Noname manuscript No.
(will be inserted by the editor)

Translation Project Adaptation for MT-Enhanced Computer Assisted Translation

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

Received: date / Accepted: date

Abstract The effective integration of MT technology into computer-assisted translation tools is a challenging topic both for academic research and the translation industry. Particularly, professional translators feel crucial the ability of MT systems to adapt to their feedback. 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 at the 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 strategy can be very effective provided the seed data used for adaptation is enough related to the remaining text to translate; otherwise, MT quality does neither improve nor worsen, 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

M. Cettolo, N. Bertoldi, M. Federico
FBK, Fondazione Bruno Kessler
38123 Povo, Trento, Italy
E-mail: LastName@fbk.eu

H. Schwenk, L. Barrault, C. Servan*
LIUM, University of Le Mans
72085 Le Mans cedex 9, France
E-mail: FirstName.LastName@lium.univ-lemans.fr
*C. Servan is now with Xerox Research Centre Europe, 38240 Meylan, France

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 work flow. 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 in fact well known that a system optimized on the text genre it is used for performs better than a generic system. This process is usually called *domain adaptation*. In addition, it is as well 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 makes the same errors over and over again.

In the translation industry, a typical scenario is that of one or more translators working 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 likely improved translations will be generated in 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 machine translation. 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, that is MT with domain and project adaptation functionality. We report here the results of our approaches to 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, 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 preced-

[1] In computer assisted translation, 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 machine translation engines.

[2] `http://www.matecat.com`

ing *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 report here also 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 from 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 critical set-up our adaptation method proved to be conservative as no degradation of the baseline reference quality was observed.

The paper is organized as follows. Section 2 discusses related works. Project adaptation methods are outlined in Section 3. In Section 4, the data used in the experiments is introduced, while lab and field tests are described and commented in Sections 5 and 6, respectively. The paper concludes with a discussion.

2 Related Work

The idea that machine translation can boost the productivity of human translators has been consolidating over the recent years thanks, in first instance, to the remarkable progresses achieved with the statistical MT approach and, in second instance, to several experimental investigations that 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 of which 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 percent fuzzy match range. A commercial statistical MT 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 second), a detailed analysis is 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 for comparing productivity of human translation versus post-editing MT outputs. The test was performed on IT documentation, on 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.[3] 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 from English into Italian or German, full IT or legal documents rather than isolated segments, without changing their working routine. Indeed, MT suggestions were provided just 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, too. Similar and more thoroughly 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 statistical significant improvements both on productivity and overall translation quality.

From the MT methods viewpoint, 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 statistical MT 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, similarly to us. However, the adaptation of the MT module involved only the language model and is performed on the MT outputs.

Niehues and Waibel (2012) compared different approaches to adapt 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

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

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 machine translation 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 compare several approaches for tuning sparse features on top of both small in-domain and larger mixed-domain systems: MERT, MIRA, Jackknife 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 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; 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. a 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 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) which works as follows. We start with a language model trained on a seed of task-specific data, a task specific small development corpus on which perplexity is optimized, and a large generic corpus from where to extract task-relevant sentences. First, the task-specific cross-entropy is computed for each sentence in the generic corpus. Then, the same sentences are scored against a language model trained on a random sample of the generic corpus; the sample size is roughly set equal to the seed corpus. Next, the difference between task-specific cross-entropy and generic cross-entropy is computed for each sentence. Finally, sentences are sorted on the basis of this score. This was shown to lead to better selection than the simple perplexity sorting, initially proposed in (Gao and Zhang, 2002).

Once the generic corpus is sorted, the optimal percentage of selected data has to be determined. The estimation is performed by minimizing the perplexity of 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 around 20% of the generic data. We were able to confirm this tendency for many different tasks. As a side effect, the models become considerably smaller which is also an important aspect when deploying MT systems in real applications.

