A machine translation system combining rule-based machine translation and statistical post-editing

Terumasa EHARA
Yamanashi Eiwa College
999, Yokone, Kofu, Yamanashi, JAPAN

eharate @ yamanashi-eiwa.ac.jp

Abstract

System architecture, experimental settings and evaluation results of the EIWA in the WAT2014 Japanese to English (ja-en) and Chinese to Japanese (zh-ja) tasks are described. Our system is combining rule-based machine translation (RBMT) and statistical post-editing (SPE). Evaluation results for ja-en task show 19.86 BLEU score, 0.7067 RIBES score, and 22.50 human evaluation score. Evaluation results for zh-ja task show 33.57 BLEU score, 0.8114 RIBES score, and 15.00 human evaluation score.

1 Introduction

One of the architectures of combining rule-based technique and statistical technique in the machine translation field is combining a rule-based machine translation (RBMT) and a statistical post-editing (SPE) (Dugast et al., 2007; Simard et al., 2007; Ehara, 2007).

The RBMT part translates source documents to target documents using rule-based machine translation. The SPE part automatically post-edits the output of the RBMT part to be more accurate target documents.

2 System architecture and experimental setting for the ja-en task

Our basic system architecture for ja-en task is shown in Figure 1 that is the same in the previous works (Ehara, 2007; Ehara, 2010; Ehara, 2011; Ehara, 2013). We use commercial based translation software for the RBMT part and the phrase based Moses (Koehn et al., 2003) for the SPE part. The distortion limit in the tuning process and decoding process for the SPE part is set to 1, because both the source and target languages on the SPE part are the same.

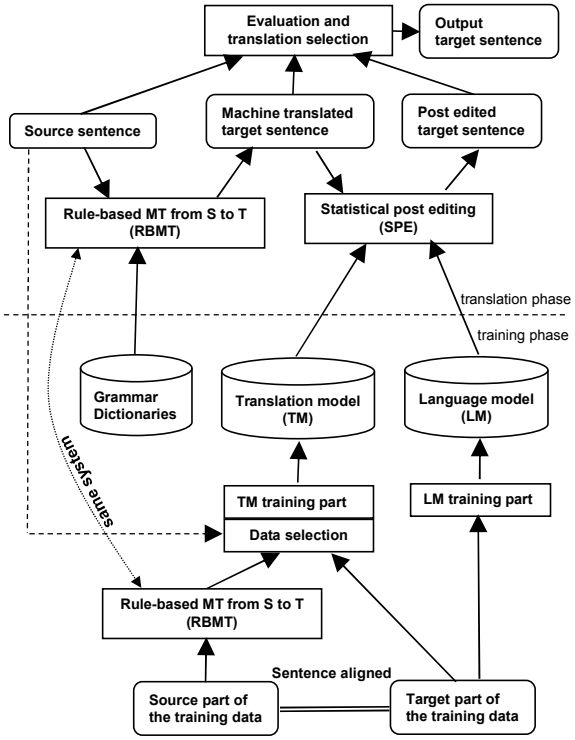

Figure 1: Basic system architecture for the ja-en task

2.1 Translation model adaptation

Translation model (TM) training for SPE part does not use whole training data for ja-en task (3,008,500 sentence pairs) but selected data adapted to the development data or the test data. For the SPE for the development data, 134,634 sentence pairs are used and for the SPE for the test data, 127,925 sentence pairs are used. The method of this filtering differ from the previous method (Ehara, 2010). The new method is as follows: At

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 50–54,
Tokyo, Japan, 4th October 2014.

a first step, we select training sentences which include a low frequency word in the development sentences (or the test sentences). The low frequency word means the count of such word in the training sentences are less than a threshold (we set 300). The word "ＳＡＲＳ", which occurs 13 times in the development set, occurs 189 times in the training set. So, it is a low frequency word in the development data. On the other hand, the word "化学", which occurs 29 times in the development set, occurs 40,964 times in the training set. So, it is not a low frequency word in the development data. At a second step, we add training sentences which include a word in the development sentences that is not in the set made by the first step. However, we do not add any sentences by the second step. For example, the word "標題" occurs 19 times in the development set and is not included in the set made by the first step but it occurs no times in the whole training set. We make a TM from this adapted training data. We refer to this TM as TM1. The training sentence pairs for the TM1 for the test data are selected by the similar method.

