Rule-based Translation With Statistical Phrase-based Post-editing

Michel Simard, Nicola Ueffing, Pierre Isabelle and **Roland Kuhn**
Interactive Language Technologies Group
National Research Council of Canada
Gatineau, Canada, K1A 0R6
firstname.lastname@nrc-cnrc.gc.ca

Abstract

This article describes a machine translation system based on an *automatic post-editing* strategy: initially translate the input text into the target-language using a rule-based MT system, then automatically post-edit the output using a statistical phrase-based system. An implementation of this approach based on the SYSTRAN and PORTAGE MT systems was used in the shared task of the Second Workshop on Statistical Machine Translation. Experimental results on the test data of the previous campaign are presented.

1 Introduction

Simard et al. (2007) have recently shown how a statistical phrase-based machine translation system can be used as an *automatic post-editing* (APE) layer, on top of a rule-based machine translation system. The motivation for their work is the repetitive nature of the errors typically made by rule-based systems. Given appropriate training material, a statistical MT system can be trained to correct these systematic errors, therefore reducing the post-editing effort. The statistical system views the output of the rule-based system as the source language, and reference human translations as the target language. Because the training material for the APE layer will typically be domain-specific, this process can be viewed as a way of *automatically adapting* a rule-based system to a specific application domain.

This approach has been shown experimentally to produce large improvements in performance not only over the baseline rule-based system that it corrects, but also over a similar statistical phrase-based MT system used in standalone mode, i.e. translating the “real” source text directly: Simard et al. report a reduction in post-editing effort of up to a third when compared to the input rule-based translation, and as much as 5 BLEU points improvement over the direct SMT approach.

These impressive results, however, were obtained in a very specific and somewhat unusual context: the training and test corpora were extracted from a collection of manually post-edited machine translations. The two corpora (one English-to-French, one French-to-English) each contained three parallel “views” of the same data: 1) the source language text, 2) a machine translation of that text into the target language, as produced by a commercial rule-based MT system, and 3) the final target-language version of the text, produced by manually post-editing the machine translation. Furthermore, the corpus was very small, at least by SMT standards: 500K words of source-language data in the French-to-English direction, 350K words in the English-to-French. Because of this, the authors were left with two important questions: 1) how would the results scale up to much larger quantities of training data? and 2) are the results related to the dependent nature of the translations, i.e. is the automatic post-editing approach still effective when the machine and human translations are produced independently of one another?

With these two questions in mind, we participated in the shared task of the Second Workshop on Statistical Machine Translation with an automatic post-editing strategy: initially translate the input text into the target-language using a rule-based system, namely SYSTRAN, and automatically post-edit the output using a statistical phrase-based system, namely PORTAGE. We describe our system in more detail in Section 2, and present some experimental results in Section 3.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 203–206,
Prague, June 2007. ©2007 Association for Computational Linguistics

2 System description

Our system is composed of two main components: a rule-based MT system, which handles the initial translation into the target language, and a statistical phrase-based post-editing system, which performs domain-specific corrections and adaptations to the output. We describe each component separately below.

2.1 Rule-based Translation

The initial source-to-target language translation is performed using the SYSTRAN machine translation system, version 6. A detailed overview of SYSTRAN systems can be found in Dugast et al. (2007). For this shared task, we used the French-to-English and English-to-French configurations of the system. Although it is possible to provide the system with specialized lexica, we did not rely on this feature, and used the system in its basic "out-of-the-box" configuration.

2.2 Statistical Phrase-based Post-Editing

The output of the rule-based MT system described above is fed into a post-editing layer that performs domain-specific corrections and adaptation. This operation is conceptually not very different from a "target-to-target" translation; for this task, we used the PORTAGE system, a state-of-the-art statistical phrase-based machine translation system developed at the National Research Council of Canada (NRC).[1] A general description of PORTAGE can be found in (Sadat et al., 2005).

For our participation in this shared task, we decided to configure and train the PORTAGE system for post-editing in a manner as much as possible similar to the corresponding translation system, the details of which can be found in (Ueffing et al., 2007). The main features of this configuration are:

- The use of two distinct phrase tables, containing phrase pairs extracted from the Europarl and the News Commentary training corpora respectively.
- Multiple phrase-probability feature functions in the log-linear models, including a joint probability estimate, a standard frequency-based conditional probability estimate, and variants thereof based on different smoothing methods (Foster et al., 2006).
- A 4-gram language model trained on the combined Europarl and News Commentary target-language corpora.
- A 3-gram *adapted language model*: this is trained on a mini-corpus of test-relevant target-language sentences, extracted from the training material using standard information retrieval techniques.
- A 5-gram truecasing model, trained on the combined Europarl and News Commentary target-language corpora.

