in the experiments is introduced, while lab and field tests are described and commented in Sects. 5 and 6, respectively. The paper concludes in Sect. 7 with a discussion on the contribution of this work, together with avenues for further investigation.

2 Related work

The idea that MT can boost the productivity of human translators has been consolidating in recent years thanks (i) to the remarkable progress achieved by SMT, and (ii) to several experimental investigations that have systematically evaluated the impact of MT on human translation. For instance, in Guerberof (2009), eight professional translators were asked to translate a fixed number of segments from English into Spanish: one third from scratch, one third from translation memory (TM) matches and one third from MT suggestions. TM matches were selected to be in the 80–90% fuzzy match range. A commercial SMT engine was trained on the content of the TM plus a core glossary. The translators used a web-based post-editing tool, supplied with the core glossary, to translate/post-edit all segments but without knowing their origin. Besides measuring and comparing productivity in terms of processing speed (words per sec-
ond), a detailed analysis was reported of the quality of the produced translations. The findings suggest that translators can achieve higher productivity and quality when post-editing MT output rather than fuzzy matches from the TM.

In Plitt and Masselot (2010), twelve professional translators were involved in an experiment which compared the productivity of human translation versus post-editing MT output. The test was performed on IT documentation, for four translation directions, and with three translators per direction. The MT engine was a specifically trained Moses engine (Koehn et al. 2007), while the post-editing tool was inspired by the CAITRA tool. Post-editing productivity was measured in terms of processing speed and edit distance. A pause analysis was carried out to compare keyboard and pause times of translation versus post-editing. Finally, a blind test was conducted to compare the quality of the segments produced with the two modalities. The most interesting outcome from our perspective is that MT allowed all translators to work faster, on average by 74%, but with a high variance across them: in fact, throughput improvements ranged from 20 to 131%.

In Federico et al. (2012), productivity was evaluated with a popular commercial CAT tool which seamlessly integrated MT suggestions within TM matches. Twelve professional translators were asked to translate full IT or legal documents rather than isolated segments, without changing their working routine, from English into Italian or German. Indeed, MT suggestions were provided in addition to TM suggestions and translators were left free to decide whether to translate segments from scratch or to post-edit the provided matches. The origin of each suggestion, TM or MT, was shown to the user with the aim of collecting more realistic figures about the potential benefits of enhancing CAT with MT. Productivity gains were observed for all translators when MT suggestions were supplied, which for ten users were statistically significant. Similar yet more thorough experiments were recently conducted by Läubli et al. (2013), which also reported statistical significant productivity gains by enhancing CAT with MT, and Green et al. (2013), whose investigation was run under more controlled conditions and reported statistically significant improvements both on productivity and overall translation quality.

From the point of view of MT methods, our work deals with MT adaptation in general, and incremental adaptation more specifically.

Bertoldi et al. (2012) presented an adaptation scenario where foreground translation and reordering models and language model of a phrase-based SMT system are incrementally trained on batches of fresh data and then paired to static background models. Similarly, the use of local and global models for incremental learning was previously proposed through a log-linear combination (Koehn and Schroeder 2007), a mixture model (linear or log-linear) (Foster and Kuhn 2007), the phrase and reordering table fill-up (Bisazza et al. 2011), or via ultraconservative updating (Liu et al. 2012).

Bach et al. (2009) investigated how a speech-to-speech translation system can adapt day-to-day from collected data on day one to improve performance on day two, in a similar way to our work. However, the adaptation of the MT module involved only the language model and was performed on the MT output.

---

http://www.caitra.org.
In an interactive setting, Hardt and Elming (2010) proposed to incrementally update SMT models by retraining a small local phrase table each time a new translation was available. Approximate alignments were obtained by means of a modified version of GIZA++ (Och and Ney 2003). Apart from the difference in the way the generic and the local phrase tables are combined—they put both them in the log-linear model while we merge them, as will be shown later—the main difference between their approach and ours regards the knowledge injected into the local table: that of single sentences in their case, and project- or even domain-related knowledge derived from larger corpora in ours.

Niehues and Waibel (2012) compared different approaches to adapting an MT system towards a target domain using small amounts of parallel in-domain data, namely the back-off, the factored, and the already mentioned log-linear and fill-up techniques. The general outcome is that each of them is effective in improving non-adapted models but none is definitely better than the others. The best performing algorithm depends on how well the test data matches the in-domain training data.

Hasler et al. (2012) focused on enhancing standard phrase-based SMT systems with word- and phrase-pair sparse features in order to bias models for the vocabulary and style of the target domain, namely TED talks. The work explores and compares several approaches for tuning sparse features on top of both small in-domain and larger mixed-domain systems: MERT (Och 2003), MIRA (Crammer et al. 2006), Jackknife (Quenouille 1956) and a retuning scheme for exploiting in-domain tuning also for mixed-domain models. Experiments were performed in the setup defined for the IWSLT 2012 shared task on two language pairs, English-to-French and German-to-English, and showed BLEU score (Papineni et al. 2002) improvements for both.

Our work deals with data selection as well, which is a problem widely investigated by the MT community, see for example Yasuda et al. (2008), Matsoukas et al. (2009), Foster et al. (2010) and Axelrod et al. (2011). We apply a standard selection technique (Moore and Lewis 2010), but in a quite different scenario where the task-specific data is extremely small and the generic corpus is actually close to the domain of the task.

3 Adaptation methods

In this section we describe the techniques applied to adapt our SMT systems, namely data selection and translation, distortion and language model combination.

3.1 Data selection

It has been believed for a long time that just adding more training data always improves performance of a statistical model, e.g. an \( n \)-gram LM. However, this is in general only true if the new data is relevant enough to the task at hand, a condition which is rarely satisfied. The typical case is that of a narrow domain, for which a small task-specific text sample can be more valuable than a very large generic text corpus, coming from sources that may be heterogeneous with respect to size, quality, domain, genre, production period, etc.
The main idea of data selection is to nevertheless try to take advantage of the generic corpus, by extracting a subset of training data that is mostly relevant to the task of interest, which in our case is a specific domain or a translation project.

In our setting, data selection is used twice: first, to adapt a generic system to a specific domain, i.e. legal or IT, before the human translator starts working; and second, to integrate it with the daily translations made by the users (project adaptation). In both cases, we apply the algorithm proposed by Moore and Lewis (2010), extended by Axelrod et al. (2011) to work on bitexts and implemented in the public tool XenC (Rousseau 2013). It assumes the availability of a seed corpus, which in our case is representative of the specific domain or project, and of a large generic corpus, from where to extract task-relevant sentences. The XenC tool provides three modes to perform data selection:

- **mode 1**: a simple filtering process based on direct perplexity computation (Gao and Zhang 2002);
- **mode 2**: monolingual cross-entropy difference (Moore and Lewis 2010);
- **mode 3**: bilingual cross-entropy difference (Axelrod et al. 2011).

