Comparative Evaluation of Statistical Post-Editing in English-Japanese MT

Alexis Kauffmann Asheesh Gulati
University of Geneva
{alexis.kauffmann, asheesh.gulati}@unige.ch

Abstract

This paper describes a comparative evaluation experiment of machine translation (MT) output for the English-to-Japanese language pair, comparing the results of statistical post-editing (SPE), linguistics-based MT (LBMT) and hierarchichal MT (HMT). Results show that SPE can be an efficient method of improving English-Japanese LBMT output, giving results equivalent to the ones given by an HMT system, in both automatic and manual evaluations.

1 Introduction

Nowadays, machine translation (MT) has mostly shifted towards statistical approaches, but rule-based and example-based approaches have yet to yield. A recent trend in the field is to combine techniques from different approaches into hybrid machine translation systems, in a bid to outperform systems based purely on these underlying approaches. One such combination is the use of statistical post-editing (SPE) to automatically correct the output of a rule-based system [6].

SPE has shown to produce very good results and the literature contains many reports of successful application, and also some more constrasted results [9]. A number of SPE experiments have been described for the Japanese-English language pair ([8], [9]). On the reverse translation direction, syntactic reordering preprocesing has been studied [4], but we could not find accounts of English-Japanese SPE experiments. In this paper, we describe a comparative evaluation experiment of machine translation output for the English-Japanese language pair, with focus on the correlation between automatic and manual evaluations.

2 Experiment

In a previous experiment on English-Japanese SPE, we found that the BLEU score of the translation was clearly higher, rising from 0.26 to 6.2 on our test set. However, human evaluation showed that there was no correlation between the increase in BLEU score and a real refinement of the translation quality, which suffered a global degradation. The experiment was conducted using a parallel corpus of 5,000 sentences as training and tuning data, an amount that happened to be insufficient and led to a high rate of ungrammatical or semantically inappropriate translations, and included a 3-gram language model.

Our present experiment is conducted using a much larger corpus as training and tuning data and a higher order language model. The experimental setup consists of 4 systems:

- Its-2, a transfer-based LBMT system developed by the Language Technology Laboratory (LATL) at the University of Geneva [10], in the English-Japanese version [2].
- Joshua, a statistical HMT system [5], including a 5-gram language model built using SRILM [7].
- Its-2 with a SPE component, trained using Moses [3] and including a 5-gram language model built using IRSTLM [1].
- Its-2 with a SPE component, trained using Joshua and including a 5-gram language model built using SRILM.

All three statistical systems (Joshua and both SPE components) have been trained on the same data sets: a parallel corpus of 48,000 sentences and 1,000 sentences for tuning (with the source side first translated using Its-2 in the case of SPE), and a monolingual Japanese corpus of 978,493 sentences for the language model. We have used the Tatoeba database[1], composed of example sentences, for the parallel corpus, and a collection of various texts for the monolingual corpus.

In addition, we have added 2 experimental conditions for the pipeline involving Its-2 and the SPE component trained using Moses:

- Reordering model, comparing unlimited reordering with the default reordering.

[1] http://tatoeba.org/eng/

- Reordering constraints, using XML markup for *walls* and *zones* in the source text[2].

2.1 Reordering model

The Moses decoder provides a way to specify a limit on distortion, affecting the cost of reordering. The default model is suitable for local reorderings and should be appropriate in the case of SPE, as both source and target sentences are in Japanese and would roughly follow the same word order. If we find an increase in translation quality by allowing unlimited reordering, then this indicates a specific shortcoming of the underlying transfer-based system.

2.2 Reordering constraints

It is sometimes useful to define parts of a given sentence that need to be translated independently of the rest of the sentence. Such parts include direct dialogues and other quoted material, and can be specified using XML markup. We add <zone> tags around Japanese quotation marks 「」, to force the decoder to translate these as a block, and we insert <wall> tags before final punctuation marks to keep them at the end of sentences.

3 Evaluation

3.1 Automatic Evaluation

The test set is composed of 1,000 sentences from the Tatoeba database. Results for all systems and conditions are given in Table 1.

We can see that the HMT system Joshua obtains the highest score, and that the different SPE systems are roughly equivalent to Joshua, with only about 0.02 points of difference between the Joshua score and the best SPE score. If we consider the Joshua HMT as a strong baseline, this result shows the validity of SPE applied on Its-2 English-Japanese output. Still, this claim needs to be confirmed by a manual evaluation of the output quality.

Among all SPE systems, Moses using both unlimited reordering and reordering constraints achieves the highest score, although the variation in score between the different SPE configurations is extremely small (between 0.0004 and 0.0061 points). The small impact of unlimited reordering vs. default reordering reassures us that Its-2 output is already of a good quality with respect to word order.

3.2 Manual Evaluation

We conduct a manual evaluation on a set of 45 sentences taken randomly from the automatic evaluation test set. Scores between 0 and 1 are given to every translation, and the score average for each MT system and condition is shown in Table 2.

System	**BLEU**
Its-2	0.0665
Its-2 + SPE (with Moses)	
- DR	0.2517
- DR + RC	0.2513
- UR	0.2569
- UR + RC	0.2578
Its-2 + SPE (with Joshua)	0.2523
Joshua	0.2788

Table 1: Comparison of BLEU scores: DR stands for ``default reordering'', UR for ``unlimited reordering'', RC for ``reordering constraints''.