2.3 Training data

Ideally, the training material for the post-editing layer of our system should consist in a corpus of text in two parallel versions: on the one hand, raw machine translation output, and on the other hand, manually post-edited versions of these translations. This is the type of data that was used in the initial study of Simard et al. (2007).

Unfortunately, this sort of training data is seldom available. Instead, we propose using training material derived directly from standard, source-target parallel corpora. The idea is to translate the source portion of the parallel corpus into the target language, using the rule-based MT component. The post-editing component can then be trained using this translation as "source" training material, and the existing target portion of the parallel corpus as "target" training material. Note how this sort of data is subtly different from the data used by Simard et al.: there, the "target" text was dependent on the "source", in the sense that it was produced by manually post-editing the machine translation; here, the two can be said to be independent, in the sense that both "source" and "target" were produced independently by man and machine (but from the same "real" source, of course). It was one of the initial motivations of the current work to verify to what extent the performance of the APE approach is affected by using two different translations (human and ma-

[1] A version of PORTAGE is made available by the NRC to Canadian universities for research and education purposes.

	en → fr	fr → en
Europarl (>32M words/language)		
SYSTRAN	23.06	20.11
PORTAGE	31.01	30.90
SYSTRAN+PORTAGE	31.11	30.61
News Commentary (1M words/language)		
SYSTRAN	24.41	18.09
PORTAGE	25.98	25.17
SYSTRAN+PORTAGE	28.80	26.79

Table 1: System performances on WMT-06 test. All figures are single-reference BLEU scores, computed on truecased, detokenized translations.

chine) instead of two versions of the same translation (raw MT versus post-edited MT).

We concentrated our efforts on the English-French language pair. For each translation direction, we prepared two systems: one for the Europarl domain, and one for the News Commentary domain. The two systems have almost identical configurations (phrase tables, log-linear model features, etc.); the only differences between the two are the *adapted language model*, which is computed based on the specific text to be translated and the parameters of the log-linear models, which are optimized using domain-specific development sets. For the Europarl domain system, we used the *dev2006* and *devtest2006* data sets, while for the News Commentary, we used the *nc-dev2007*. Typically, the optimization procedure will give higher weights to Europarl-trained phrase tables for the Europarl domain systems, and inversely for the News Commentary domain systems.

3 Experimental Results

We computed BLEU scores for all four systems on the 2006 test data (*test2006* for the Europarl domain and *nc-devtest2007* for the News Commentary). The results are presented in Table 1. As points of comparison, we also give the scores obtained by the SYSTRAN systems on their own (i.e. without a post-editing layer), and by the PORTAGE MT systems on their own (i.e. translating directly source into target).

The first observation is that, as was the case in the Simard et al. study, post-editing (*SYSTRAN+PORTAGE* lines) very significantly increases the BLEU scores of the rule-based system (*SYSTRAN* lines). This increase is more spectacular in the Europarl domain and when translating into English, but it is visible for all four systems.

For the News Commentary domain, the APE strategy (*SYSTRAN+PORTAGE* lines) clearly outperforms the direct SMT strategy (*PORTAGE* lines): translating into English, the gain exceeds 1.5 BLEU points, while for French, it is close to 3 BLEU points. In contrast, for the Europarl domain, both approaches display similar performances. Let us recall that the News Commentary corpus contains less than 50K sentence pairs, totalling a little over one million words in each language. With close to 1.3 million sentence pairs, the Europarl corpus is almost 30 times larger. Our results therefore appear to confirm one of the conjectures of the Simard et al. study: that APE is better suited for domains with limited quantities of available training data. To better understand this behavior, we trained series of APE and SMT systems on the Europarl data, using increasing amounts of training data. The resulting learning curves are presented in Figure 1.[2]

As observed in the Simard et al. study, while both the SMT and APE systems improve quite steadily with more data (note the logarithmic scale), SMT appears to improve more rapidly than APE. However, there doesn't seem to be a clear "crossover" point, as initially conjectured by Simard et al. Instead, SMT eventually catches up with APE (anywhere between 100K and 1M sentence pairs), beyond which point both approaches appear to be more or less equivalent. Again, one impressive feature of the APE strategy is how little data is actually required to improve upon the rule-based system upon which it is built: around 5000 sentence pairs for English-to-French, and 2000 for French-to-English.

