Statistical Post-Editing of a Rule-Based Machine Translation System*

A.-L. Lagarda, V. Alabau, F. Casacuberta
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, Spain
alagarda@iti.upv.es

R. Silva, and **E. Díaz-de-Liaño**
Celer Soluciones, S.L.
Madrid, Spain

Abstract

Automatic post-editing (*APE*) systems aim at correcting the output of machine translation systems to produce better quality translations, i.e. produce translations can be manually post-edited with an increase in productivity. In this work, we present an *APE* system that uses statistical models to enhance a commercial rule-based machine translation (*RBMT*) system. In addition, a procedure for effortless human evaluation has been established. We have tested the *APE* system with two corpora of different complexity. For the *Parliament* corpus, we show that the *APE* system significantly complements and improves the *RBMT* system. Results for the *Protocols* corpus, although less conclusive, are promising as well. Finally, several possible sources of errors have been identified which will help develop future system enhancements.

1 Introduction

Current machine translation systems are far from perfect. To achieve high-quality output, the raw translations they generate often need to be corrected, or post-edited by human translators. One way of increasing the productivity of the whole process is the development of automatic post-editing (*APE*) systems (Dugast et al., 2007; Simard et al., 2007).

* Work supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, by the Spanish research programme Consolider Ingenio 2010:MIPRCV (CSD2007-00018), and by the i3media Cenit project (CDTI 2007-1012).

Many of these works propose a combination of rule-based machine translation (*RBMT*) and statistical machine translation (*SMT*) systems, in order to take advantage of the particular capabilities of each system (Chen and Chen, 1997).

A possible combination is to automatically post-edit the output of a *RBMT* system employing a *SMT* system. In this work, we will apply this technique into two different corpora: *Parliament* and *Protocols*. In addition, we will propose a new human evaluation measure that will deal with the impact of the automatic post-editing.

This paper is structured as follows: after a brief introduction of the *RBMT*, *SMT*, and *APE* systems in Section 2, Section 3 details the carried out experimentation, discussing its results. Finally, some conclusions and future work are presented in Section 4.

2 Systems description

Three different systems are compared in this work, namely the *RBMT*, *SMT*, and *APE* systems.

Rule-based machine translation. *RBMT* was the first approach to machine translation, and thus, a relatively mature area in this field. *RBMT* systems are basically constituted by two components: the rules, that account for the syntactic knowledge, and the lexicon, which deals with the morphological, syntactic, and semantic information. Both rules and lexicons are grounded on linguistic knowledge and generated by expert linguists. As a result, the build process is expensive and the system is difficult to maintain (Bennett and Slocum, 1985). Furthermore, *RBMT* systems fail to adapt to new domains.

Proceedings of NAACL HLT 2009: Short Papers, pages 217–220,
Boulder, Colorado, June 2009.

Although they usually provide a mechanism to create new rules and extend and adapt the lexicon, changes are usually very costly and the results, frequently, do not pay off (Isabelle et al., 2007).

Statistical machine translation. In *SMT*, translations are generated on the basis of statistical models, which are derived from the analysis of bilingual text corpora. The translation problem can be statistically formulated as in (Brown et al., 1993). In practice, several models are often combined into a *log-linear* fashion. Each model can represent an important feature for the translation, such as *phrase-based*, *language*, or *lexical* models (Koehn et al., 2003).

Automatic post-editing. An *APE* system can be viewed as a translation process between the output from a previous MT system, and the target language. In our case, an *APE* system based on statistical models will be trained to correct the translation errors made by a *RBMT* system. As a result, both *RBMT* and *SMT* technologies will be combined in order to increase the overall translation quality.

3 Experiments

We present some experiments carried out using the introduced *APE* system, and comparing its performance with that of the *RBMT* and *SMT* systems. In the experimentation, two different English-to-Spanish corpora have been chosen, *Parliament* and *Protocols*, both of them provided by a professional translation agency.

Corpora. The *Parliament* corpus consists of a series of documents from proceedings of parliamentary sessions, provided by a client of the translation agency involved in this work. Most of the sentences are transcriptions of parliamentary speeches, and thus, with the peculiarities of the oral language. Despite of the multi-topic nature of the speeches, differences in training and test perplexities indicate that the topics in test are well represented in the training set (corpus statistics in Table 1).