The procedure sketched above works well to select monolingual data. It can be also used to select the most appropriate bitexts either by choosing one of the two sides or by working on both and then combining somehow the two computations. In our particular case, there are several options to build the seed for data selection. After one day of work on a project, we can use the source text and the human translations produced either by post-editing or direct translation, to perform data selection on a bilingual seed. On the other side, 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 using a source-side only but larger seed.

This data selection method was proposed by Axelrod et al (2011) and is available in the XenC (Rousseau, 2013) and IRSTLM (Federico et al, 2008) toolkits.

3.2 Adaptation of SMT models

Translation and distortion models: Data selection is a very effective method to adapt the translation model on 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. Therefore, 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 adapted foreground phrase table by adding only phrase pairs that do not appear in the foreground table. Notice that all the original scores of both the background and foreground phrase and distortion tables are preserved. Moreover, only for the translation table, an additional indicator feature is introduced to signal whether each phrase entry stems from the foreground or from the background phrase table.

We chose the fill-up technique because it performs as good 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 simplified version of the fill-up method, called *back-off*, in which the indicator feature is omitted.

Language model: As concerns the language model adaptation, we employed the mixture of model which consists of the convex combination of one or more background language models with a foreground language model.

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, 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. They are all publicly available. In addition, a proprietary large translation memory (TM), that is a collection of parallel entries, 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 purposes. In particular, the `train` entries refer to the whole generic training texts, the `project selection` entries refer to the subset of `train` data that was selected for project adaptation to the specific document to translate, and development sets to additional data on which the parameters of the phrase-based MT model were optimized.

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 corpora[4] (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 language pair (15.4/17.9 million words) is about four times smaller than for the other tasks. Similarly, also for the legal en→de task,

[4] Available from http://www.statmt.org/wmt13/translation-task.html.

domain specific data (again, refer the entry `domain selection`) was selected from various generic linguistic resources, namely, Europarl, JRC-Acquis and proprietary TMs.

domain	pair	corpus	segments	tokens	
				source	target
IT	en→it	train	5.4 M	57.2M	59.9M
		project selection	0.36M	3.8M	4.0M
		development set	2,156	26,080	28,137
	en→fr	train	1.1 M	15.4M	17.9M
		domain selection	1.2 M	20.0M	22.2M
		project selection	0.53M	8.6M	9.5M
		development set	4,755	26,747	30,100
Legal	en→it	train	2.7 M	61.4M	63.2M
		project selection	0.18M	5.4M	5.4M
		development set	181	5,967	6,510
	en→fr	train	2.8 M	65.7M	71.1M
		project selection	0.18M	5.5M	5.8M
		development set	600	17,737	19,613
	en→es	train	2.3 M	56.1M	62.0M
		project selection	0.18M	5.6M	6.1M
		development set	700	32,271	36,748
	en→de	train	5.8 M	140.3M	131.4M
		domain selection	1.9 M	49.3M	45.4M
		project selection	2.3 M	62.2M	57.2M
		development set	925	35,270	32,277

Table 1: Overall statistics on parallel data used for training and development (tuning) purposes: number of segments and running words of source and target sides. Symbol M stands for 10^6.

4.2 Evaluation data

For the IT domain, data were supplied by the industrial partner of the MateCat project. 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 (1614 segments).

For the legal domain a document was taken from the European Commission website, 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 few segments not available for all languages.

domain	pair	test set	segments	tokens	
				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

Table 2: Overall statistics on parallel data used for evaluation purposes: number of segments and running words of source and target sides.

5 Lab Test

5.1 MT systems

The SMT systems have been built upon 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 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 interpolation model on development sets whose content is coherent to training data and of adequate size (entries `development set` of Table 1).

For each task, two different SMT engines have been tested, the reference `domain-adapted` (DA) system and the `project-adapted` (PA) system.

The models of DA engines were estimated on domain-specific training data, i.e. entries `train` (IT English-to-Italian, legal English-to-Italian/French/Spanish systems), `domain selection` (legal English-to-German system) or both (IT English-to-French system) of Table 1.

