Preliminary Experiments on Using Users' Post-Editions to Enhance a SMT System

Marion Potet[1], **Emmanuelle Esperança-Rodier**[1],
Hervé Blanchon[2], **Laurent Besacier**[1]
[1]UJF-Grenoble 1 ; [2]UPMF-Grenoble 2
LIG UMR 5217, Grenoble, 38041, France
FistName.Name@imag.fr

Abstract

The aim is to develop an interactive and iterative translation system able to adapt and improve itself rapidly using user's feedbacks (collected as translation post-editions). We present a preliminary study in which three methods for integrating feedbacks are tested. The results show that, despite the difficulty to measure the system improvement with automatic metrics (given the small size of our collected post-edition corpus) a small-sized subjective analysis reveals the contribution and prospects of such methods.

1 Introduction

New challenges and uses have emerged thanks to the development of the collaborative Web. The Wikipedia encyclopedia, the Google Image Labeler, the Amazon Mechanical Turk platform or Waze are well-known examples. The aim of these collaborative "services" is to solicit input from users to create content, solve problems and enrich database. Thus, in the field of Machine Translation, the same idea has recently appeared in order to overcome the current flaws of MT systems by using the users themselves and enhance them as time goes on.

There are two typical scenarios of the use of MT, namely assimilation (to get the gist) and dissemination. The second scenario is seen as a scenario for professional translators (it requires post-edition and revision). Our long-term goal is to propose post-edition in both scenarios in order for unskilled users to obtain better translations taking advantage of skilled users' post-editions that will further improve the MT outputs. To sum up, we argue that MT systems for both assimilation and dissemination can "learn" from their users' knowledge in order to improve themselves overtime through post-edition. For assimilation, post-editors would be either native speakers of the target language with a good knowledge of the source language or native speakers of the source language with a good knowledge of the target language. For dissemination, we aim at "dissemination for every one", post-editors would be professionals or have a good knowledge of the target language and MT would be used to improve productivity.

In this paper, we experimented with collecting post-editions from native speakers of the source language with a good knowledge of the target language and integrating these contributions into the system in order to improve it. We have initially created a statistical machine translation system (Section 3.1). Then, we have collected translations correction (post-editions) made by volunteer annotators (section 3.2). Finally, we propose and experiment with several methods, to use these post-editions to improve our system: adding data to the training corpus of the system (section 4.1); automatically correcting the outputs of the system (section 4.2); and re-tuning the weights of the log-linear model (Section 4.3).

2 Context

2.1 Related Work

Statistical machine translation (MT) techniques regained interest in the 1990s (Brown et al., 1990). Since then, the evolution of the models, in particular the phrase-based approach (Marcu and Wong, 2002; Koehn et al., 2003), and the proliferation of parallel corpora have pushed these approaches to the forefront of the MT research. Although the fully automatic approach has shown some effectiveness, the tools are intrinsi-

© 2011 European Association for Machine Translation.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 161–168
Leuven, Belgium, May 2011

cally limited and the quality of the translations produced is generally deemed insufficient to be used as is. Therefore, in recent years, many solutions have been proposed to integrate the human in the SMT framework (human in the loop).

In this trend, for example, active learning methods (Callison-Burch, 2003) were proposed. In practice, the idea is to select the most relevant examples and integrate them into the system in order to improve it for a given task. While recent studies have demonstrated the value and effectiveness of active learning for several applications, in order to reduce the annotation cost without damaging the quality, this technique is rarely used despite huge efforts spent annotating big corpora (Tomanek and Olsson, 2009).

Another way to reach the same goal (using human expertise to improve a system) is to use human post-editions. In the field of computer-assisted translation or for the evaluation of machine translation systems, post-editing is frequently used. However, to date there are few studies on the use of post-editions as feedback to correct and improve a machine translation system. For example, Simard et al. (2007a,b) used manual post-editions of machine translation proposals to create an automatic post-editor which can "fix" the system outputs, and the FAUST project (User Feedback Analysis for Adaptive Statistical Translation) organizes the collection and analysis of users' feedbacks from the online Reverso translation service in order to develop techniques to exploit them (Déchelotte, 2010).