We used modes 2 and 3. In mode 2, each word sequence $s$ is scored with the difference of the cross-entropy computed on two LMs. The first LM is estimated from the whole task-specific corpus, while the second LM is estimated from a random subset of the generic corpus, with a number of tokens similar to the specific one. In language modeling, cross-entropy is defined as in (1):

$$H_{LM}(s) = -\frac{1}{N} \sum_{i=1}^{N} \log P_{LM}(w_i | w_1 \cdots w_{i-1})$$  \hspace{1cm} (1)$$

where $P_{LM}(x)$ is the probability assigned to the event $x$ by the language model LM, $s = w_1 \cdots w_N$ is a word sequence and $w_1 \cdots w_{i-1}$ represents the history of the word $w_i$.

Formally, let $FG$ (foreground) be our task-specific corpus and $BG$ (background) our generic one. $H_{FG}(s)$ will be the cross-entropy of a sentence $s$ of $BG$ given by the LM estimated from $FG$, while $H_{BG}(s)$ will be the cross-entropy of the sentence $s$ of $BG$ given by the LM estimated from the subset of $BG$. The sentences $s$ from the generic corpus $BG$ will then be evaluated by $H_{FG}(s) - H_{BG}(s)$ and sorted by their score.

Mode 3 is based on the bilingual cross-entropy difference. Formally, let $FG_S$ and $FG_T$ be our task-specific corpus in the source $S$ and target $T$ languages, and $BG_S$ and $BG_T$ the generic corpus for the same language pair. For each language, first the monolingual cross-entropy difference is computed as described in the preceding paragraph. The final score will then be the sum of the two cross-entropy differences, as in (2):

$$[H_{FG_S}(s_S) - H_{BG_S}(s_S)] + [H_{FG_T}(s_T) - H_{BG_T}(s_T)]$$  \hspace{1cm} (2)$$

where $s_S$ is a word sequence from the generic corpus in the source language and $s_T$ is the corresponding word sequence in the target language.
Once the generic corpus is sorted, the optimal percentage of data to select has to be determined. The estimation is performed by minimizing the perplexity of a development set on language models trained on increasing amounts of the sorted corpus (10, 20, ..., 100%). Moore and Lewis (2010) reported that the perplexity decreases when less but more appropriate data is used, typically reaching a minimum at around 20% of the sorted generic data. As a side effect, the models become considerably smaller, which is also an important aspect when deploying MT systems in real applications. In our experiments, either we selected a fixed amount of about 10-15% of the sorted generic corpus, or set the percentage to the value which minimized the perplexity of a development set.

The choice of the seed corpus depends on the adaptation task. For domain adaptation, a domain-specific corpus was employed. For project adaptation several options are available (Cettolo et al. 2013): after one day of work on a project, we can use the portion of the source text processed and the corresponding human translation—produced either by post-editing automatic translations or by translating from scratch—to perform data selection on a bilingual seed. In contrast, the source text of the whole translation project is usually available at the beginning of the process; then, it could be used to select project-specific data by a source-side-only but larger seed. Not surprisingly, preliminary investigations suggested to use the largest possible seed by including the entire source document and the portion post-edited by translators in the period before the currently processed day.

3.2 Adaptation of SMT models

3.2.1 Translation and distortion models

Data selection is a very effective method to adapt the translation model on the most relevant data. However, by discarding some of the available resources, we take the risk to miss some translations which are not present in the selected data. To avoid that risk, we adapt the translation model with the fill-up technique, initially proposed in Nakov (2008) and then refined in Bisazza et al. (2011). In practice, the fill-up technique merges the generic background phrase table with the specific foreground phrase table by adding only phrase pairs that do not appear in the foreground table.

Formally, the translation model assigns a feature vector $\phi(\tilde{f}, \tilde{e})$ to each phrase pair, where $\tilde{f}$ and $\tilde{e}$ are respectively the source and target phrases. Let $T_{FG}$ and $T_{BG}$ be the foreground and the background phrase tables, respectively. The filled-up model $T_{fillUp}$ is defined as in (3):

\[
\forall (\tilde{f}, \tilde{e}) \in T_{FG} \cup T_{BG}: \quad \phi_{fillUp}(\tilde{f}, \tilde{e}) = \begin{cases} 
(\phi_{FG}(\tilde{f}, \tilde{e}), \exp(0)) & \text{if } (\tilde{f}, \tilde{e}) \in T_{FG} \\
(\phi_{BG}(\tilde{f}, \tilde{e}), \exp(1)) & \text{otherwise}
\end{cases}
\]

(3)

The entries of the filled-up model correspond to the union of the two phrase tables, while the scores are taken from the more reliable source whenever possible. To keep
track of a phrase pair’s provenance, a binary feature is added\(^4\) that fires if the phrase pair comes from the background table. The weight assigned to this feature acts as a scaling factor for the scores from the generic phrase table. It is worth noting that the phrase-table fill-up technique can be seamlessly applied to the lexicalized reordering model (Koehn et al. 2005; Galley and Manning 2008) we used in our experiments, with the only difference that no additional feature is introduced.

We chose the fill-up technique because it performs as well as other popular adaptation techniques (Niehues and Waibel 2012), but generates models that are more compact and easier to tune. Actually, we applied an even more simplified version of the fill-up method, called back-off, in which the indicator feature is omitted in the filled-up phrase table as well: this means that no additional weight (that of the binary feature) has to be estimated when the combined phrase table is used, while in practice no significant performance variation is observed (Cettolo et al. 2013).

3.2.2 Language model

As concerns the language model adaptation, we employed the mixture of models proposed by Kneser and Steinbiss (1993), which consists of the convex combination of two or more language models. Formally, let \( P_k(w_i \mid w_{i-n+1} \cdots w_{i-1}) \) be the conditional probability of observing the word \( w_i \) after the words \( w_{i-n+1} \cdots w_{i-1} \) for the \( k^{th} \) \( n \)-gram LM. Each of the \( K \) LMs contributes to the mixture according to (4):

\[
P_{\text{mixLMs}}(w_i \mid w_{i-n+1} \cdots w_{i-1}) = \sum_{k=1}^{K} \lambda_k P_k(w_i \mid w_{i-n+1} \cdots w_{i-1})
\]

where \( 0 \leq \lambda_k \leq 1 \) and \( \sum_{k=1}^{K} \lambda_k = 1 \); \( \lambda_k \)'s can be interpreted as the probability that \( w_i \) is generated by the \( k^{th} \) LM. \( \lambda \)'s are computed via maximum-likelihood estimation. In our setting, we interpolate the foreground and the background language models.