Project-specific data (entries `project selection` of Table 1) were selected from generic training corpora by means of the *data selection* method described in Section 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 translate during the field-test session. Project-specific models were trained on the concatenation of texts selected this way and of warm-up documents. The models of PA engines are the combination of project-specific and domain-adapted models, via the *back-off* and *LM mixture* methods of Section 3.

The adaptation scheme used for English-to-German system differs slightly. 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). This resulted in a selection of 35.1% of the bilingual and 43.0% of the monolingual data. The LM is a 4-gram. 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 44.3% 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 eventual overfitting by MERT.

5.2 Results

Table 3 provides BLEU, TER and GTM scores computed on the field test documents with respect to human references of the domain and project adapted systems for each of the six translation tasks.

pair	MT engine	IT domain			Legal domain		
		BLEU	TER	GTM	BLEU	TER	GTM
en→it	DA	55.3	29.2	77.8	31.0	53.1	61.8
	PA	57.5	26.3	78.6	35.0	49.1	64.6
en→fr	DA	41.3	38.3	69.5	33.9	52.2	63.0
	PA	41.4	37.9	69.9	36.4	49.1	65.1
en→es	DA	–	–	–	35.5	50.7	65.7
	PA	–	–	–	36.4	50.2	65.6
en→de	DA	–	–	–	19.3	65.0	52.6
	PA	–	–	–	20.1	64.7	52.8

Table 3: Overall performance of MT engines with respect to human references on the documents of the proper field-test session.

For the legal domain, the improvements over the reference system yielded by the adaptation technique is quite effective. As an example, the BLEU score improves by 12.9% (31.0 to 35.0) when translating into Italian, by 7.4% (33.9 to 36.4) into French, by 2.5% (35.5 to 36.4) into Spanish and by 4.1% (19.3 to 20.1) into German.

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 pretty literal and the verb *"take"*, the adjective *"appropriate"* and the noun *"measures"* are respectively translated as *"prendere"*, *"adeguate"* and *"misure"*, which differ from 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 allows 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.

src	take appropriate measures to prevent a recurrence ;
DA	prende adeguate misure per evitare il ripetersi ;
PA	adottare opportune misure per impedire che i fatti si ripetano ;
ref	adottare i provvedimenti opportuni per impedire che i fatti si ripetano ;
src	It shall not form part of the inspection report .
DA	Esso non fanno parte della relazione di controllo .
PA	Esso non fanno parte del rapporto di ispezione .
ref	Essa non fa parte del rapporto di ispezione .
src	. . . on the **security rules** and procedures for protecting EUCI
DA	. . . über die **Regeln** und Verfahren zum Schutz von EU-Verschlusssachen
PA	. . . über die **Sicherheitsvorschriften** und Verfahren zum Schutz von EU-Verschlusssachen
ref	. . . über die **Sicherheitsvorschriften** und -verfahren für den Schutz von EU-VS
src	This Decision shall repeal and replace Council Decision 2011/292/EU.
DA	Diese Entscheidung **und ersetzt und hebt die Entscheidung** des Rates 2011/292/EU.
PA	Diese Entscheidung **wird aufgehoben und ersetzt den Beschluss** des Rates 2011/292/EU.
ref	Der Beschluss 2011/292/EU des Rates wird durch den vorliegenden Beschluss aufgehoben und ersetzt.

Fig. 1 Examples comparing improvements due to project adaptation over the translation from the domain adapted system (legal domain, English-to-Italian and English-to-German directions). Source (`src`) and reference (`ref`) texts are also shown.

Also in the second English-to-Italian example, the correct DA translation *"relazione di controllo"* of *"inspection report"* is appropriately replaced by *"rapporto di ispezione"* in the PA translation; however, the errors are not recovered: the lack of context prevents to translate *"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 have to 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 the PA 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.

Differently than the legal domain, in the IT domain the only significant gain is observed for the Italian direction, where the BLEU score increases by 4% (55.3 to 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. It is reported in the following section.