2.2 Foreseen Scenario

In these experiments we would like to build an SMT system, available online for example, which allows the native speakers of the source language with a good knowledge of the target language (1) to input a text to be translated; (2) to get a translation output; (3) to post-edit the output. The post-edition is subsequently taken into account and used by the system to improve itself.

In order to implement this scenario, the first step is to propose an efficient technique to integrate human post-editions into the system in order to improve it. The aim of this study is to:

- create a state-of-the-art phrase-based SMT system;
- allow the user to post-edit a small set of hypotheses generated by our translation system;
- propose solutions to take into account those manual post-editions to improve the system.

3 Translation system and collection of post-editions

Initially, we created a state-of-the-art SMT System and collected a corpus of 175 post-edited segments.

3.1 Baseline translation system

The choice of the topic area has been motivated by the desire to create a generic system in order to validate the proposed methods: the system perform translations of "news" from French into English (Potet et al., 2010).

Our baseline system is a phrase-based SMT where translation units are segments (sequence of *n* consecutive words). First, the corpora were word aligned with the GIZA++ toolkit (Och and Ney, 2003) and then, the pairs of source and corresponding target phrases were extracted from the word-aligned bilingual training corpora using the scripts provided with the Moses decoder (Koehn et al., 2007). The result is a phrase-table containing all the aligned segments. This phrase-table, produced by the translation modeling, is used to extract the 14 default features functions combined in a weighted log-linear model. These weights can be adjusted with the Minimum Error Rate Training (MERT) strategy (Och, 2003) where an error criterion is minimized on a development corpus. Nevertheless, this was not done on our baseline system because this adjustment deteriorates our system's performance scores on WMT 2010 development data (consequently, no MERT is used in section 4.1 and 4.2).

This baseline, "state-of-the-art", system has been validated during our participation to the international evaluation campaign WMT 2010[1]. It is described in (Potet et al., 2010) as *system (3)*. The translation model was trained on the union of Europarl (≈46 MWords) and the News Commentary corpus (≈2 MWords) containing 1,640,463 utterances[2] preliminary normalized. The normalization phase includes case removal, tokenization and the transformation of the euphonious *t*. The target language model is a standard 4-gram language model trained using

[1] www.statmt.org/wmt10/translation-task.html

[2] We will call "utterance" the translation units of the corpora. An utterance is most often a sentence, but can also be a title or a set of two, or more rarely, three sentences.

the SRI Language Modeling toolkit (Stolcke, 2002) on a monolingual training corpus of 48M sentences. The smoothing technique we applied is the Kneser-Ney discounting with interpolation (Kneser and Ney, 1995). This system will be referred to as "reference system" or "baseline" in the following.

The parallel English/French corpora used are described in Table 1: the training data set is called TRAIN, the test set is called TEST and the post-editing set is called PE. TEST and PE are disjoint and contain utterances from several newspaper websites (Libération, Le Figaro, Les Echos, Etc..) translated by professional translators. This translation task is quite difficult because the domain of the training data (mostly parliament transcriptions) is different from the domain of the test (news). The given translation of each utterance will be referred to as the gold-standard translation.

Name	*Use*	*# of utterances*
TRAIN	learning corpus	1 640 643
PE	post-edition corpus	175
TEST	test corpus	2852

Table 1. Parallel corpora used

3.2 Collecting Users' Post-editions

In our scenario, the user feedback is provided as a post-edition of the system translation hypothesis, by a human annotator.

Instructions given to the post-editors Post-editing is an expensive task in terms of human time and effort. For a preliminary study, we chose to annotate a small corpus of 175 utterances (≈5,000 words). The PE corpus was translated by our baseline system and the translation hypotheses were post-edited with SECtra_w our in-house post-editing environment (Blanchon et al., 2009). Post-editors were asked to enhance the translation hypotheses with as few corrections as possible to produce a correct translation of the source. In some cases, no corrections were necessary. The post-editors were neither professional translators nor native speakers but they had a fairly good knowledge of English. This post-editors' profile fits with our scenario. This relies upon the fact that MT errors can be post-edited with high accuracy by non-expert speakers with little knowledge of the source language, as shown in (Llitjos and Carbonell, 2006).