4 Data

Two domains and four language pairs, for a total of six different tasks, are involved in our experiments: translation from English into Italian and French for the IT domain, and from English into Italian, French, Spanish and German for the legal domain. In the following two sections, training and evaluation data prepared for each task are described.

4.1 Training data

For training purposes we relied on several language resources, including parallel corpora and translation memories. For the IT domain, software manuals from the OPUS
corpus (Tiedemann 2012), namely KDE4, KDE4-GB, KDEdoc, and PHP were used, which are all publicly available. In addition, a proprietary large TM was employed. It mostly consists of real projects on software documentation commissioned by a specific customer.

For the legal domain, the publicly available JRC-Acquis collection (Steinberger et al. 2006) was used, which mostly includes EU legislative texts translated into 22 languages.

Table 1 provides detailed statistics on the actual bitexts used for training translation and reordering models; language models were estimated on the target side. The in-domain train entries refer to the entire generically in-domain training texts, while the project selection entries refer to the subset of in-domain train data that was selected for project adaptation to the specific document to translate.

The domain selection entry of the IT en→fr task refers to data selected from out-of-domain texts (Giga English–French, United Nation, and Common Crawl corpora5 (Bojar et al. 2013)) by using the in-domain text as seed; this was done to augment the amount of training data, since the size of in-domain text available for that

---

5 Available from http://www.statmt.org/wmt13/translation-task.html.
language pair (15.4/17.9 million words) is about four times smaller than for the other tasks.

Development sets are additional corpora on which the parameters of the phrase-based MT model were optimized. They are collections of excerpts from real IT/legal translation projects commissioned to the MateCat industrial partner Translated.S.r.l. with no overlap with training or evaluation sets.

4.2 Evaluation data

For the IT domain, data was supplied by Translated S.r.l. An already executed translation project from English was selected for which translations into Italian and French were available. As translations in the two languages were carried out with different CAT tools, some manual pre-processing was necessary to normalize the text segmentations of the documents across the two translation directions. Moreover, the texts were cleaned to remove formatting tags and software code excerpts, which are not relevant for our field test. Finally, a single source document of 1,956 segments and about 17,800 source words was created, and split into two portions: one for the warm-up session (342 segments), and one for the actual field-test session (1,614 segments).

For the legal domain a document was taken from the European Union law, for which translations into the four languages of interest were available. The document was pre-processed as well so that the segments of the four versions were all aligned. The full document consists of 605 segments and 13,900 words, and was split into two portions: one for the warm-up session (133 segments) and one for the actual field-test session (472 segments).

Table 2 provides some statistics on the texts to be translated during the warm-up session and the proper field-test session. The target word counts refer to the human references. Note that for each domain, the document to translate is shared among all language-pairs. The small difference between warm-up legal texts is due to a few segments being unavailable for all languages.

5 Lab test

5.1 MT systems

The SMT systems were built using the open-source MT toolkit Moses (Koehn et al. 2007). The translation and lexicalized reordering models were trained on parallel training data (Table 1). Back-off $n$-gram language models ($n = 5$ unless otherwise specified) were built on the target side of the training bitexts using improved Kneser-Ney smoothing (Chen and Goodman 1999). The standard MERT procedure provided within the Moses toolkit was used to optimize the weights of the log-linear interpola-

---

6 2013/488/EU: “Council Decision of 23 September 2013 on the security rules for protecting EU classified information”.

7 [http://eur-lex.europa.eu/](http://eur-lex.europa.eu/).
Table 2  Overall statistics on parallel data used for evaluation purposes: number of segments and running words of source and target sides

| Domain | Pair    | Test set  | Segments | Source | Target |
|--------|---------|-----------|----------|--------|--------|
| IT     | en→it   | Warm-up   | 342      | 3,435  | 3,583  |
|        |         | Field-test| 1,614    | 14,388 | 14,837 |
|        | en→fr   | Warm-up   | 342      | 3,435  | 3,902  |
|        |         | Field-test| 1,614    | 14,388 | 15,860 |
| Legal  | en→it   | Warm-up   | 133      | 3,082  | 3,346  |
|        |         | Field-test| 472      | 10,822 | 11,508 |
|        | en→fr   | Warm-up   | 134      | 3,084  | 3,695  |
|        |         | Field-test| 472      | 10,822 | 12,810 |
|        | en→es   | Warm-up   | 131      | 3,007  | 3,574  |
|        |         | field-test| 472      | 10,822 | 12,699 |
|        | en→de   | Warm-up   | 133      | 3,082  | 3,125  |
|        |         | Field-test| 472      | 10,822 | 10,963 |

For each task, two different SMT engines were tested: the reference in-domain (IN) system, and the corresponding project-adapted (PA\textsubscript{IN}) system.

The models of IN engines were estimated on domain-specific training data, i.e. entries in-domain train (plus domain selection for the IT English-to-French task) in Table 1.

Project-specific data (entries project selection of Table 1) were selected from the in-domain training corpora by means of the data selection method described in Sect. 3, using as a seed corpus the source/target sides of the document to be translated during the warm-up session and the source side of the document to be translated during the field-test session. Project-specific models were trained on the concatenation of texts selected this way in addition to warm-up documents. The models of PA\textsubscript{IN} engines are the combination of project-specific and in-domain models, via the back-off and LM mixture methods of Sect. 3.

5.2 Results

Table 3 provides BLEU, TER (Snover et al. 2006) and GTM (Turian et al. 2003) scores computed on the field test documents with respect to human references of the in-domain and project-adapted systems for each of the six translation tasks.

