Continuous Adaptation to User Feedback for Statistical Machine Translation

Frédéric Blain, Fethi Bougares, Amir Hazem, Loïc Barrault and Holger Schwenk
LIUM - University of Le Mans (France)
firstname.surname@lium.univ-lemans.fr

Abstract

This paper gives a detailed experiment feedback of different approaches to adapt a statistical machine translation system towards a targeted translation project, using only small amounts of parallel in-domain data. The experiments were performed by professional translators under realistic conditions of work using a computer assisted translation tool. We analyze the influence of these adaptations on the translator productivity and on the overall post-editing effort. We show that significant improvements can be obtained by using the presented adaptation techniques.

1 Introduction

Language service providers (LSP) and human professional translators currently use machine translation (MT) technology as a tool to increase their productivity. For this, MT is closely integrated into computer-assisted translation (CAT) tool. The MT system suggests an automatic translation of the input sentence which is then post-edited by the human professional translators. They generally work on a project-based pace, *i.e.* a set of documents (the project) have to be translated in a certain period of time. It is well know that an MT system has to be adapted to the target task and domain in order to achieve the best performances. This process of adaptation can be separated into two different steps. First, an adaptation is performed before the beginning of the translation process,. This aims to specialize the system to the targeted domain: we will to this adaptation as *domain adaptation*.

Then, another adaptation is performed during the translation process with the aim of iteratively integrating users' feedback into the MT system. The adaptation can be performed at two different frequencies: (i) the system can continuously learn from post-edited segments, the models being immediately updated, or (ii) all the available project-specific data is used after each day of work to adapt the MT engine. This scheme is more related to document level adaptation; we will refer to it as *project adaptation*. The experimental work described in this paper fits into the latter adaptation scheme.

As part of the MATECAT project[1], we analyze project adaptation performed over several days. All experiments were performed with professional human translators under realistic conditions of work. The motivations of this work are detailed in section 2 and related work is discussed. In sections 3 and 4 we present both the experimental protocol and framework before presenting the corresponding results in section 5.

2 Motivations

This work is a continuation of earlier research on adaptation of a statistical MT (SMT) system Cettolo et al., 2014). More precisely, it was motivated by remaining opened questions. First, what does the learning curve look like for an iterative usage of the daily adaptation procedure? Even if the efficiency of the project adaptation scheme has been established, it has not been tested yet over multiple days. Does it reaches a plateau or do the translation

[1] www.matecat.com

quality continue to improve? What are the causes for the observed gains? Are they due to the familiarization of the users with both the system and the task, or are they due to real efficiency of the adaptation scheme? In previous work, the protocol did not allow to clearly measure the adaptation performance. In order to avoid this issue, a specific experimental protocol has been defined as described in section 3. Moreover, in addition to answer these new questions, we assessed a project adaptation scheme which take advantage of continuous space language modeling (CSLM) as explained in section 4. As far as we know, this is the first time that a neural network LM is integrated into a professional environment workflow, and that adaptation in such an approach is considered.

3 Evaluation Protocol

We defined an adaptation protocol with the goal to assess the same task with and without adaptation procedure. Like in (Guerberof, 2009; Plitt and Masselot, 2010), three professional translators were involved in a two parts experiment: during the first part, translators receive MT suggestions from a state-of-the-art domain-adapted engine built with the Moses toolkit (Koehn et al., 2007), without being adapted with the data generated during the translation of the project.For the second part, the MT suggestions are provided by a MT system which was previously adapted to the current project using the human translations of prior working days. Since we asked the same translators to post-edit the same document twice (*i.e.* with and without MT adaptation), the second run was launched after a sufficient delay: the human memory impact is reduced since translators worked on other projects in between.

To measure the user productivity, we considered two performance indicators: (i) the post-editing effort measured with TER (Snover et al., 2006) which corresponds to the number of edits made individually by each translator, (ii) the time-to-edit rate expressed in number of translated words per hour. In addition to these two key indicators, we evaluated the translation quality using an automatic measure, namely BLEU score (Papineni et al., 2002). This measure is used to make sure that no regression in the translation quality is observed after several days of work due to overfitting of the project adaptation (since previous working days are used to adapt the models).

Moreover, in order to respect realistic working conditions, we decided to set up a unique user-specific Moses engine per translator. By these means, any inter-user side-effects due to personal choices or stylistic edits are avoided. In addition, we obtain multiple references for assessing the results of the test. Consequently, it was required for the assessment that human translators work in a synchronized manner, *i.e.* the same amount of data is translated every day by each translator. The systems are then adapted, individually for each translator, using previous days of work, and used by the translators during the next day, and so on.