Finally, we obtained a corpus of 175 triples:

- the translation hypothesis produced by the baseline system;
- its post-edition performed by a human annotator;
- the gold-standard translation given as a reference in the original corpus.

We call these three translations respectively: "System translation" (or trans.), "Post-edited translation" (or pe.), and "gold standard" (or std.). We denote similarly, $PE_{std.}$ the corpus of the baseline system translations aligned with the gold-standard translations and $PE_{pe.}$ the corpus of the baseline system translations aligned with the post-edited translations.

Examples of post-editions As we can see from the examples given in Table 2, the gold-standard translations prove to be very free translations, far from the source utterance (sometimes containing errors) while the post-edited translations are close to the system outputs and seem to be usable to correct the system outputs.

Source Utt.	*Gold-Standard*	*System Output*	*Post Edited Output*
Les pierres sont sales.	The stone **is** dirty.	The stones **are** vile.	The stones are **dirty**.
J'ai lu cela dans mes mangas.	I read **about this** in my manga.	I **have read it** in my **mangas**.	I have read **about this** in my mangas.
Il a eu de la peine pour obtenir un oscar.	**And awarding him with** an oscar **had been quite** hard	**It was** hard **to get** an oscar.	It was hard **for him** to get an oscar.

Table 2. Post-edition examples

Distance between the three families of translations For a given source utterance from the PE corpus, we have three different translations: the baseline system translation, the gold-standard reference translation, and the post-edited translation. We computed the BLEU score (Papineni et al., 2002), for each translation pair: this score may be interpreted as a proximity measure (between 0 and 100) between the translations. Figure 1 illustrates a distance as calculated by d = 100 – BLEU.

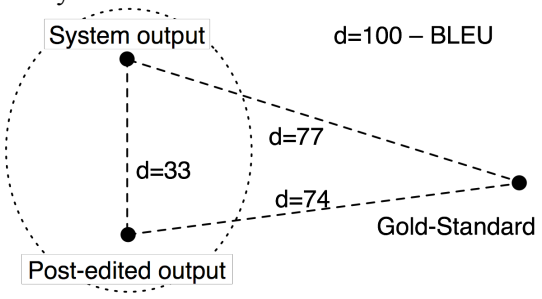

Figure 1. Distances between different kinds of translations

We found that the distance between the gold-standard and the baseline system translation is bigger than the one between the baseline system translation and its post-edition. Similarly, gold-standard and post-edited translations are dramatically remote while both are supposed to represent correct translations of a source utterance.

4 Using Post-Editions to improve statistical machine translation

We collected post-editions of SMT outputs from human users. We ended up with a preliminary corpus of 175 post-edited translations. In this section, we propose and evaluate ideas to exploit these corrected data into our SMT system in order to improve it. These data are used at three different levels of the translation process:

- to enrich the training corpus: we propose to add post-edited translations in the training corpus and re-train the SMT system;
- to automatically correct the system outputs: the corpus of post-edited translations is used here to train an automatic statistical post-editor (SPE)[3];
- to adjust the weights of the log-linear model: the idea is to use post-edited translations as references during the minimum error rate training (MERT) process.

4.1 Adding the Post-Edited Corpus to the MT Training Data

The idea is to add the post-edited translations (we call $PE_{pe.}$ the utterances of the PE corpus, which are post-edited translations of the SMT system output) to the SMT training corpus. The problem is that the amount of post-edited utterances is very small (175 Utts.) compared to the initial parallel corpus used for training (1,640,463 Utts.). To give more weight to these data, we duplicated them 10, 100 and 1000 times before adding them into the training corpus. In other words, we create a corpus gathering the training data of the reference system plus N times the post-edited data ($0 \leq N \leq 1000$). N may be interpreted as the weight given to the post-edited data.

The evaluation results of the different systems are given in Table 3. The column "different (≠) translations" shows the proportion of utterances whose translation produced by the system *Baseline* + $N*PE_{pe.}$ is different from the one produced by the reference *Baseline* system (corresponding to N=0). The log-linear weights are not changed; only the phrase-table is different. It is important to notice that if one adds the corpus of 175 post-edited utterances by duplicating it 1000 times (this is equivalent to increasing the training corpus by 11%), 90% of the PE corpus translations are modified, and the BLEU score goes from 23.50 to 25.73. This gain is significant with high certainty according to Koehn's (2004) resampling method. The gain is more modest but also observable on the TEST corpus, which moves from 25.27 to 25.51 BLEU with 65% of the translations being modified.