For the legal domain, the improvements over the reference system show the adaptation technique to be quite effective. As an example, we see a relative improvement in BLEU score of 12.9 % (31.0–35.0) when translating into Italian, of 7.4 % (33.9–36.4) into French, of 2.5 % (35.5–36.4) into Spanish, and of 7.7 % (18.3–19.7) into German.
Table 3 Overall performance of MT engines with respect to human references on the documents of the proper field-test session

| Pair  | MT engine | IT domain | Legal domain |
|-------|-----------|-----------|--------------|
|       |           | BLEU | TER | GTM | BLEU | TER | GTM |
| en→it | IN        | 55.3 | 29.2 | 77.8 | 31.0 | 53.1 | 61.8 |
|       | PA<sub>TN</sub> | 57.5 | 26.3 | 78.6 | 35.0 | 49.1 | 64.6 |
| en→fr | IN        | 41.3 | 38.3 | 69.5 | 33.9 | 52.2 | 63.0 |
|       | PA<sub>TN</sub> | 41.4 | 37.9 | 69.9 | 36.4 | 49.1 | 65.1 |
| en→es | IN        | –    | –    | –    | 35.5 | 50.7 | 65.7 |
|       | PA<sub>TN</sub> | –    | –    | –    | 36.4 | 50.2 | 65.6 |
| en→de | IN        | –    | –    | –    | 18.3 | 68.4 | 50.5 |
|       | PA<sub>TN</sub> | –    | –    | –    | 19.7 | 66.6 | 52.3 |

Different from the legal domain, in the IT domain the only significant gain is observed for the Italian direction, where the BLEU score increases by 4% relative (55.3–57.5). Improvements on the French task are negligible.

As the methods perform quite differently across domains and language pairs, a deeper analysis was conducted, which we report in Sect. 5.5.

5.3 Domain adaptation: the legal English-to-German task as a case study

It is known that translation into German is a rather hard task, posing special challenges in terms of word order and compounding of words. Our experimental results on the legal domain confirm this: the translation of the same English document by means of companion SMT systems trained on approximately the same amount of text from the same source yield a BLEU score for German which is up to 40-50% worse than for Italian, French and Spanish. Hence, we chose this challenging task to validate a variant of the adaptation scheme proposed so far, which relies on a preliminary selection of domain-specific data from in- and out-of-domain corpora, as described in the following sections.

5.3.1 Training data

For training English-to-German models tailored to the legal domain, domain-specific data was selected from various generic linguistic resources, namely Europarl (Koehn 2005), JRC-Acquis and proprietary TMs. Table 4 provides detailed statistics on the actual bilingual texts used for training and development purposes.

5.3.2 MT system

As stated above, the adaptation scheme used here for the English-to-German task differs slightly from that followed in the development of IN and PA<sub>TN</sub> systems in
Table 4 Overall statistics on data used for training and development (tuning) of the domain-adapted en→de engines: number of segments and running words of source and target sides

| Domain       | Pair       | Corpus  | Segments | Tokens Source | Tokens Target |
|--------------|------------|---------|----------|---------------|---------------|
| Legal        | en→de      | Train   | 5.8M     | 140.3M        | 131.4M        |
|              |            | Domain selection | 1.9M  | 49.3M        | 45.4M        |
|              |            | Project selection | 2.3M  | 62.2M        | 57.2M        |
|              |            | Development set   | 925   | 35,270       | 32,277       |

Symbol \( M \) stands for \( 10^6 \)

Table 5 Overall performance of English-to-German engines with respect to human references on the document of the proper field-test session

| Pair        | MT engine | Legal domain |
|-------------|-----------|--------------|
|             | BLEU      | BP           | TER        | GTM        |
| en→de       | IN        | 18.3         | 0.98       | 68.4       | 50.5       |
|             | PA\(_{IN}\) | 19.7         | 0.97       | 66.6       | 52.3       |
|             | DA        | 19.3         | 0.93       | 65.0       | 52.6       |
|             | PA\(_{DA}\) | 20.1       | 0.95       | 64.7       | 52.8       |
|             | PA\(_{DA}\) + warm-up | 21.1     | 0.95       | 65.2       | 52.8       |

BP stands for BLEU’s brevity penalty

Sects. 5.1 and 5.2. First, we built a domain-adapted system by performing data selection in all available corpora using a development set which is supposed to be representative for the domain (about 35k source words).\(^8\)

This resulted in a selection of 34.6% of the bilingual and 43.0% of the monolingual data, employed to train a 4-gram LM. This same development set was also used for MERT. Project adaptation was performed similarly, but using data produced by the human translator during the warm-up session as seed for monolingual and bilingual data selection. This resulted in a selection of 43.5% of the bilingual and 52.4% of the monolingual data. Since this warm-up data is rather small (about 3k words), we combined it with the generic domain development set to circumvent potential overfitting by MERT.

5.3.3 Results

In Table 5, MT scores of all English-to-German engines are reported. The first two rows replicate the en→de entries in Table 3 for easing the comparison. The DA and PA\(_{DA}\)

\(^8\) It is a report by the European Parliament, not included in the training data, containing a proposal for financial regulations in the European Union, available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A7-2013-0039+0+DOC+XML+V0//EN.
entries refer to the domain- and project-adapted systems described in the previous section.

First of all it has to be noted that the domain-adapted engine (DA) definitely outperforms that trained on the in-domain corpus only (IN) by 1.0 absolute BLEU points, 3.4% TER and 2.1% GTM. As the two systems were trained on approximately the same amount of text, this result clearly indicates the ability of the selection process to find additional data in the generic corpora which is appropriate for the domain.

Concerning the project adaptation scheme of Sect. 5.3.2, it yields a relative BLEU improvement of 4.1% (19.3–20.1), while minor gains are observed in terms of TER and GTM. Evidently, the DA system already fits well the translation project—as can be seen from the fact that its performance is equivalent to that of the PA\textsubscript{IN} system—and it is not easy to further adapt it to the project using only 3,000 words of new data. Nevertheless, we do achieve improvements. It is also interesting to note that, on one hand, the TER of the second adaptation scheme is lower than for the schemes IN and PA\textsubscript{IN}. On the other hand, the BLEU score is handicapped by a brevity penalty below 0.95. It may be argued that TER is more closely related to the editing effort of a human translator than BLEU.

Finally, we added the parallel data obtained from the warm-up session to the bitexts and built a new system (last line in Table 5). Normally, we would not expect any impact on the translation performance by adding just 3,000 words to a training corpus of more than 50 million words. Surprisingly, however, the BLEU score increased by one point. Apparently, the warm-up session does contain new words or expressions which also appear in the second day of the field test.

We conclude from these experiments that data selection is very effective in extracting the most relevant parallel and monolingual data to adapt an SMT system to a particular domain, and to a translation project running over several days. The data obtained after one day of work—typically around three thousand words—is already enough to guide this selection process. In addition, it can be helpful to add this data to the bitexts, but we do not anticipate that such a small amount of data will improve the translation process in all cases.

5.4 Examples

Figure 1 provides four examples showing the impact of project adaptation on the quality of automatic translations, two from the legal/English-to-Italian task and two from the legal/English-to-German task.

