Automatic post-editing

Kuhn, Roland; Isabelle, Pierre; Goutte, Cyril; Senellart, Jean; Simard, Michel; Ueffing, Nicola

Publisher's version / Version de l'éditeur:

Multilingual, 21, 1, pp. 43-46, 2010-03-01

Recent Advances in Automatic Post-Editing

Roland Kuhn, Pierre Isabelle, Cyril Goutte, Jean Senellart, Michel Simard, and Nicola Ueffing.

The term "post-editing" is often used to refer to the correction by a human being of translations output by a machine translation (MT) system. Although there are some documented cases of success, MT followed by human post-editing has never become a mainstream practice among professional translators. Most translators consider that MT output is more of a hindrance than a help to their work, because it contains too many errors. They find it highly frustrating that MT systems fail to learn from their tedious post-editing work and keep repeating the same mistakes over and over again.

Translators who do use MT followed by post-editing typically try to reduce the number of errors by customizing the MT system to the particular translation domain. Today's MT systems can be divided into two classes, "rule-based machine translation" (RBMT) systems and "statistical machine translation" (SMT) systems. Most commercial MT offerings are RBMTs; they are usually provided in a generic version that can be used to translate in any domain. However, the performance of the generic version within a specialized domain will typically not be impressive. The user can obtain better performance by adapting the RBMT to the domain, typically by providing a domain-specific dictionary.

The arrival of commercial SMT systems offers another possible solution to the problem: since SMT systems learn directly from existing translations, they can be automatically customized to new domains. However, this depends on having available a large amount of training data consisting of existing translations; even if such data are available, the MT output may still fail to reach sufficiently high quality.

A third possible solution – at first sight, a paradoxical one – is to have a first MT system produce an initial solution which is then post-edited by a second system. The post-editing in this case is called "automatic post-editing" (APE). The first mention of this idea we have found is in [Knight and Chander 1994], though these authors did not implement the idea. Subsequently, [Allen and Hogan 2000] also suggested a version of the idea. [Elming 2006] used a technique called "transformation-based learning" to correct the output of an RBMT system called Patrans and reported a significant improvement in translation performance.

We have implemented a version of APE in which the system that does the post-editing of MT output is itself an MT system. In our work, the system that produces the initial MT output is an RBMT system; the system that does the post-editing is also an MT system, but one based on SMT. This is shown in **Figure 1**, with T_1, T_2, … denoting the initial translation into the target language and T_1', T_2', … denoting the translations that result from APE. Consider the task of translating from French to English. The RBMT system produces an initial translation into English from the source document. In the second system (based on SMT), this initial English translation is treated as if it were a foreign language that must be translated into real English. The hope is that by "translating"

RBMT output English to English, the SMT system will correct some of the errors in the translation generated by the RBMT system.

As shown in **Figure 2**, we tried this approach out in two different kinds of setting. In the first setting, we train the SMT post-editor on a corpus consisting of initial output from the RBMT system on a given source-language text in parallel with a version of this output that has been corrected by human translators. In the second setting, we train the SMT post-editor on a corpus consisting of initial output from the RBMT system on a given source-language text in parallel with translations of the same source-language text independently produced by human translators. The first setting is obviously preferable, since the SMT post-editor has a chance to learn from corrections made to output from the RBMT system; however, it's hard to get hold of corpora of post-edited MT output. The second setting does not pose this practical problem: one creates a training corpus by finding a source-language corpus that has already been translated by human beings and then running it through the RBMT to obtain the RMBT output. However, at first glance the second setting looks much less likely to produce useful results. In carrying out our experiments, we used our SMT system, PORTAGE; PORTAGE is a "phrase-based" SMT system that is described in more detail in [Sadat *et al*. 2005]. Surprisingly, we obtained good results for automatic post-editing in both settings.

Our experiments in the first setting are described in [Isabelle *et al*. 2007] and [Simard *et al*. 2007a]. They were carried out with data from the Canadian government's department of Human Resources and Social Development (HRSDC). This department maintains a web site called "Job Bank" (www.jobbank.gc.ca) where potential employers can post ads for open positions; over one million ads are posted on the site each year, totaling more than 180 million words. By law, each ad must be posted in both English and French, which means that ads submitted in English must be translated into French, and vice versa. Employers often post ads that are either identical or very similar to previous ads; these are handled by consulting a database of previous ads or are translated by human translators using a translation memory. The remainder consists of about a third of the ads submitted in either language. These remaining ads are translated by an RBMT system. The output of the system is then post-edited by a human; HRSDC employs as many as 20 post-editors working full-time. There are two versions of the RBMT: a generic version and a version with a dictionary customized to the Job Bank task; this dictionary contains about 18,000 entries and took approximately 18 person-months to construct. One objective of our experiments was to determine whether an SMT-based post-editor could make construction of a customized dictionary unnecessary.