Thus, it seems that adding post-edited data, even in small quantity, to the training corpus, improves not only the performance if one re-translates the same utterances (results on the column PE in Table 3) but also improves the performance on new utterances (results on the column TEST in Table 3). Some examples of TEST utterances translated with the *Baseline* system versus the *Baseline* + $1000*PE_{pe.}$ are given in Figure 2. We can observe, in these examples, that the problem of unknown words (*in italics*) is obviously not solved by this method (which is not surprising given the size of the corrected corpus) but apart from that, some translations are more consistent with the source utterance.

4.2 Using the Post-edited Corpus to Train a Statistical Post-Editor

Another approach is to use the post-edited corpus to automatically correct system outputs. Re-

$PE_{pe.}$ Weight	# utt. of the learning corpus	PE corpus		TEST Corpus	
		≠ translations	BLEU score	≠ translations	BLEU score
0	1 640 463	0%	23.50	0%	25.27
1	1 640 638 (+0.01%)	85%	25.17	42%	25.28
10	1 643 213 (+0.1%)	86%	25.28	44%	25.30
100	1 657 963 (+1.06%)	90%	25.49	49%	25.38
1000	1 815 463 (+11%)	90%	25.73	65%	25.51

Table 3. Results of adding corrected data to the MT training data

[3] We are aware that the amount of data is small, but we believe it is worth to present our results, especially the obtained SPE phrase table.

Source Utterance:	Au terme des échanges, la bourse de Prague bascula dans le négatif.
Baseline :	In terms of trade, the stock market in Prague in the negative *bascula*.
+ 1 000 $PE_{pe.}$:	At the end of trade, the Prague Stock Exchange *bascula* in the negative.
Source Utterance:	Paulson : le plan doit être efficace.
Baseline :	Paulson 's plan is to be effective.
+ 1 000 $PE_{pe.}$:	Paulson : the plan must be efficient.
Source Utterance	On vous conseillera, comment choisir.
Baseline :	You can choose *conseillera*.
+ 1 000 $PE_{pe.}$:	We *conseillera*, how to choose.

Figure 2. Examples of translation from the TEST corpus : *Baseline* versus *Baseline + 1000*$PE_{pe.}$*

cently, statistical post-edition (SPE) has been used to improve MT output.

Statistical post-editing (SPE) Post-editing is originally a human task. A MT system translation is edited via an interface and then corrected by a human annotator. However, manual annotation is time-consuming and costly to implement. To facilitate this, the idea of SPE is to automate this task, i.e. to learn the behavior of human annotators to be able to automatically propagate their corrections on other translations. The post-editing task is then seen as a translation task between the raw outputs of a MT system and the corrections of these outputs. Some recent studies showed that statistical post-edition can be used for improving significantly a rule-based machine translation system (Dugast et al., 2009) (Lagarda et al., 2009) or for a domain adaptation task (Diaz de Ilarraza et al., 2008) but SPE has been little (if any) used as a post-processing of statistical machine translation.

Training the statistical post-editor We trained a translation model, on the PE corpus, from the hypotheses suggested by the translation system and the post-editions made by human annotators. We got a translation table of 22 938 segments. After removing from this table source segments that were identical to their correction (where the annotator has made no correction) we got 9523 parallel segments. Examples of segments obtained are shown in Figure 3.

This lookup table enables us to distinguish four different types of corrections made by the post-editors:

- Substitution (s1→s2): a sequence of words s1 is replaced by a new sequence s2 (examples 1, 3, 5, 4, 6 in Figure 3). This includes the translation by the post-editor, of a segment unknown by the system (examples 2, 7, 8 and 11 in Figure 3);
- Reordering (s1→s1'): the segment is present in both the hypothesis and its correction but not at the same position (10, 13 in Figure 3);
- Deletion (s→ø): deletion of a segment (example 9 in Figure 3);
- Insertion (ø→s): adding a segment (exemple 12 in Figure 3).