On the other hand, the *Protocols* corpus is a collection of medical protocols. This is a more difficult task, as its statistics reflect in Table 1. There are many factors that explain this complexity, such as the different companies involved in training and test sets, out-of-domain test data (see perplexity and out-of-vocabulary words), non-native authors, etc.

Table 1: Corpus statistics for *Parliament* and *Protocols*. OOV stands for out-of-vocabulary words.

		Parliament		*Protocols*	
		En	Sp	En	Sp
Training	Sentences	90K	90K	154K	154K
	Run. words	2.3M	2.5M	3.2M	3.6M
	Vocabulary	29K	45K	41K	47K
	Perplexity	42	37	21	19
Test	Sentences	1K	1K	3K	3K
	Run. words	33K	33K	54K	71K
	OOVs	157	219	2K	1.7K
	Perplexity	44	43	131	173

Evaluation. In order to assess the proposed systems, a series of measures have been considered. In first place, some state-of-the-art automatic metrics have been chosen to give a first idea of the quality of the translations. These translations have been also evaluated by professional translators to assess the increase of productivity when using each system.

Automatic evaluation. The automatic assessment of the translation quality has been carried out using the *BiLingual Evaluation Understudy* (BLEU) (Papineni et al., 2002), and the *Translation Error Rate* (TER) (Snover et al., 2006). The latter takes into account the number of edits required to convert the system output into the reference. Hence, this measure roughly estimates the post-edition process.

Human evaluation. A new human evaluation measure has been proposed to roughly estimate the productivity increase when using each of the systems in a real scenario, grounded on previous works for human evaluation of qualitative factors (Callison-Burch et al., 2007). One of the desired qualities for this measure was that it should pose little effort to the human evaluator. Thus, a binary measure was chosen, the *suitability*, where the translations are identified as suitable or not suitable. A given translation is considered to be suitable if it can be manually post-edited with effort savings, i.e., the evaluator thinks that a manual post-editing will increase his productivity. On the contrary, if the evaluator prefers to ignore the proposed translation and start it over, the sentence is deemed not suitable.

Significance tests. Significance of the results has been assessed by the *paired bootstrap resampling* method, described in (Koehn, 2004). It estimates how confidently the conclusion that a system outperforms another one can be drawn from a test result.

Experimental setup. Rule-based translation was performed by means of a commercial *RBMT* system. On the other hand, statistical training and translation in both *SMT* and *APE* systems were carried out using the Moses toolkit (Koehn et al., 2007). It should be noted that *APE* system was trained taking the *RBMT* output as source, instead of the original text. In this way, it is able to post-edit the *RBMT* translations.

Finally, the texts employed for the human evaluation were composed by 350 sentences randomly drawn from each one of the two test corpora described in this paper. Two professional translators carried out the human evaluation.

3.1 Results and discussion

Experimentation results in terms of automatic and human evaluation are shown in this section.

Automatic evaluation. Table 2 presents *Parliament* and *Protocols* corpora translation results in terms of automatic metrics. Note that, as there is a single reference, this results are somehow pessimistic.

In the case of the *Parliament* corpus, *SMT* system outperforms the rest of the systems. *APE* results are slightly worse than *SMT*, but far better than *RBMT*.

However, when moving to the *Protocols* corpus, a more difficult task (as seen in perplexity in Table 1), the results show quite the contrary. *SMT* and *APE* systems show how they are more sensitive to out-of-domain documents. Nevertheless, the *RBMT* system seems to be more robust under such conditions. Despite of the degradation of the statistical models, *APE* manages to achieve much better results than the other two systems. It is able to conserve the robustness of *RBMT*, while its statistical counterpart deals with the particularities of the corpus.

Human evaluation. Table 3 shows the percentage of translations deemed suitable by the human evaluators. Two professional evaluators analysed the suitability of the output of each system

In the *Parliament* case, *APE* performance is found much more suitable than the rest of the systems. In fact, this difference between *APE* and the rest is statistically significant at a 99% level of confidence. In addition, significance tests show that, on average, *APE* improves *RBMT* on 59.5% of translations.

Table 2: Automatic evaluation for *Parliament* and *Protocols* tests.

	Parliament		*Protocols*	
	BLEU	TER	BLEU	TER
RBMT	29.1	46.7	29.5	48.0
SMT	49.9	34.9	22.4	59.6
APE	48.4	35.9	33.6	46.2

Regarding to the *Protocols* corpus, it must be noted that a first review of the translations pointed out that the *SMT* system performed quite poorly. Hence, *SMT* was not considered for the human evaluation on this corpus.