We make another TM. Adding to the training sentences of TM1, we add additional training sentences from the RBMT outputs of the development data (or the test data). This means SPE part makes easier not to rewrite RBMT outputs. We refer to this TM as TM2. The training data for the TM2 of the development set includes 134,634 sentence pairs and the training data for the TM2 of the test set includes 129,737 sentence pairs.

2.2 Language model

For language model (LM) training for SPE part uses 2,000,000 sentences from train-1.txt and train-2.txt. We use SRILM (Stolcke, 2002) with 5-grams order and modified Kneser-Ney discounting.

2.3 Translation selection

We use a translation selection method with a quality estimation method described in Ehara (2013). Translation candidates are from three translations: RBMT output and two SPE outputs. One SPE method uses TM1 and the other SPE method uses TM2, which are described in 2.1. The bonus score (Ehara, 2013) for two SPE outputs are set to 0.1. As the result, 126, 850 and 836 outputs are selected from the outputs of RBMT, SPE with TM1 and SPE with TM2, respectively.

3 System architecture and experimental setting for the zh-ja task

For the zh-ja task, we use the same base system of ja-en task. But, we do not use TM adaptation by the data selection part in the Figure 1 and translation selection by the translation evaluation and selection part in the Figure 1, as the ja-en task. We, however, add some new things to our base system:

- To use a user dictionary in the RBMT part.
- To adapt a language model.
- To make additional sentence segmentation by the full-width space position.

3.1 User dictionary

A user dictionary for the RBMT part is made from the zh-ja training data. We make the phrase table from the training data using phrase-based Moses and extract high scored phrase pairs. We filter these selected phrase pairs using part of speech. As a result, we get the user dictionary having 1,246,274 entries all of which are nouns.

3.2 Translation model

We use whole 672,315 zh-ja training sentence pairs to make the TM for the SPE part. After additional sentence segmentation described the section 3.4, we get 679,292 training segment pairs. We do not use training data filtering like in the ja-en task.

3.3 Language model adaptation

For LM training, in addition to zh-ja training data, we add more sentences from the Japanese side of the ja-en training data. The method for adding is like TM adaptation for the ja-en task. For example, LM for the development set is adapted as follows. In the first step, we select training sentences of ja-en task (only Japanese side) which include a low frequency word in the RBMT output of the Chinese side of the development data. The low frequency word means the count of such word in the training sentences are less than a threshold (we set 300). In a second step, we add training sentences of ja-en task (only Japanese side) which include a word in the RBMT outputs of the Chinese side of the development data and that is not in the set made by the first step. LM adaptation for the test set is similar.

As the result, we get 1,109,647 Japanese sentences to train LM for the development set and 1,092,850 Japanese sentences to train LM for the

test set. By this adaptation, mean value of the perplexity for the Japanese side of the development set drops from 66.72 to 59.69.

3.4 Additional sentence segmentation

Several sentences in the zh-ja task include a full-width space character (" ") which is used to separate segments. For example, the following sentences in the development set include such full-width space.

「北方的水生食物连锁污染数据库　调查工具」
「北方の水食物連鎖汚染データベース　調査道具」

So, we make an additional segmentation to the training, development and test sets. When the number of the full-width spaces between Japanese side and Chinese side in training or development sentence pair is the same, the sentences are segmented at the full-width space position. For the test set, we make this segmentation at all full-width space position in Chinese sentences.

4 Issues for context-aware machine translation

TM adaptation described in 2.1 and LM adaptation described in 3.3 can be considered context-aware machine translation. Our method can be extended to document level adaptation. However, we do not make experiments that use non adapted models or adapted models not with the test set level but the test document level. One shortcoming of our adaptation method is that it needs re-training of TM and/or LM adapted to the input document, which is time consuming.