In the first English-to-Italian example, the translation from the domain-adapted system is correct, but it is very literal and the verb “take”, the adjective “appropriate” and the noun “measures” are respectively translated as “prendere”, “adeguate” and “misure”, which differ from the reference translations (“adottare”, “opportuni” and “provvedimenti”, respectively) that better suit the specific document at hand; also the phrase “a recurrence” is translated literally as “il ripetersi”, while the emphatic “che i fatti si ripetano” is preferred in the manual translation. The project adaptation helps us to reduce the difference between the machine-generated translation and the reference translation: the verb, the adjective and the phrase are translated as expected, while...
only the noun “measures” is not recovered, which also yields the mismatch between the masculine “opportuni” and the feminine “opportune” plural declensions of the adjective.

In the second English-to-Italian example also, the correct IN translation “relazione di controllo” of “inspection report” is appropriately replaced by “rapporto di ispezione” in the PA_IN translation. However, the errors are not fully recovered: the lack of context prevents the translation of “it” with the feminine pronoun “essa” (the masculine “esso” is used instead), while an intrinsic weakness of the models yields the number mismatch between the subject “esso” (singular) and the verb “fanno” (plural), which should be “fa”.

The first English-to-German example shows how the system has learned the translation of “security rules”. This term does not appear in our development data, but several times with the corresponding translation in the warm-up texts. Overall, the translation project (warm-up and field test) are more focused on security issues than the more general development set. The corresponding terms were correctly introduced by our adaptation scheme. In the second English-to-German example, it is interesting to see that the PA_DA system obtains a much better translation although the English word “repeal” appears only in the field test data. However, the word “Decision” and the expression “Council Decision” is much more frequent in the warm-up data.

5.5 Analysis and discussion

First of all, we tried to evaluate the potential of project-adapted models over the starting models. For this purpose, all engines—whether project-adapted or not—were tuned on the evaluation set, that is the field-test document, without changing the models
Table 6 Overall “upperbound” (*) performance of MT engines on field-test documents

| Pair   | MT engine | IT domain | Legal domain |
|--------|-----------|-----------|--------------|
|        |           | BLEU      | TER          | GTM          |
|        |           |           | BLEU         | TER          | GTM          |
| en→it  | IN*       | 59.0      | 25.5         | 79.5         | 34.2         | 49.5         | 64.4         |
|        | PA<sub>IT</sub>* | 58.8      | 26.2         | 79.0         | 37.6         | 46.5         | 66.5         |
| en→fr  | IN*       | 43.8      | 36.5         | 71.0         | 35.7         | 50.9         | 64.5         |
|        | PA<sub>IT</sub>* | 44.2      | 36.3         | 71.3         | 38.7         | 48.6         | 66.3         |
| en→es  | IN*       | –         | –            | –            | 37.8         | 48.8         | 66.9         |
|        | PA<sub>IT</sub>* | –         | –            | –            | 39.8         | 46.8         | 67.8         |
| en→de  | IN*       | –         | –            | –            | 19.6         | 66.7         | 51.7         |
|        | PA<sub>IT</sub>* | –         | –            | –            | 20.7         | 67.0         | 52.3         |
|        | DA*       | –         | –            | –            | 21.9         | 66.1         | 53.8         |
|        | PA<sub>DA</sub>* | –         | –            | –            | 21.7         | 65.5         | 53.2         |

Tuning means the estimation via MERT of the log-linear interpolation weights which maximize the BLEU score. By these means, we obtain an estimate of the upperbound performance of our systems. The aim of this experiment was to understand whether the problems observed on the IT domain and reported in Sect. 5.2 are due to a bad choice of development set, or too big a mismatch between the warm-up and field-test documents.

Table 6 collects for each SMT system the automatic scores obtained at the end of the MERT procedure run on the field test documents.

For the IT domain, the “upperbound” experiments show that there is no room for improving the baseline reference performance – already quite a strong baseline engine – with the project-adapted models. On the English-to-Italian task, the fair outcomes (Table 3) were optimistic since the gains between the IN and PA<sub>IT</sub> models in terms of BLEU and TER observed there (about 2 and 3 absolute points, respectively) vanished completely in the unfair experiment. On the English-to-French task, the negligible increase in performance in the upperbound experiment justifies the only marginal improvement of the fair project-adapted engine over the baseline engine. The additional outcome of this experiment is that the fair tuning is really effective: in fact, the use of upperbound weights improves the BLEU scores of the fairly estimated weights by no more than 7% relative, the maximum being 6.8% of the PA<sub>IT</sub> en→fr task (41.4–44.2).

Concerning the legal domain, with the exception of the English-to-German domain-adapted engines, this experiment shows that: (i) there is quite a large potential for project-adapted models to improve performance of baseline models: from Table 6, 10% for English-to-Italian (34.2–37.6), 8% for English-to-French (35.7–38.7) and 5% for both English-to-Spanish (37.8–39.8) and English-to-German (19.6–20.7); (ii) such a potential is well exploited, as the upperbound performance is not much larger than the fair improvements reported in Sect. 5.2; (iii) as in the IT domain, the fair tuning is effective for both baseline and project-adapted legal engines, since the upperbound
BLEU scores (Table 6) are better by at most 10% than fair scores (Table 3), the larger difference being for the IN en -> it task: from 31.0 to 34.2, i.e. 10.3% relative.

The upperbound experiments with the English-to-German domain-adapted engines show a different pattern: the DA* and PADA* scores are almost identical, slightly better in fact than the PADA system. We explain this by the fact that the DA model was already better adapted to the domain since we already applied data selection of the DA model, with help of a representative development corpus. Nevertheless, our project adaptation method is robust since we are still able to improve the DA system.

5.5.1 Warm-up versus field test documents

A second set of experiments was designed to investigate why some of the project-adapted engines, especially for the IT domain, had no potential to improve baseline engines, differently than in most of the legal tasks. As described in the previous sections, project adaptation is performed on the warm-up document with the aim of adapting the in-domain models towards the specific text to be translated. The underlying assumption is that the text translated during the warm-up period is a good representation of what will be translated during the proper field-test session. Tables 7 and 8 provide some statistics computed to investigate the representativeness of warm-up documents with respect to the documents translated in the field test session.

Table 7 provides statistics on the source-side overlap between the warm-up and the field-test documents of the two considered domains, whose sizes are provided in Table 2. Each entry gives the percentage rate of the \( n \)-gram types of the corresponding document (either warm-up or field-test) which occur in both documents. The higher the number, the higher the overlap, i.e. the greater the similarity of the two documents. It is clear that legal documents are definitely closer to one another than the IT documents, as evidenced by the rate of shared 4-grams between legal documents being three times higher than for IT documents.