HRSDC kindly provided us with English-to-French and French-to-English data from the Job Bank, consisting of parallel "blocks" of text each containing: 1. text in the source language, S. 2. a translation T1 of S produced by the generic RBMT system. 3. a translation T2 of S produced by the version of the system with the customized dictionary. 4. a reference translation T_R that had been manually post-edited. For English-to-French experiments, we picked a subset of about 1,000 blocks (9,700 English words) as test data and another subset of about 30,000 blocks (321,000 English words) as training data to estimate our post-edition model. For the French-to-English experiments we again picked

a subset of about 1,000 blocks (13,500 French words) as test data, and a block of about 37,000 blocks (509,000 French words) as training data. The main metric we used for evaluating the quality of post-editing was the Translation Edit Rate (TER) [Snover *et al*. 2006]. TER reflects the number of edit operations (inserting, deleting, or substituting words or shifting blocks of words) needed to change a hypothesized translation into the reference translation as a proportion of the length of the reference. Lower TER indicates lower post-editing effort and suggests better translation quality.

The most interesting experiments involved training the APE system to post-edit translations T2 produced by the version of the RBMT system with the customized dictionary. For English to French test data, the TER was 53.5% for T2, meaning that slightly over 53% of the words coming from the customized English-to-French system needed to be changed. For French to English experiments, the TER was 59.3% for T2, meaning that nearly 60% of the words coming from the customized French-to-English system had to be changed. When we trained the APE system to post-edit T2 output, the TER dropped dramatically: to 47.3% in the English to French case (a drop of 6.2% TER) and to 41.0% in the French to English case (a drop of 18.3% - nearly a third). Thus, APE can significantly reduce the amount of subsequent human post-editing needed.

To see if it might be possible to dispense with the RBMT system altogether, we tried training the SMT system to translate directly from English to French and from French to English. But standalone SMT proved to be inferior to the RBMT + APE combination, with a TER of 53.7% for English to French, and 43.9% for French to English. Apparently, the RBMT system contributes useful information that was not learned by the standalone SMT system (note, however, that this conclusion may depend on the amount of data available for training the SMT system)

Finally, we looked at what happens when we train our SMT system to post-edit output of the generic system (T1). In this case, the TER was only very slightly worse than for APE on output of the customized system (48.6% instead of 47.3% for English-to-French and 41.5% instead of 41.0% for French-to-English). Thus, it seems that the heavy cost in person-hours required for customization may be unnecessary: one can train an SMT post-editor on manually post-edited translations from a generic system instead.

We now turn to the second setting for our APE experiments. In this setting, the SMT-based APE is trained on sentence pairs such that one member of the pair is a target-language translation of a source-language sentence produced by an RBMT system, and the other member of the pair is a target-language translation of the same sentence independently produced by a human translator. We have explored this setting with two different language pairs: English-French (in both directions) and Chinese-to-English translation. Our goal has been to find out whether RBMT followed by APE using an SMT system (PORTAGE) is superior to either RBMT alone or SMT alone. For both language pairs, the RBMT component was provided by Systran (Systran was not the RBMT system used in the Job Bank experiments). Although it is possible to customize the Systran system by providing a specialized dictionary, we did not rely on this feature – we used Systran in its “out of the box” configuration.

The English-French work was part of an evaluation associated with the Second Workshop on Statistical Machine Translation – see [Callison-Burch *et al*. 2007] and [Simard *et al*. 2007b]. An advantage of participating in this evaluation was that system outputs were evaluated not only by automatic metrics, but also by human evaluators. One of the most interesting results from this evaluation involves a corpus called Europarl derived from the proceedings of the European parliament and containing nearly 1.3 million sentence pairs. Some results for this corpus are given in **Table 1**:

System	Eng→Fr: (*BLEU*) **Rank/8**	Fr→Eng: (*BLEU*) **Rank/7**
Systran	(*23.3*) **6**	(*21.1*) **6**
PORTAGE	(*29.4*) **5**	(*31.2*) **5**
Systran→PORTAGE	(*30.1*) **1**	(*31.3*) **2**

Table 1: Europarl BLEU score and human rankings for Eng→Fr and Fr→Eng systems

The table shows two different kinds of scores for the three systems (the RBMT Systran system, the SMT PORTAGE system, and the hybrid system where PORTAGE post-edits Systran output). The numbers in brackets and italics are the BLEU scores. BLEU is an automatic metric for evaluating MT systems – the higher the value of BLEU, the better the system [Papineni *et al*. 2001]. The numbers in bold show how human evaluators ranked a system compared to other MT systems in the evaluation. According to the automatic BLEU metric, the quality of the hybrid system's output is only slightly better than that of the pure SMT system (PORTAGE). Human evaluators disagree: they rank the hybrid system as the best of the eight systems competing on the English-to-French task and second of the seven systems competing on the French-to-English task, while they give the pure RBMT system and the pure SMT system worse rankings. Thus it seems there is a complementarity between the RBMT system and the SMT post-editor: together, they make a much more favourable impression on human evaluators than either alone.