Figures show that *APE* complements and improves *RBMT*, although differences between them are tighter than in the *Parliament* corpus. However, significance tests still prove that these improvements are statistically significant (68% of confidence), and that the average improvement is 6.5%.

Table 3: Human evaluation for *Parliament* and *Protocols* corpora. Percentage of suitable translated sentences for each system.

	Parliament	*Protocols*
RBMT	58	60
SMT	60	–
APE	94	67

It is interesting to note how automatic measures and human evaluation seem not to be quite correlated. In terms of automatic measures, the best system to translate the *Parliament* test is the *SMT*. This improvement has been checked by carrying out significance tests, resulting statistically significant with a 99% of confidence. However, in the human evaluation, *SMT* is worse than *APE* (this difference is also significant at 99%). On the other hand, when working with the *Protocols* corpus, automatic metrics indicate that *APE* improves the rest (significant improvement at 99%). Nevertheless, human evaluators seem to think that the difference between *APE* and *RBMT* is not so significant, only with a confidence of 68%. Previous works confirm this apparent

discrepancy between automatic and human evaluations (Callison-Burch et al., 2007).

Translator's commentaries. As a subproduct of the human evaluation, the evaluators gave some personal impressions regarding each system performance. They concluded that, when working with the *Parliament* corpus, there was a net improvement in the overall performance when using *APE*. Changes between *RBMT* and *APE* were minor but useful. Thus, *APE* did not pose a system degradation with respect to the *RBMT*. Furthermore, a rough estimation indicated that over 10% of the sentences were perfectly translated, i.e. the translation was human-like. In addition, some frequent collocations were found to be correctly post-edited by the *APE* system, which was felt very effort saving.

With respect to the *Protocols* corpus, as expected, results were found not so satisfactory. However, human translators find themselves these documents complex.

Finally, in both cases, *APE* is able to make the translation more similar to the reference by fixing some words without altering the grammatical structure of the sentence. Finally, translators would find very useful a system that automatically decided when to automatically post-edit the *RBMT* outputs.

4 Conclusions

We have presented an automatic post-editing system that can be added at the core of the professional translation workflow. Furthermore, we have tested it with two corpora of different complexity.

For the *Parliament* corpus, we have shown that the *APE* system complements and improves the *RBMT* system in terms of suitability in a real translation scenario (average improvement 59.5%). Results for the *Protocols* corpus, although less conclusive, are promising as well (average improvement 6.5%). Moreover, 67% of *Protocols* translations, and 94% of *Parliament* translations were considered to be suitable.

Finally, a procedure for effortless human evaluation has been established. A future improvement for this would be to integrate the process in the core of the translator's workflow, so that on-the-fly evaluation can be made. In addition, several possible sources of errors have been identified which will help develop future system enhancements. For example, as stated in the translator's commentaries, the automatic selection of the most suitable translation among the systems is a desirable feature.

References

W. S. Bennett and J. Slocum. 1985. The lrc machine translation system. *Comp. Linguist.*, 11(2-3):111–121.

P. F. Brown, S. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comp. Linguist.*, 19(2):263–312.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (meta-) evaluation of machine translation. In *Proc. of the 2nd Workshop on SMT*, pages 136–158, Prague, Czech Republic. ACL.

K. Chen and H. Chen. 1997. A hybrid approach to machine translation system design. In *Comp. Linguist. and Chinese Language Processing 23*, pages 241–265.

L. Dugast, J. Senellart, and P. Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proc. of the 2nd Workshop on SMT*, pages 220–223, Prague, Czech Republic. ACL.

P. Isabelle, C. Goutte, and M. Simard. 2007. Domain adaptation of mt systems through automatic post-editing. In *Proc. of MTSummit XI*, pages 255–261, Copenhagen, Denmark.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL-HLT*, pages 48–54, Edmonton, Canada.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, pages 177–180, Prague, Czech Republic.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP 2004*, Barcelona, Spain.

K. Papineni, S. Roukos, T. Ward, and W.-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Philadelphia, PA, USA.

M. Simard, C. Goutte, and P. Isabelle. 2007. Statistical phrase-based post-editing. In *Proc. of NAACL-HLT2007*, pages 508–515, Rochester, NY. ACL.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, pages 223–231.