system's segment ||| *post-edited segment*
1 : demanded of personal ||| asked for personal
2 : survivante ||| survivor
3 : the city in the five lakes ||| the city with its five lakes
4 : fairness ||| global equity
5 : a command easy ||| an easy command
6 : a decline in sales of 16 % ||| a decrease in sales of 16 %
7 : troncs cerebral similar ||| similar brain stems
8 : cafards ||| cockroaches
9 : camera of moments of personality of ||| camera moments of personality of
10 : comparison nervous ||| nervous comparison
11 : fêtera ||| celebrate
12 : i live in rome for 25 ||| i have lived in rome for 25
13 : in the system of blood in quebec ||| in the blood system in quebec

Figure 3. Entries from the SPE phrase-table

Results obtained with the statistical post editor The next step is to apply the trained SPE for post-editing our MT outputs on both the PE and TEST corpora (Table 4).

While the results are logically improved when the same corpus is re-translated (PE from 23.5 to 24.58 with 85% of the utterances being translated differently), the SPE approach degrades the results on new data (TEST from 25.27 to 24.32

System	*PE Corpus*		*TEST Corpus*	
	≠ translations	*BLEU score*	*≠ translations*	*BLEU score*
Baseline	0%	23.50	0%	25.27
+ post-editions	85%	24.58	40%	24.32

Table 4. Automatic statistical post-editor results

Source utterance:	Les lecteurs, par contre, ont l' avantage d' avoir une commande facile.
Baseline :	The readers, however, have the advantage of having a command easy.
+ post-edition:	The readers, however, have the advantage of having an easy command.
Source utterance:	Le costume tyrolien bon marché semble atteindre son objectif.
Baseline :	The suit tyrolean cheap seems to achieve its goal.
+ post-edition:	The cheap tyrolean suit seems to achieve its goal.
Source utterance:	Les pierres sont sales.
Baseline :	The stones are vile.
+ post-edition:	The stones are dirty

Figure 4. Samples of automatically post-edited utterances of the PE corpus

with 40% of different outputs). To explain this, it is important to recall that the automatic post-editor has been trained on a corpus of only 175 utterances, which is clearly insufficient to effectively model the behavior of human post-editors. Therefore we are currently focusing on the collection of a corpus of several thousands of post-edited sentences.

Back to the PE corpus, although the improvement in terms of BLEU score may be seen as low, a brief manual analysis of the results shows that the automatic post-editor allows to correct the translation of some sentences which are presented again as input of the system (at least, we can consider that the system has learnt how to correct some of its previous errors). Figure 4 shows some examples to illustrate this.

4.3 Using the Post-Edited Corpus to Optimize the log-linear Weights

It has been often mentioned, in the past, that the MERT strategy to adjust the combination weights of a system is tricky and not always effective. This can be explained in particular by the distance between the translations produced by the system and the gold-standard translations usually given as a reference for tuning.

Our idea, here, is to replace (in the tuning corpus) the gold-standard references by the corrected translations obtained through post-editing, as the latter are "closer" to those of the system (as shown in Figure 2). We hope that this will make MERT converge faster while still providing adequate log-linear weights.

Using gold-standards versus post-edited data for MERT on the PE corpus The weights of the baseline model were optimized using MERT with, as references, either gold-standard translations ($PE_{std.}$) or post-edited translations ($PE_{pe.}$). A detailed analysis of the differences between the weights obtained after optimization on $PE_{std.}$ versus optimization on $PE_{pe.}$, shows that :

- penalties related to the number of words in the sentence (+ 6%) and the cost of reordering the words in the sentence (+ 7%) are bigger in the model optimized on post-editions ($PE_{pe.}$);
- there are significant changes in the weights of the 6 functions related to words reordering in the sentence (form +/- 4% to 11%) between the two models.

This can be interpreted by the fact that the post-edited translations are closer to the system outputs, compared to the gold-standard translations. As a result, they have globally the same number of words (cf. penalty on the number of words in the sentence) and the word order is usually kept between a source sentence and its translation.

We ran the deterministic MERT optimization process (mert-moses.pl and not mert-new-moses.pl). The optimization process converges faster and more efficiently with the use of post-edited translations ($PE_{pe.}$), as shown in Figure 5.