5.3 Analysis and discussion

First of all, we tried to evaluate the potential of project-adapted models over the baseline models. For this purpose, all six engines were tuned on the evaluation set, that is the field-test document, without changing the models themselves. By these means, we hoped to get an estimate of the "oracle" performance of our systems. The aim of this experiment was to understand if the problems observed on the IT domain are due to a bad choice of the development set, or a too large mismatch between the warm-up and field-test documents.

Table 4 shows the automatic scores got at the end of the MERT procedure run on the field test documents.

pair	MT engine	IT domain			Legal domain		
		BLEU	TER	GTM	BLEU	TER	GTM
en→it	DA.oracle	59.0	25.5	79.5	34.2	49.5	64.4
	PA.oracle	58.8	26.2	79.0	37.6	46.5	66.5
en→fr	DA.oracle	43.8	36.5	71.0	35.7	50.9	64.5
	PA.oracle	44.2	36.3	71.3	38.7	48.6	66.3
en→es	DA.oracle	–	–	–	37.8	48.8	66.9
	PA.oracle	–	–	–	39.8	46.8	67.8
en→de	DA.oracle	–	–	–	21.9	66.1	53.8
	PA.oracle	–	–	–	21.7	65.5	53.2

Table 4: Overall "oracle" performance on field-test documents of MT engines tuned on the same texts.

For the IT domain, the "oracle" experiments show that there is no room for improving the baseline reference performance with the project-adapted models. On the English-to-Italian task, the fair outcomes (Table 3) are too lucky since the gains between the DA and PA 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 of performance in the oracle 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 oracle weights improves the BLEU scores of the fairly estimated weights by no more than 7% relative, the maximum being 6.8% of the $\mathrm{PA}_{\mathrm{en}\rightarrow\mathrm{fr}}$ task (41.4 to 44.2).

Concerning the legal domain, with the exception of the English-to-German pair, this experiment shows that: (i) there is quite a large potential for project-adapted models to improve performance of baseline models: from Table 4, 10% for English-to-Italian (34.2 to 37.6), 8% for English-to-French (35.7 to 38.7) and 5% for English-to-Spanish (37.8 to 39.8); (ii) such a potential is well exploited, being the upper-bound performance of oracle experiments not much larger than fair improvements reported in Section 5.2; (iii) as in the IT domain, the fair tuning is effective for both baseline and project-adapted legal engines, since the oracle BLEU scores (Table 4) are better by at most 10% than fair scores (Table 3), the larger difference being for $\mathrm{DA}_{\mathrm{en}\rightarrow\mathrm{it}}$ task: from 31.0 to 34.2, i.e. 10.3% relative.

The oracle experiments for the English-to-German task show a different pattern: the oracle BLEU scores for the DA and PA models are almost identical – in fact slightly better for the PA model. 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.

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 translate. The underlying assumption is that the text translated during the warm-up period well represents what will be translated during the proper field-test session. Table 5 provides some statistics computed to investigate the representativeness of warm-up documents with respect to the documents involved in the field test session.

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-grams (n=1...4). In order to make the rates comparable across different sized

pair	test set	IT domain				
		RR	PP/OOV%			
			DA	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

pair	test set	Legal domain				
		RR	PP/OOV%			
			DA	PS	WU+PS	PA
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	169/0.00	na	na	161/0.00
	field-test	14.4	210/0.00	na	na	198/0.00

Table 5: 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 domain-adapted (DA) 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).

corpora, statistics are collected on a fixed-size sliding window (one thousand words), and properly averaged. More formally:

$$\mathtt{RR} = \left(\prod_{n=1}^{4} \frac{\sum_S \left(\ V(n) - V(n,1)\ \right)}{\sum_S V(n)} \right)^{1/4} \tag{1}$$

where S is the sliding window, $V(n,1)$ is the number of singleton n-grams in S, and $V(n)$ is the total number of different n-grams 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 5, the RR of warm-up documents differs from the RR of field-test documents a bit more in the IT domain than in the legal domain, which is a first (weak) warning on whether to adopt IT warm-up texts as seeds for data selection. But the strong 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 en→fr/es pairs): the adapted IT language model improves PP of field-test document 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 well fitting the field test document, as proved by looking at the perplexity on language models estimated on selected text (PS). Moreover, by adding the warm-up document to selected text, the resulting language model (WU+PS) does not change 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 warm-up text (3.5 thousand words, see Table 2). These outcomes undoubtedly show that the warm-up documents well represent the field test texts of legal en→it/fr/es tasks, but not of IT tasks.