We also explored the application of the APE idea to Chinese-to-English translation. This is a very challenging language pair, as the two languages involved are completely unrelated and have very dissimilar properties. To train the Chinese-English system, the Chinese portion of a huge bilingual corpus (8.8 million Chinese sentences aligned with their English translations, comprising 278 million English words) is translated into English by the Systran system. The result is a corpus consisting of "Systran English" sentences in parallel with English sentences; this parallel corpus is used to train the SMT-based APE. Although we have not yet carried out a formal human evaluation of the resulting hybrid system, we are very pleased with some of the output we have seen. In many cases, the hybrid system succeeds in producing understandable, fluent output in cases where the two "parent" systems – the rule-based Systran system and the pure SMT system PORTAGE – have produced output that is difficult to read.

For instance, a certain Chinese sentence is translated by Systran as "What the EU officials most were worried now is soon the migratory bird which flies back from

Africa". PORTAGE translated the same sentence as "EU officials are most concerned about is coming from Africa flew back to the migratory birds" which is incomprehensible. However, when the Systran output was post-edited by PORTAGE, the result was "The EU officials are most worried about the migratory birds that fly back from Africa" which is a perfect translation of the original Chinese sentence. What seems to have happened in this case, and in many similar cases we have seen, is that the Systran system's deeper knowledge of syntax is complemented by PORTAGE's extensive statistical knowledge of the surface forms of English sentences.

The experiments described above show that automatic post-editing (APE) by an SMT system of output from an RBMT system can produce translations that will be superior to those produced by either "parent" system. We have also shown that in this setup, the APE system does not need to be trained on output of RBMT that has been manually post-edited – one can train the APE system on RBMT translations in parallel with independently produced human translations of the same source sentences.

We recently learned of an application of the ideas described above to a different pair of systems and targeting a different language pair. Inspired by [Simard *et al*. 2007a] and [Simard *et al*. 2007b], a group of researchers at BBN Inc., Sakhr Inc., and MIT built an Arabic-to-English system in which the Sakhr RBMT system produces an initial English translation of the Arabic input; this initial translation is then converted into better English by an SMT system from BBN called HierDec. As in our experiments, the resulting hybrid system yields output at least as good as the pure SMT system; there is also evidence that the hybrid system may handle some phenomena better than the SMT system. This research has not yet been published.

Meanwhile, Systran has recently released its first commercial hybrid MT system, based on a first pass of RBMT followed by SMT post-editing, as explained above. As the Systran site describing this product, http://www.systransoft.com/translation-products/server/systran-enterprise-server, notes: "it combines the strengths of rule-based and statistical machine translation. It merges the predictability and language consistency of rule-based MT with the fluency and flexibility of statistical MT." One of Systran's largest customers, Symantec, has been exploring the application of APE techniques to its translation needs. In a recent presentation, a researcher from Symantec concluded that for some of the company's large production projects, the approach has yielded a good return on investment http://summitxii.amtaweb.org/summitxii-keynote-roturier.pdf .

We have now begun to explore deeper forms of integration between RBMTs and SMTs. For instance, the RBMT system's output can be divided up into subsentential chunks, each annotated with a confidence level; the SMT system might post-edit only chunks that have lower confidence. In cases where the RBMT system had particularly low confidence, it might provide a number of alternative translations and allow the SMT post-editor to choose between them. Our experiments have already shown that this approach yields even better performance than the original Systran-PORTAGE combination. Finally, another idea we are considering is to have the SMT system "look" at the original

source-language text, and combine its predictions with the predictions of the RBMT layer.

References

[Allen and Hogan 2000] Jeffrey Allen and Christopher Hogan. "Toward the development of a post-editing module for Machine Translation raw output: a controlled language perspective." In *Third International Controlled Language Applications Workshop (CLAW 2000)*, Washington DC, USA, 2000.

[Callison-Burch *et al*. 2007] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. "(Meta-) Evaluation of Machine Translation." In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007.

[Elming 2006] Jakob Elming. "Transformation-based corrections of rule-based MT." In *Proceedings of the EAMT 11th Annual Conference*, Oslo, Norway, 2006.

[Isabelle *et al.* 2007] Pierre Isabelle, Cyril Goutte, and Michel Simard. "Domain adaptation of MT systems through automatic post-editing." In *MT Summit XI*, Copenhagen, Denmark, 2007.

[Knight and Chander 1994] Kevin Knight and Ishwar Chander. "Automated postediting of documents." In *Proceedings of National Conference on Artificial Intelligence*, Seattle, USA, 1994.

[Papineni *et al*. 2001] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a Method of Automatic Evaluation of Machine Translation". In *Proceedings of the ACL,* Philadelphia, USA, 2001.

[Sadat *et al.* 2005] Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Roland Kuhn, Joel Martin, and Aaron Tikuisis. "PORTAGE: a phrase-based machine translation system." In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, USA, 2005.

[Simard *et al*. 2007a] Michel Simard, Cyril Goutte, and Pierre Isabelle. "Statistical phrase-based post-editing."In *Proceedings of HLT-NAACL*, Rochester, NY, USA 2007.

[Simard *et al*. 2007b] Michel Simard, Nicola Ueffing, Pierre Isabelle and Roland Kuhn. "Rule-based translation with statistical phrase-based post-editing." In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007.

[Snover *et al*. 2006] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. "A study of translation edit rate with targeted human annotation." In *Proceedings of AMTA*, Cambridge, USA, 2006.