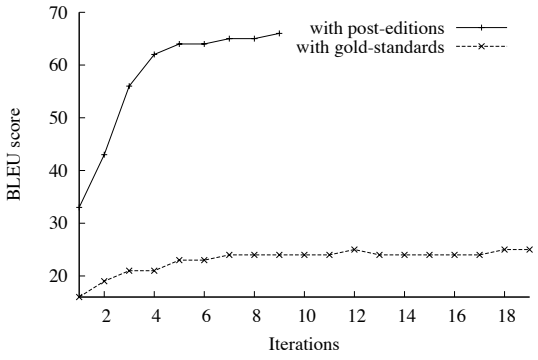

Figure 5. Evolution of the BLEU scores during weight optimization with MERT

BLEU score increases from 33 to 66 in 9 iteration, while using gold-standards lead to a small BLEU increase from 23 to 25 in 19 iterations. Using post-edited translations as reference during MERT seems to be a promising way to help the convergence, leaving room for the use of more advanced optimization methods in the future.

Automatic evaluation of systems on the TEST set according to the reference used for

MERT The two sets of weights, as well as the language/translation models were evaluated to translate the TEST set. The 2852 utterances were translated: first, with the system tuned using gold-standards ($PE_{std.}$) and, secondly, with the system tuned using corrected translations ($PE_{pe.}$). Although both models only differ in their log-linear weights, we end up with 65% of different MT hypotheses between the two systems.

However, the automatic evaluation results presented in Table 6 do not show any significant difference between the two optimization procedures. That being said, the results show that post-editing machine translation results can be an effective and relatively inexpensive way to build an in-domain development corpus in the case where no representative data is available for optimization.

Corpus	*Without optimization*	*With optimization*	
		On $PE_{pe.}$	*On $PE_{std.}$*
TEST	25.27	25.36	25.43

Table 5. Score on the TEST set according to the references used for MERT: ($PE_{std.}$) vs ($PE_{pe.}$)

Subjective evaluation of systems according to the reference used for MERT In order to better understand our results, we conducted a subjective evaluation that involved three voluntary evaluators. For each sentence to be translated, we asked the evaluators if they preferred the hypothesis given by the system optimized using $PE_{pe.}$ or the one given by the system optimized using $PE_{std.}$, or if they regarded them as equivalent. The evaluators rated each of the 175 utterances of the PE corpus, and rated 928 utterances (among 2852) of the TEST corpus.

We are interested in utterances for which at least two evaluators give the same decision by taking a majority vote. This represents approximately 95% of the evaluated utterances. The results are given in Table 7. We can observe the same trend as the automatic evaluation: for the PE corpus, evaluators significantly prefer (64%) optimization on post-edited translations ($PE_{pe.}$). On the TEST corpus, evaluators judge, by a majority the two optimizations as equivalent.

Corpus	*# Utt.*	*$PE_{std.}$*	*$PE_{pe.}$*	*No preference*
TEST	859	33%	24%	43%
PE	158	16%	64%	19%

Table 6. Subjective evaluation of systems according to the reference used for MERT

5 Conclusion and Future Work

We presented our preliminary study concerning interactive and iterative methods for improving an automatic statistical MT system using user's feedback. We have created a baseline phrase-based SMT system and collected, in a first step, a corpus of 175 post-edited translations corresponding to human corrections of our system outputs. Those post-editions were produced by native speakers of the source language with good knowledge of the target language.

The purpose of the experiments that followed has been to try to improve our translation system using these post-edited data. These data were used at three different levels of the translation process: a) in the training corpus, b) to automatically post-edit the system outputs and c) to adjust the weights of the log-linear model.

These experiments have shown that those three techniques are efficient for correcting and improving the system when it translates again the same data (the system learns from its errors); but it is difficult to propagate these corrections/improvements on new data (no great generalization). Indeed, the corpus of post-edited data (consisting of 175 utterances) is too small to allow us to draw a definitive conclusion.

Therefore, we are currently focusing our efforts on collecting a corpus of manual post-edition of 12 000 utterances. On one hand, we will analyze in detail the corrections made by the post-editors on the outputs of our MT system (quantitative and qualitative study of the post-edition data collected). On the other hand, we will re-evaluate and extend the methods proposed here with this larger amount of post-edited data. We plan, for example, to directly modify the translation table using human post-editions: this can be done by filtering/replacing lines corresponding to corrected outputs, while giving more confidence to segments considered as correct by human annotators.