In Table 8, the column RR reports the repetition rate of warm-up and field-test documents on the target side; the other columns show the perplexity (PP) and out-of-vocabulary (OOV) word percentage of those documents on four different LMs (see caption).

The repetition rate (Bertoldi et al. 2013) is a measure of the repetitiveness of a text. It is defined as the geometric mean of the rate of non-singleton \( n \)-gram types (\( n=1, \ldots, 4 \)). In order to make the rates comparable across different sized corpora, statistics are collected on a fixed-size sliding window of 1,000 words, and properly averaged. Formally, the RR in a document can be expressed as in (5):

| Table 7 | Rate of shared \( n \)-gram types between the warm-up and field-test documents, for each domain |
|---------|--------------------------------------------------|
| Domain | Test set | 1-gram | 2-gram | 3-gram | 4-gram |
| IT     | Warm-up  | 53.2   | 23.1   | 9.8    | 6.4    |
|        | Field-test | 22.1   | 6.6    | 2.4    | 1.5    |
| Legal  | Warm-up  | 70.5   | 45.4   | 25.9   | 15.5   |
|        | Field-test | 29.8   | 15.0   | 7.7    | 4.5    |
Table 8  Repetition rate (RR) of warm-up and field-test documents and their perplexity (PP) and out-of-vocabulary (OOV) word percentage, computed with respect to language models of the in-domain (IN) and of the project-adapted (PA) systems, and to language models trained only on data selected for project adaptation (PS) or on the concatenation of warm-up document and selected texts (WU+PS)

| Pair | Test set | IT domain | Legal domain |
|------|----------|-----------|--------------|
|      |          | RR        | PP/OOV%      |              |
|      |          | IN        | PS           | WU+PS        | PA            |
| en→it| Warm-up  | 30.2      | 202/2.25     | 131/2.32     | 5.9/0.00      | 34.5/0.00     |
|      | Field-test| 22.7      | 266/5.54     | 239/6.23     | 236/6.18      | 246/5.41      |
| en→fr| Warm-up  | 29.5      | 242/4.2      | 178/4.98     | 6.2/0.00      | 37/0.00       |
|      | Field-test| 23.9      | 441/5.5      | 436/7.14     | 419/6.88      | 406/5.27      |
| en→it| Warm-up  | 17.1      | 73.6/0.14    | 50.1/0.29    | 4.5/0.00      | 18.8/0.00     |
|      | Field-test| 15.7      | 105/0.32     | 87.6/0.84    | 70.1/0.82     | 76.5/0.31     |
| en→fr| Warm-up  | 19.0      | 52.6/0.26    | 36.6/0.49    | 4.6/0.00      | 14.5/0.00     |
|      | Field-test| 16.4      | 67.5/0.50    | 56.1/0.76    | 44.3/0.63     | 47.6/0.38     |
| en→es| Warm-up  | 19.7      | 68.0/2.00    | 51.4/2.27    | 4.7/0.00      | 15.9/0.00     |
|      | Field-test| 16.4      | 80.3/1.52    | 70.3/1.91    | 52.0/0.80     | 56.5/0.40     |
| en→de| Warm-up  | 15.2      | 180/2.94     | 126/3.19     | 4.7/0.00      | 27.3/0.00     |
|      | Field-test| 14.4      | 236/2.69     | 197/3.45     | 133/2.04      | 153/1.37      |

\[
RR = \left( \prod_{n=1}^{4} \frac{\sum_S (V(n) - V(n, 1))}{\sum_S V(n)} \right)^{1/4}
\] (5)

where \(S\) is the sliding window, \(V(n, 1)\) is the number of singleton \(n\)-gram types in \(S\), and \(V(n)\) is the total number of different \(n\)-gram types in \(S\). The highest and lowest values, \(RR=1\) and \(RR=0\), are achieved when all distinct \(n\)-grams observed in all sliding windows occur, respectively, more than once and exactly once.

Looking at Table 8, the RR of the warm-up documents differs from the RR of the field-test documents a bit more in the IT domain than in the legal domain, which is a first (weak) warning as to whether to adopt IT warm-up texts as seeds for data selection. However, the strongest reason against the use of IT warm-up documents for data selection is given by results in terms of perplexity. Let us focus on English-to-Italian as an example (analogous considerations hold for the en→fr pair): the adapted IT language model improves the PP of the field-test documents over the baseline language model by only 7.5% (from 266 to 246), whereas in the legal domain the relative improvement is 27.1% (from 105 to 76.5). Such results are due to the fact that data selected in the IT domain are far from being good fits to the field test documents, as demonstrated by looking at the perplexity on language models estimated on the
selected text (PS). Moreover, by adding the warm-up document to the selected text, the resulting language model (WU+PS) hardly changes at all in the IT domain (PP goes from 239 to 236), but improves a lot in the legal domain (from 87.6 to 70.1), especially taking into account the small size of the warm-up text (about 3,000 words, see Table 2).

The outcomes in terms of \( n \)-gram overlap, repetition rate and perplexity undoubtedly show that the warm-up documents are good representatives of the field test texts of legal en→it/fr/es/de tasks, but not of IT tasks. This explains why the potential of the project-adapted models is significant only for the legal tasks (Table 6) and therefore the lack of improvement for IT tasks in fair experiments (Sect. 5.2). Nevertheless, our adaptation method proved to be quite robust as in the worst case it does not cause the quality of the baseline to deteriorate.

6 Field test

In this section, we report on the field tests run by the MateCat project to evaluate the impact of MT project adaptation on the productivity of professional translators. The field tests were run on the IT and legal domains for the English-to-Italian direction.

6.1 Protocol

The field test post-editing experiments were executed with the MateCat tool, an open-source web-based CAT tool, under development within the same project, integrating new MT functions such as the self-tuning presented in this paper, and built on top of state-of-the-art MT and CAT technologies.

The translation environment is shown in Fig. 2. It has been designed so as to be as fast and easy to use as possible for professional translators. One of the key goals was to minimize the learning curve so that translators could be as efficient with the MateCat tool as they would be with their standard CAT tool without extensive training and/or experience with the tool. Hence, the text is presented in a tabular view where the document is broken down into minimal units (segments). For the active segment (i.e. the segment the translator is editing), the three best hypothesised translations, from either the TM or the MT, are presented, ranked according to their quality; the quality is given by the fuzzy match value in case of suggestions coming from the TM, and by a default value for MT-generated suggestions. During the field test, the default MT quality was set to 85% by the organizers; such a setting yielded MT suggestions to be ranked better than any TM match most of the time, resulting in a clear preference of translators to edit automatic translations (97–98% of the cases).