Concerning the legal English-to-German task, its adaptation scheme (Section 5.1) allows to compare only the language models of the domain-adapted and of the project-adapted engines. The perplexity of both the warm-up and the field-test documents does not improve significantly from the PA to the DA language model (from 169 to 161, and from 210 to 198, respectively). The DA system already fits well the translation project. This explains why the translations quality of the project-adapted engine improves less than for the other language pairs; data selected for project adaptation is only slightly more focused to the project than data selected for domain adaptation.

We have so explained the experimental results reported in Section 5.2: the reason for the lack of improvement of the IT adapted models over the domain-adapted baseline models is the mismatch between the adaptation text and the actual test document. Nevertheless, our adaptation method proved to be quite robust as in the worst case it does not affect the quality of the baseline.

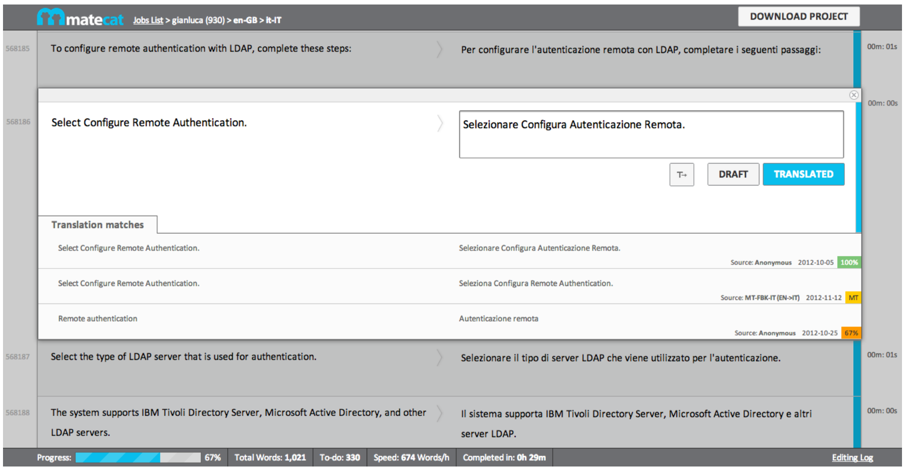
6 Field Test

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 Figure 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 machine-generated translations are presented, ranked on their quality; the quality is given by the fuzzy match value in case



Jobs List > gianluca (930) > en-GB > it-IT

568185 To configure remote authentication with LDAP, complete these steps: | Per configurare l'autenticazione remota con LDAP, completare i seguenti passaggi: | 00m: 01s

568186 Select Configure Remote Authentication. | Selezionare Configura Autenticazione Remota. | 00m: 00s

Select Configure Remote Authentication.	Selezionare Configura Autenticazione Remota.	Source: Anonymous 2012-10-05	100%
Select Configure Remote Authentication.	Seleziona Configura Remote Authentication.	Source: MT-FBK-IT (EN->IT) 2012-11-12	MT
Remote authentication	Autenticazione remota	Source: Anonymous 2012-10-25	67%

568187 Select the type of LDAP server that is used for authentication. | Selezionare il tipo di server LDAP che viene utilizzato per l'autenticazione. | 00m: 01s

568188 The system supports IBM Tivoli Directory Server, Microsoft Active Directory, and other LDAP servers. | Il sistema supporta IBM Tivoli Directory Server, Microsoft Active Directory e altri server LDAP. | 00m: 00s

Progress: 67% | Total Words: 1,021 | To-do: 330 | Speed: 674 Words/h | Completed in: 0h 29m

Fig. 2 The GUI of the MateCat tool. The central pane contains the current segments and translation matches supplied by the translation memory and the machine translation engine.

of suggestions coming from the TM, by a default value for MT generated suggestions. During the field test, the default MT quality was set to 85% by the organizers, resulting in a clear preference to edit MT suggestions (97-98% of the times).