4 Experimental framework

We ran contrastive experiments by asking the translators to post-edit translations of a Legal document from English into French (about 15k words) over five days (*i.e.* about 3k words/day). An in-domain adapted (DA) system was used as baseline system for the first day, before project adapted (PA) systems have taken over.

4.1 Domain adapted system

Before the human translator starts working, our DA system is trained using an extracted subset of bilingual training data that is mostly relevant to our specific domain. The extraction process, widely known as *data selection*, is applied using cross-entropy difference algorithm proposed by (Axelrod et al., 2011)[2]. In order to augment the amount of training data[3] (about 22M words) we also select a bilingual subset from *Europarl*, *JRC-Acquis*, *news commentary*, *software manuals of the OPUS corpus*, *translation memories* and the *United Nations corpus*. About 700M additional newspaper monolingual data selected from WMT evaluation campaign are also used for language modeling.

4.2 Project adapted system

Our project-adaptation scenario, which is repeated iteratively during the lifetime of the translation

[2]We used the XenC tool for data selection

[3]DGT+ECB corpora (see http://opus.lingfil.uu.se)

project, is achieved as follows: the new daily amount of specific data is added to the development set, and new monolingual and bilingual data selections are performed with it. The new SMT system built on these selected data is then optimized on the new development set.

When performing project adaptation of an SMT system, we assume that the documents of a project are quite close and then, adapting the SMT system using the n-th days could be helpful to translate the $n+1$ day. However, we need to be careful to not overfit to a particular day of the project. This is particularly risky since the daily amount of specific data is relatively small (about 3k words). Therefore, we chose to add three times the daily data to our existing in-domain development set. This factor of three was empirically determined during prior lab tests. Also, all the previous days are used, *i.e.* when we adapt after three days of work, we used all the data from the first three days.

4.3 Continuous Space Language Model

Over the last years, there has been significantly increasing interest in using neural networks in SMT. As mentioned above, we used this technology into our project adaptation scheme. Fully integrated to the MT systems, it was used by our three SMT systems dedicated to the translators.

A continuous space LM (CSLM) (Schwenk, 2010; Schwenk, 2013) is trained on the same data than a classical n-gram back-off LM and is used to rescore the n-best list. In our case, and after each day of work, the daily generated data (3k words) is used to perform the adaptation of the CSLM by continuing its training (see (Ter-Sarkisov et al., 2014) for details). An important advantage of this approach is that the adaptation can be performed in a couple of minutes.

5 Experimental results and discussion

All the results presented in this section have been extracted from the edit logs provided by the MATECAT CAT tool.

5.1 Post-editing effort

In terms of post-editing effort, the results for each translator according to several SMT systems are shown in Table 1. Several TER scores are computed between the SMT system output and various sets of references. This score reveals the number of edits performed by the translator in order to obtain a suitable translation. The first column indicates the day of the experiment. The second column represents three SMT systems, namely: the baseline system adapted to the domain (DA), the same system with a CSLM (DA+CSLM) and the project adapted system (all models were updated, including the CSLM) noted "PA+CSLM-adapt". The third, fourth and fifth columns represent respectively the TER scores for the three translators. The first score is calculated using the reference produced by the translator himself. It could be considered as HTER (Snover et al., 2009). The second score (in parenthesis) is calculated using the three references produced by the translators. The third score (in brackets) is calculated according to an official "generic" reference provided by the European Commission. By these additional results, we aim to assess whether their is a tendency of the systems to adapt strongly to the particular style of one translator, or whether they still perform well with respect to independent references. On day 1, only the DA and DA+CSLM systems are presented since the project adaptation can only start after the first working day.

First of all, we can notice that the use of CSLM significantly decrease the TER scores for all translators. We can also remark that the third translator has a much higher TER than the two other translators during the first two days. Then, the system seems to learn his style and the TER reaches a comparable level at day 3. We can observe that project adaptation always lowers the TER with respect to the individual reference. The only exception can be observed for the first translator for days 2, 4 and 5. However, the project-adapted system is better or identical in most cases when multiple references are used. It is also interesting to note that our adaptation procedure improves the post-editing effort with respect to the independent reference translation in nine out of twelve cases. Overall, it can be clearly seen that the adaptation scheme is very effective. The difference between the baseline system (DA+CSLM) and the fully adapted system (PA+CSLM-adapt) reaches 9 TER points in some conditions.