The GUI was designed to allow translators to focus their attention on the active segment and on the supplied suggestions. While translators usually work on the text segment by segment, the MateCat Tool allows them to also move across segments, edit or proofread their output more times, without any restriction. For each interaction with a segment, the cumulative time needed to elaborate the final version of the translation is collected. The time taken to edit is shown on the right-hand side of each segment and then collected in the editing log.
The field test was organized over two days in which a document had to be translated by four professional translators. During the first day—the warm-up session—for the translation of the first half of the document, translators received MT suggestions by the DA engine; during the second day—the field-test session—MT suggestions came from the PA system, adapted to warm-up and (source of) field-test texts following the scheme proposed in this paper. The impact of the project adaptation was measured by comparing productivity of translators during the first and the second day. Productivity was measured by two key performance indicators described in the following section.

6.2 Key performance indicators

We used two key performance indicators to measure the effectiveness of our adaptation scheme, namely the time-to-edit and the post-editing effort.

**Time-to-edit (TTE)**, the average translation drafting speed by the translators. TTE aims at measuring the average productivity of translators. In particular, we measure the average time taken by the translator to complete a segment in seconds per word. Those segments whose cumulative editing time is either too large or too short are considered outliers and then discarded. For instance, an outlier can occur when a translator stops working, yet leaves the segment active (i.e. not accepted), or when the suggestion is accepted too quickly to indicate that it has really been checked.

**Post-editing effort (PEE)**, the average percentage of word changes applied by the translators to suggestions provided by the CAT tool. PEE aims at defining the quality of suggestions. We measured the percentage of words edited in a segment by comparing the CAT suggestion and the edited segment submitted by the translator. A proprietary function was used which compares two segments and provides a
Table 9  Time-to-edit (TTE) and Post-editing effort (PEE) for each translator in warm-up and field-test sessions (IT and legal domain, English-to-Italian pair)

| Domain | user | TTE (sec/word) | PEE | p value | Δ (%) | p value | Δ (%) |
|--------|------|----------------|-----|---------|-------|---------|-------|
|        |      | Warm-up | Field-test |       |       | Warm-up | Field-test |       |       |
| IT     | t1   | 4.70    | 3.36 | 0.001  | 28.51 | 34.27   | 30.99  | 0.060  | 9.57  |
|        | t2   | 2.26    | 2.47 | 0.220  | -9.29 | 38.50   | 39.52  | 0.330  | -2.65 |
|        | t3   | 3.17    | 3.11 | 0.450  | 1.89  | 32.53   | 30.17  | 0.133  | 7.25  |
|        | t4   | 4.77    | 3.64 | 0.006  | 23.69 | 32.22   | 28.44  | 0.040  | 11.73 |
| Legal  | t1   | 5.20    | 5.63 | 0.222  | -8.27 | 26.47   | 24.57  | 0.212  | 7.18  |
|        | t2   | 5.42    | 3.92 | 0.002  | 27.68 | 29.11   | 26.25  | 0.140  | 9.82  |
|        | t3   | 5.86    | 4.32 | 0.000  | 26.28 | 35.65   | 34.11  | 0.247  | 4.32  |
|        | t4   | 6.60    | 3.73 | 0.000  | 43.48 | 22.72   | 18.07  | 0.011  | 20.47 |

The difference of these measures achieved in the two sessions and its significance p-value are also reported.

match percentage score; this score resembles TER but is computed by assigning less weight to punctuation and casing errors.

6.3 Results

Table 9 reports results in terms of key performance indicators for all translators involved in the English-to-Italian IT and legal tasks. Significant TTE and PEE improvements can be observed between warm-up and field-test sessions together with the corresponding p-values computed with a randomized permutation test (Noreen 1989).

On the IT domain, two translators out of four improved significantly according to both figures (t1 and t4), while on the legal domain this was the case for three out of four (t2–t4). Most TTE reductions (five out of eight) were statistically significant (p value < 0.05), while the same hold only for two of the observed PEE variations. By looking at the average productivity gains, on the IT domain we observed an 11.2% gain in TTE and a 6.5% in PEE, while on the legal domain we observed a 22.2% gain in TTE and a 10.7% improvement in PEE. Finally, the good correlations observed between PEE and TTE under the different conditions indicate it to be very likely that the translators were able to take advantage of the MT suggestions, and that the adapted MT engine suggestions were in general better. In fact, better PEE effort was observed for seven translators out of eight.

7 Conclusions

In this paper we addressed a hot research topic for the CAT industry, namely how self-tuning capability can be added to SMT systems incorporated in CAT tools. Self-tuning can be viewed on two different scales: at the domain level, or simply at the project level. On a larger scale, the goal is to focus general-purpose models towards the specific domain of interest; for example, this could be applied for preparing the MT
system to be employed at the beginning of the translation process once the domain of the translation project is known. On a lower scale, the goal is to further focus in-domain models towards the specific translation project, once the source text is available and the post-edits start to arrive: this kind of self-tuning can be applied at any time, provided that enough fresh data is at one’s disposal for updating the models according to the needs of the methods employed.

To handle this type of self-tuning, we have proposed an adaptation scheme which has been tested in an extensive experimental framework, consisting not only of lab tests but also field tests which involved professional translators and the industrial partner of MateCat, the project in which this work has been conducted.

The experimental results demonstrate the effectiveness of the proposed scheme used to integrate project-adapted SMT systems into the CAT workflow; gains of human translator productivity up to over 43% were measured.

Nevertheless, the method works if the seed used for data selection is a good representative of the document to be translated. In fact, where this condition is not satisfied, as in our IT tasks, the adapted engines are unable to outperform the reference baseline systems; in this case performance does not decrease, underlining the conservative nature of the adaptation scheme.

Several issues remain open and deserve to be investigated in the future. First of all, the prediction of the behaviour of the adapted models is extremely important: is it possible to forecast whether an adapted engine is effectively able to generate better suggestions than those of the reference system for a document whose source side is given, and if so how?

Another issue regards the iterative application of the proposed daily adaptation procedure: what does the learning curve look like? Does it (soon) reach a plateau? Is the daily frequency the optimal rate?

Finally, gains observed in our field-test experiments could be partially due to the familiarization of the users with the system and with the specific project, or to a different translation difficulty of the documents used in the two sessions. Actually, we are aware of this issue, and in the MateCat field tests following those reported in this paper, such effects were mitigated by generating suggestions in the field-test session by both the systems under investigation as well as the reference system, so that the net contribution of the tested methods could be measured.