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 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.

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 assigns a match percentage based on factors such as same words in the two segments and word order.

domain	user	TTE (sec/word)				PEE			
		warm-up	field-test	p-value	*Δ*	warm-up	field-test	p-value	*Δ*
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%

Table 6: 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). The difference of these measures achieved in the two sessions and its significance p-value are also reported.

6.3 Results

Table 6 reports results in terms of key performance indicators for all translators involved in the IT and legal, English-to-Italian 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 of four improved significantly both figures (t1 and t4), while on the legal domain this was the case for three 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 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% in PEE. Finally, the good correlation observed between PEE and TTE under the different conditions show that very likely the translators were able to took advantage of MT suggestions, and that the adapted MT engine suggestions were in general better. In fact, better PEE effort was observed for seven translators of eight.

7 Conclusions

In this paper we have faced a hot research topic for CAT industry: how to add self-tuning capability to SMT systems equipping CAT tools. Self-tuning can be seen at two different scales: at the domain level or simply at the project level. At the 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. At the 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 coming; this kind of self-tuning can be applied at any time, provided that enough fresh data is at disposal for updating the models according to the needs of the methods employed.

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

The collected experimental results proved 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 well represents the document to translate. In fact, where this condition is not satisfied, like in our IT tasks, adapted engines are unable to outperform the reference baseline systems; anyway, performance does not decrease, proving the conservativeness of the adaptation scheme.

Several issues remain open and deserve to be investigated in the future.

First of all, the prediction of the behavior of adapted models would be extremely important: is it possible and how to forecast if an adapted engine is effectively able to generate better suggestions than those of the reference system for a document whose source side is given?

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

Finally, gains observed in our 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, aware of this issue, in the MateCat field tests successive to that reported in this paper, those effects were mitigated by generating suggestions in field-test session by both the systems under investigation and the reference system, so that the net contribution of the tested methods could be measured.

Acknowledgements This work was supported by the MateCAT project, which is funded by the EC under the 7^{th} 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, UK, 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 MachineTranslation Systems. In: Proceedings of the Workshop on Statistical Machine Translation (WMT), Montréal, Canada, 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, France, 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, Soricut R, Specia L (2013) Findings of the 2013 Workshop on Statistical Machine Translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Sofia, Bulgaria, pp 1–44, URL `http://www.aclweb.org/anthology/W13-2201`

Chen SF, Goodman J (1999) An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language 4(13):359–393

Federico M, Bertoldi N, Cettolo M (2008) IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In: Proceedings of Interspeech, Melbourne, Australia, pp 1618–1621

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)

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

Foster G, Goutte 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

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, pp 439–448

Guerberof A (2009) Productivity and quality in MT post-editing. In: Proceedings of the MT Summit XII, Beyond Translation Memories: New Tools for Translators Workshop

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 (China), pp 268–275

Koehn P, Schroeder J (2007) Experiments in Domain Adaptation for Statistical Machine Translation. In: Proceedings of the Workshop on Statistical Machine Translation (WMT), Prague, Czech Republic, pp 224–227

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, Czech Republic, pp 177–180

Läubli S, Fishel M, Massey G, Ehrensberger-Dow M, Volk M (2013) Assessing Post-Editing Efficiency in a Realistic Translation Environment. In: Sharon O'Brien MS, (eds) LS (eds) Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, France, 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, Korea, 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, 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

Plitt M, Masselot F (2010) A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. Prague Bulletin of Mathematical Linguistics 93:7–16

Rousseau A (2013) Xenc: An open-source tool for data selection in natural language processing. The Prague Bulletin of Mathematical Linguistics 100(1):73–82

Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T, Tufiş D, Varga D (2006) The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In: In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 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, Turkey, pp 2214–2218

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, India