A quite similar tendency can be observed when

day	method	translator 1	translator 2	translator 3
1	**DA**	33.34 (28.10) **[54.59]**	32.99 (28.10) **[54.59]**	48.62 (28.10) **[54.59]**
	DA+CSLM	**31.13 (25.73)** [54.94]	**31.43 (25.73)** [54.94]	**48.50 (25.73)** [54.94]
2	**DA**	35.33 (30.73) [56.63]	37.44 (30.73) [56.63]	49.03 (30.73) [56.63]
	DA+CSLM	**33.06 (28.86)** [56.30]	36.24 (28.86) **[56.30]**	49.12 **(28.86) [56.30]**
	PA+CSLM-adapt	34.31 (29.07) **[56.18]**	**30.48 (27.21) [56.30]**	**47.29** (29.62) [56.53]
3	**DA**	30.76 (26.68) [55.49]	35.09 (26.68) [55.49]	38.05 (26.68) [55.49]
	DA+CSLM	27.87 (24.70) [55.09]	33.86 (24.70) [55.09]	36.72 (24.70) [55.09]
	PA+CSLM-adapt	**25.24 (20.04) [54.13]**	**27.48 (20.40) [54.16]**	**27.42 (20.99) [53.77]**
4	**DA**	33.01 (29.07) [55.90]	38.31 (29.07) **[55.90]**	41.96 (29.07) [55.90]
	DA+CSLM	**29.79** (27.12) [56.78]	37.92 (27.12) [56.78]	41.03 (27.12) [56.78]
	PA+CSLM-adapt	30.47 **(25.87) [55.21]**	**30.15 (25.53)** [56.12]	**32.70 (24.03) [55.86]**
5	**DA**	31.34 (26.31) [54.78]	34.38 (26.31) [54.78]	39.41 (26.31) [54.78]
	DA+CSLM	**29.52** (24.88) **[52.59]**	33.94 (24.88) [54.74]	38.85 (24.88) [54.74]
	PA+CSLM-adapt	31.52 **(24.43)** [53.08]	**26.19 (22.34) [53.16]**	**30.46 (23.71) [54.31]**

Table 1: TER scores for English-French data of the Legal domain for the three translators over 5 days. Parenthesized scores are calculated using the references of all three translators, while scores in brackets are calculated using a generic reference provided by the European Commission.

analyzing translation quality in terms of BLEU score (results not presented here). Like for the prior TER results, the BLEU scores for translator 3 are much worse than the scores of the two other ones. After the third day, the scores reach the same level. Again, this could indicate that the adaptation process has learned his particular style.

5.2 Translation speed

Table 2 reports, for each translator, the translation speed, expressed in number of post-edited words per hour. The results are given for the two conditions of our experiment, along with the percentage of relative improvement. We can observe a very high productivity gain for all translators between the two sessions of our test, from 18.5% to 38.3%. The huge gain for translator T3 could be biased by the low working speed of the translator, even if we had confirmed that all the translators are experts with the post-editing process. We assume that either T3 had some difficulties with the legal domain or he had just taken his time to perform the test, or both. This could partially explain the huge improvement in productivity which is doubled.

User ID	Translation speed (words/hour)		
	DA+CSLM	PA+CSLM-adapt	Δ
T1	928	1283	38.3%
T2	1533	1816	18.5 %
T3	308	704	128.5%

Table 2: Overall translation speed for all translators. Measurements are taken on post-edits performed with the domain-adapted MT system (DA+CSLM) and the project-adapted MT system (PA+CSLM-adapt).

6 Conclusion

Several studies have also shown that the close integration of MT into a CAT tool can increase the productivity of human translators. In this work, we extended these works in several aspects. We have observed systematic improvements of the translation quality and speed when adapting the systems with data generated during the translation project (spanning several days). The MT system does not only adapt to the style of the human translator who post-edit the automatic translations. In all cases, we observed improved translation quality with respect to an independent reference translation. Finally, we have shown that neural network LMs can be used in an operational SMT system and that they can be adapted very quickly to small amount of data. Although the use of neural networks in SMT is drawing a lot of attention, we are not aware at any other

deployment in real applications.

Acknowledgments

We thank the post-editors who took part to this experiment, as well as our anonymous reviewers for their feedback and suggestions. This work has been partially supported by the EC-funded project MATECAT (ICT-2011.4.2-287688).

References

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

Cettolo, M., Bertoldi, N., Federico, M., Schwenk, H., Barrault, L., and Servan, C. (2014). Translation project adaptation for mt-enhanced computer assisted translation. *Machine Translation*, 28(2):127–150.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93(-1):7–16.

Schwenk, H. (2010). Continuous space language models for statistical machine translation. In *The Prague Bulletin of Mathematical Linguistics*, number 93, pages 137–146.

Schwenk, H. (2013). Cslm - a modular open-source continuous space language modeling toolkit. *Interspeech*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

Ter-Sarkisov, A., Schwenk, H., Bougares, F., and Barrault, L. (2014). Incremental adaptation strategies for neural network language models. *Available at http://arxiv.org/abs/1412.6650*.