Acknowledgments This work was supported by the MateCAT project, which is funded by the EC under the 7th Framework Programme.

References

Axelrod A, He X, Gao J (2011) Domain adaptation via pseudo in-domain data selection. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, pp 355–362

Bach N, Hsiao R, Eck M, Charoenpornsawat P, Vogel S, Schultz T, Lane I, Waibel A, Black AW (2009) Incremental adaptation of speech-to-speech translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL HLT) Conference: Short Papers. Boulder, US-CO, pp 149–152
Bertoldi N, Cettolo M, Federico M, Buck C (2012) Evaluating the learning curve of domain adaptive statistical machine translation systems. In: Proceedings of the Workshop on Statistical Machine Translation (WMT). Montréal, pp 433–441

Bertoldi N, Cettolo M, Federico M (2013) Cache-based online adaptation for machine translation enhanced computer assisted translation. In: Proceedings of the MT summit XIV. Nice, pp 35–42

Bisazza A, Ruiz N, Federico M (2011) Fill-up versus interpolation methods for phrase-based SMT adaptation. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT). San Francisco, US-CA, pp 136–143

Bojar O, Buck C, Callison-Burch C, Federmann C, Haddow B, Koehn P, Monz C, Post M, Soricet R, Specia L (2013) Findings of the 2013 workshop on statistical machine translation. In: Proceedings of the eighth workshop on statistical machine translation. Sofia, pp 1–44

Cettolo M, Servan C, Bertoldi N, Federico M, BarraULT L, Schwenk H (2013) Issues in incremental adaptation of statistical mt from human post-edits. In: Proceedings of the MT summit XIV Workshop on Post-editing Technology and Practice (WPTP-2). Nice, pp 111–118

Chen SF, Goodman J (1999) An empirical study of smoothing techniques for language modeling. Comput Speech Lang 4(13):359–393

Crammer K, Dekel D, Keshet J, Shalev-Shwartz S, Singer Y (2006) Online passive–aggressive algorithms. J Mach Learn Res 7:551–585

Federico M, Cattelan A, Trombetti M (2012) Measuring user productivity in machine translation enhanced computer assisted translation. In: Proceedings of conference of the Association for Machine Translation in the Americas (AMTA). San Diego, US-CA

Foster G, Kuhn R (2007) Mixture-model adaptation for SMT. In: Proceedings of the Workshop on Statistical Machine Translation (WMT). Prague, pp 128–135

Foster G, Gouette C, Kuhn R (2010) Discriminative instance weighting for domain adaptation in statistical machine translation. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP). Cambridge, US-MA, pp 451–459

Galley M, Manning CD (2008) A simple and effective hierarchical phrase reordering model. In: Proceedings of the Conference on empirical methods in natural language processing (EMNLP). Honolulu, US-HI, pp 848–856

Gao J, Zhang M (2002) Improving Language model size reduction using better pruning criteria. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL). Philadelphia, US-PA, pp 176–182

Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, Paris, pp 439–448

Guéberof A (2009) Productivity and quality in MT post-editing. In: Proceedings of the MT summit XII, Beyond translation memories: new tools for translators workshop. Ottawa, Canada

Hardt D, Elming J (2010) Incremental re-training for post-editing SMT. In: Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA). Denver, US-CO

Hasler E, Haddow B, Koehn P (2012) Sparse lexicalised features and topic adaptation for SMT. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT). Hong Kong, pp 268–275

Kneser R, Steinbiss V (1993) On the dynamic adaptation of stochastic language models. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), vol II, Minneapolis, US-MN, pp 586–588

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: Proceedings of the MT summit X. Phuket, pp 79–86

Koehn P, Schroeder J (2007) Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Workshop on Statistical Machine Translation (WMT). Prague, pp 224–227

Koehn P, Axelrod A, Mayne AB, Callison-Burch C, Osborne M, Talbot D (2005) Edinburgh system description for the 2005 IWSLT speech translation evaluation. In: Proceedings of the international workshop on spoken language translation (IWSLT). Pittsburgh, US-PA

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, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL): Companion volume proceedings of the demo and poster sessions. Prague, pp 177–180

123
Läubli S, Fishel M, Massey G, Ehrensberger-Dow M, Volk M (2013) Assessing post-editing efficiency in a realistic translation environment. In: Proceedings of the MT summit XIV, workshop on post-editing technology and practice. Nice, pp 83–91
Liu L, Cao H, Watanabe T, Zhao T, Yu M, Zhu C (2012) Locally training the log-linear model for SMT. In: Proceedings of the joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Jeju Island, pp 402–411
Matsoukas S, Rosti AVI, Zhang B (2009) Discriminative Corpus weight estimation for machine translation. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, pp 708–717
Moore RC, Lewis W (2010) Intelligent selection of language model training data. In: Proceedings of the annual meeting of the Association of Computational (ACL): Short Papers. Uppsala, pp 220–224
Nakov P (2008) Improving English-Spanish Statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In: Proceedings of the Workshop on Statistical Machine Translation (WMT). Columbus, US-OH, pp 147–150
Niehues J, Waibel A (2012) Detailed Analysis of different strategies for phrase table adaptation in SMT. In: Proceedings of the conference of the Association for Machine Translation in the Americas (AMTA). San Diego, US-CA
Noreen EW (1989) Computer intensive methods for testing hypotheses: an introduction. Wiley Interscience, New York
Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of the annual meeting of the Association for Computational (ACL). Sapporo, pp 160–167
Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51
Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the annual meeting of the Association of Computational (ACL). Philadelphia, US-PA, pp 311–318
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull Math Linguist 93:7–16
Quenouille MH (1956) Notes on bias in estimation. Biometrika 43:353–360
Rousseau A (2013) XenC: an open-source tool for data selection in natural language processing. Prague Bull Math Linguist 100(1):73–82
Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the Conference of the association for machine translation in the Americas (AMTA). Cambridge, US-MA, pp 223–231
Steinberger R, Pouliquen B, Widiger A, Iğnat C, Erjavec T, Tuğr D, Varga D (2006) The JRC-acquis: a multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the international conference on language resources and evaluation (LREC). Genoa, pp 2142–2147
Tiedemann J (2012) Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the international conference on Language Resources and Evaluation (LREC). Istanbul, pp 2214–2218
Turian JP, Shen L, Melamed ID (2003) Evaluation of machine translation and its evaluation. In: Proceedings of MT summit IX, New Orleans, US-LA, pp 386–393
Yasuda K, Zhang R, Yamamoto H, Sumita E (2008) Method of Selecting training data to build a compact and efficient translation model. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). Hyderabad, pp 655–660