4 Conclusions

We have presented a combination MT system based on a post-editing strategy, in which a statistical phrase-based system corrects the output of a rule-based translation system. Experiments confirm the

[2] The systems used for this experiment are simplified versions of those described in Section 2, using only one phrase table, a trigram language model and no rescoring; furthermore, they were optimized and tested on short sentences only.

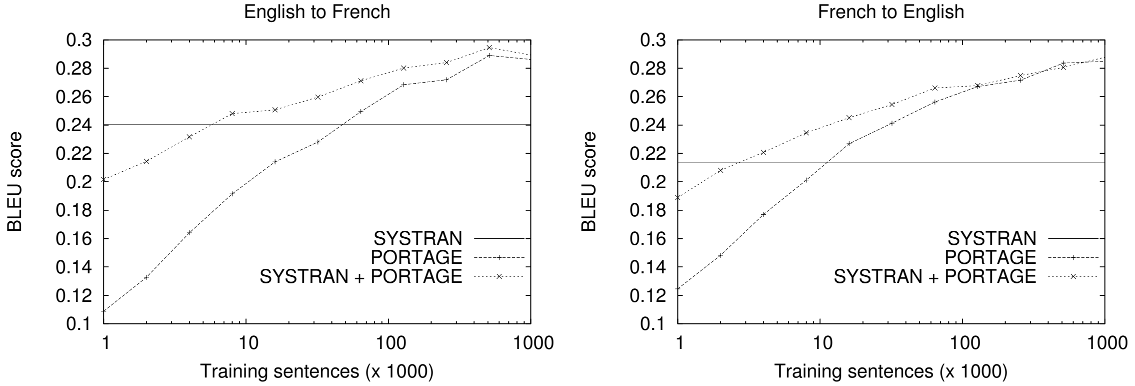

Figure 1: BLEU scores on Europarl data under increasing amounts of training data for PORTAGE SMT alone and SYSTRAN MT with PORTAGE APE.

conclusions of earlier studies: not only can phrase-based post-editing significantly improve the output of a rule-based MT system (in terms of BLEU score), but when training data is scarce, it also outperforms a direct phrase-based MT strategy. Furthermore, our results indicate that the training data for the post-editing component does not need to be manually post-edited translations, it can be generated from standard parallel corpora. Finally, our experiments show that while post-editing is most effective when little training data is available, it remains competitive with phrase-based translation even with much larger amounts of data.

This work opens the door to a number of lines of investigation. For example, it was mentioned earlier that phrase-based APE could be seen as a form of automatic domain-adaptation for rule-based methods. One thing we would like to verify is how this approach compares to the standard “lexical customization” method proposed by most rule-based MT vendors. Also, in the experiments reported here, we have used identical configurations for the APE and direct SMT systems. However, it might be possible to modify the phrase-based system so as to better adapt it to the APE task. For example, it could be useful for the APE layer to “look” at the real source-language text, in addition to the MT output it is post-editing. Finally, we have so far considered the front-end rule-based system as a “black box”. But in the end, the real question is: Which part of the rule-based processing is really making things easier for the phrase-based post-editing layer? Answering this question will likely require diving into the internals of the rule-based component. These are all directions that we are currently pursuing.

Acknowledgements

This work was done as part of a collaboration with SYSTRAN S.A. Many thanks go to Jean Senellart, Jens Stephan, Dimitris Sabatakakis and all those people behind the scene at SYSTRAN.

References

L. Dugast, J. Senellart, and P. Koehn. 2007. Statistical Post-Edition on SYSTRAN Rule-Based Translation System. In *Proceedings of the Second Workshop On Statistical Machine Translation*, Prague, Czech Republic.

G. Foster, R. Kuhn, and H. Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of EMNLP 2006*, pages 53–61, Sydney, Australia.

F. Sadat, H. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin, and A. Tikuisis. 2005. PORTAGE: A Phrase-Based Machine Translation System. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 129–132, Ann Arbor, USA.

M. Simard, C. Goutte, and P. Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, USA.

N. Ueffing, M. Simard, S. Larkin, and H. Johnson. 2007. NRC’s PORTAGE system for WMT 2007. In *Proceedings of the Second Workshop On Statistical Machine Translation*, Prague, Czech Republic.