References

Blanchon Hervé, Boitet Christian and Huynh Cong-Phap. 2009. A Web Service Enabling Gradable Post-edition of Pre-translations Produced by Existing Translation Tools: Practical Use to Provide High-quality Translation of an Online Encyclopedia. *Beyond Translation Memories: New Tools for Translators Workshop at Machine Translation Summit XII*, Ottawa, Canada. 20–27

Brown Peter F., Cocke John, Della Pietra Stephen A., Della Pietra Vincent J., Jelinek Fredrick, Lafferty

John D., Mercer Robert L., Roossin Paul S. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Callison-Burch Chris. 2003. Active learning for statistical machine translation. *PhD proposal, Edinburgh University, UK*.

Déchelotte Daniel. 2010. Analysis of translation suggestions on Reverso translation engines : initial findings. *FAUST project report, LIMSI-CNRS. 13 p.*

Diaz de Ilarraza Arantza, Labaka Gorka and Sarasola Kepa. 2008. Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages. *Mixing Approaches to Machine Translation Workshop*, Donostia-San Sebastian, Spain. 35–40

Dugast Loic, Senellart Jean and Koehn Philipp. 2009. Statistical Post Editing and Dictionary Extraction: Systran/Edinburg submissions for ACL-WMT2009. *4th Workshop on Statistical Machine Translation*, Athens, Greece. 110–114.

Kneser Reinhard and Herman Ney. 1995. Improved Backing-Off for M-Gram Language Modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, USA. 181–184.

Koehn Philipp. 2004. Statistical Significance Tests for Machine Translation Evaluation. *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 388–395.

Koehn Philipp, Hoang Hieu, Birch Alexandra, Callison-Burch Chris, Federico Marcello, Bertoldi Nicola, Cowan Brooke, Shen Wade, Moran Christine, Zens Richard, Dyer Chris, Bojar Ondřej, Constantin Alexandra and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *45st Annual Meeting of the Association for Computational Linguistics on Human Language Technology*, demonstration session, Prague, Czech Republic. 177–180.

Koehn Philipp, Och Franz Josef and Marcu Daniel. 2003. Statistical Phrase-Based Translation. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada. 48–54.

Lagarda Antonio L., Alabau Vincente, Casacuberta Francisco, Silva Roberto and Díaz-de-Liaño Enrique. 2009. Statistical Post-Editing of a Rule-Based Machine Translation System. *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, Boulder, CO, USA. 217–220.

Llitjos Ariadna Font and Carbonell Jaime G. 2006. Automating Post-Editing To Improve MT Systems. *Automated Post-Editing and Apllications Workshop*, *7th biennial Association for Machine Translation in the Americas Conference*, Cambridge, MA, USA. 10 p.

Marcu Daniel and Wong William. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. *Empirical Methods for Natural Language Processing*, Philadelphia, PA, USA. 133–139.

Och Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. *41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan. 160–167.

Och Franz Josef and Ney Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29:19–51.

Papineni Kishore, Roukos Salim, Ward Todd and Zhu Wei-Jing. 2002. BLEU : A Method for Automatic Evaluation of Machine Translation. *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA. 311–318.

Potet Marion, Besacier Laurent and Blanchon Hervé. 2010. The LIG Machine Translation System for WMT 2010. *Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden. 167–172.

Simard Michel, Ueffing Nicola, Isabelle Pierre and Roland Kuhn. 2007a. Rule-Based Translation with Statistical Phrase-Based Post-Editing. *2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic. 203–206.

Simard Michel, Goutte Cyril and Isabelle Pierre. 2007b. Statistical Phrase-based Post-editing. *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, USA. 508–515

Stolcke Andreas. 2002. SRILM - an Extensible Language Modeling Toolkit. *International Conference on Spoken Language Processing*, Denver, CO, USA. 901–904.

Tomanek Katrin and Olsson Fredrik. 2009. A Web Survey on the Use of Active Learning to Support Annotation of Text Data. *Workshop on Active Learning for Natural Language Processing*, *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, Boulder, CO, USA. 45–48.