We can see that the best score average is obtained by the hierarchical SPE (Joshua SPE), and that the HMT system Joshua almost obtains the same score average (only 0.013 point difference). This confirms that SPE, and especially hierarchical SPE can be an efficient way of improving Its-2 output, reaching results equivalent to those of Joshua.

Among all SPE systems, Joshua SPE seems clearly more effcient (+0.7 or +0.8 points ahead) than the different configurations of phrase-based SPE (Moses SPE), a fact that did not appear in the automatic evaluation results. As shown in Table 3, the percentage of cleary improved translations is 60% with Joshua SPE, whereas it only reaches 44% with Moses SPE. This gain in syntactic and semantic accuracy may be a consequence of the hierarchical phrase-based model of Joshua, that seems more powerful than the classical statistical phrase-based model of Moses.

Having a closer look at the translation outputs, comparison between produced sentences can be sorted in different types of cases depending on the source sentence.

In some cases, SPE, especially Joshua SPE, refines the Its-2 translation and produces a better translation than the Joshua translation, as in Examples 1 and 2.

(1) Source sentence: The teenage friends stayed up talking all night.

Its-2 translation: １０代の 友達は すべての 夜話して とどまった。

Joshua SPE translation: １０ 代 の 友人 は 一晩 中 おしゃべり を して いた 。

Joshua translation: 友達teenage?oov を言って一晩中起きて いた。

[2] www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc17

System	score average
Its-2	0.488
Its-2 + SPE (Moses)	
- DR (+RC)	0.522
- UR	0.513
- UR + RC	0.522
Its-2 + SPE (Joshua)	0.593
Joshua	0.580

Table 2: Comparison of manual scores computed on a random sample of 45 sentences: DR stands for ``default reordering'', UR for ``unlimited reordering'', RC for ``reordering constraints''.

	+	=	-
Moses SPE (DR)	44%	23%	33%
Joshua SPE	60%	18%	22%

Table 3: Translation improvement rates for Moses SPE with default reordering (DR) and Joshua SPE: + stands for ``improved translations'', = for ``translations of equivalent quality'', and - for ``deteriorated translations''.

(2) Source sentence: I don't want dinner.
Its-2 translation: 私は 夕食が ほしくない。
Joshua SPE translation: 私 は 夕食 は いり ませ ん 。
Joshua translation: 私 は 夕食 たく あり ませ ん 。

In other cases, as in Example 3, the Its-2 tanslation is too wrong or too broken for the SPE to improve it well, and the Joshua translation, even if not perfect, is clearly the best.

(3) Source sentence: At the beginning of the fifth year, Tony said he was going to have to charge more.
Its-2 translation: 第5の 年 の 初め で,
Joshua SPE translation: 第 五 年 の 初め に 、
Joshua translation: 5 年 の 初め に 、 彼 は トニー は 言った 担当 し なければ なら ない だろう 。

In other cases, the Its-2 translation remains more appropriate than the other ones, as in Example 4.

(4) Source sentence: She cherished his old love letters.
Its-2 translation: 彼女は 彼の 昔からの ラブレターを 持った。
Joshua SPE: 彼女 は 自分 の 古い ラブ レター を 開いた 。
Joshua translation: 彼女 は 古い 愛 を 秘めて 通 。

4 Conclusion and future work

SPE applied to the English-Japanese language pair drastically improves the overall BLEU score of the translation while preserving adequacy, as seen in the manual evaluation, but we also notice that the actual increase in translation quality depends upon the source sentence, and some sentences are not improved, some are even degraded. Hence, the next step would consist in training a system able to predict which sentence will benefit from SPE and selectively apply it (as in [8]), or return the original LBMT/HMT translation.

With respect to the amount of training data, ``more is better'', but we would like to investigate how much data is required to clearly outperform a statistical hierarchical phrase-based system, and if the domain may have an influence.

Finally, we plan to conduct a more in depth experiment on reordering constraints, using an appropriate test set composed of direct dialogues.

References

[1] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedingsof the 9th Annual Conference of the International Speech Communication Association*, pages 1618--1621, 2008.

[2] Alexis Kauffmann, Daisuke Kawahara, and Sadao Kurohashi. Treatment of Complex Sentences, Modality and Verbal Structures in Linguistics-Based MT. In *Proceedings of NLP 2011, Toyohashi, Japan*, 2011.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177--180, June 2007.

[4] Young-Suk Lee, Bing Zhao, and Xiaoqiang Luo. Constituent Reordering and Syntax Models for English-to-Japanese Statistical Machine Translation. In *Proceedings of the 23rd International*

Conference on Computational Linguistics (Coling 2010), pp. 626-634, Beijing, 2010.

[5] Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar F. Zaidan. Joshua: An Open Source Toolkit for Parsing-based Machine Translation. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, StatMT '09, pages 135--139, 2009.

[6] 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, pp. 203-206, Prague*, 2007.

[7] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901--904, 2002.

[8] Hirokazu Suzuki. 統計的後編集手法を適用したルールベース翻訳と文レベルの自動品質評価との融合 . In *Proceedings of NLP 2011, Toyohashi, Japan, pp.1119-1122*, 2011.

[9] Keisuke Toue, Tatsuya Izuha, and Jin'Ichi Murakami. 日英方向におけるハイブリッド翻訳とルールベース翻訳の人手評価 . In *Proceedings of NLP 2011, Toyohashi, Japan, pp.1127-1130*, 2011.

[10] Eric Wehrli, Luka Nerima, and Yves Scherrer. Deep Linguistic Multilingual Translation and Bilingual Dictionaries. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 90--94, 2009.