5 Evaluation results

Automatic evaluation results for the whole test sentences (zh-ja: 2107, ja-en: 1812) and human evaluation results for the selected test sentences (400) by the organizer are shown in Table 1 with our system rank and the number of all evaluated systems up to September 14th.

Task	BLEU	RIBES	HUMAN
ja-en	19.86 (9/27)	0.7067 (5/27)	22.50 (7/16)
zh-ja	33.53 (12/19)	0.8114 (6/19)	15.00 (5/11)

Table 1: Evaluation results

5.1 Human evaluation results

Comparing with the baseline system, the number of wins, ties and losses in human evaluation in the zh-ja task are 197, 66 and 137, respectively. An example of EIWA's win case is shown in Table2[1]. The EIWA's result is very similar to the reference.

EIWA	この発見は角野の推測に新たな根拠を与えていると考えられる。
base line	この発見により与えられると考えられる対角野の推測のために新しい。
reference	このことは角野の推測に新たな根拠を与えるものと考えられる。
source	可以认为这个发现对角野的推测给予新的根据。

Table 2: Example of EIWA's win case

Some examples of EIWA's loss cases are shown in Table 3.

EIWA	灯油漏洩を防止処置を扱った研究は非常に多く，化学処理方法の検討中の段階であった。
base line	石油流出の防止処理対策についての研究は非常に多く，化学の防犯処理方法は研究段階にある。
reference	流出油の防除手法についてはかなり研究が進んでいるが，ケミカルの防除手法はまだ調査研究の途中段階である。
source	关于石油泄漏的防范处理措施的研究非常多，通过化学的防范处理方法正处在研究阶段。

(a) Lexical mistranslation

EIWA	着目し，地盤材料として利用する人工石炭灰粒子として，すべての粒子の特性に基づいて，粒子の形状，物理的特性と単粒子の粉砕特性について検討を行った。
base line	個々の粒子の特性に基づいて，地盤材料としての利用の人工石炭灰粒子に着目し，粒子の形状，物理特性および単個の粒子の破砕特性について検討した。
reference	造粒石炭灰を地盤材料として利用することに着目し，粒子個々の特性に基づいて，粒子の形状や物理的特性および単粒子破砕特性について検討した。
source	着眼于作为地基材料予以利用的人造煤灰颗粒，基于每个粒子的特性，对粒子的形状、物理特性以及单颗粒子的破碎特性进行了讨论。

(b) Syntactic mistranslation

Table 3: Examples of EIWA's loss cases

[1] All sample sentences (source and reference sentences) in this document are provided by the Asian Scientific Paper Excerpt Corpus.

In Table 3 (a), EIWA's translation of the Chinese expression "石油泄漏" is "灯油漏洩" compared with the baseline translation "石油流出". This case has lexical mistranslation in EIWA.

In Table 3 (b), Chinese subordinate clause including "着眼" is incorrectly parsed by EIWA's RBMT part. This case has syntactic mistranslation in EIWA.

5.2 Correlation between automatic and human evaluation results

For human evaluated 400 sentences of zh-ja task, we conduct automatic evaluations with RIBES (Isozaki et al., 2010) and IMPACT (Echizen-ya and Araki, 2007).

Here we use "human average score (HUM)" which means the average value of scores of the three human annotators. DIFF_RIBES is the difference of RIBES scores of EIWA's output and baseline output. DIFF_IMPACT, also, means the difference of IMPACT scores of EIWA's output and baseline output. All these scores are in the interval [-1,1]. The scattering graph between HUM and DIFF_RIBES is shown in the Figure 2. The scattering graph between HUM and DIFF_IMPACT is shown in the Figure 3.

Pearson's correlation coefficient between HUM and DIFF_RIBES is 0.3618 and Pearson's correlation coefficient between HUM and DIFF_IMPACT is 0.3686. They are weak correlations.

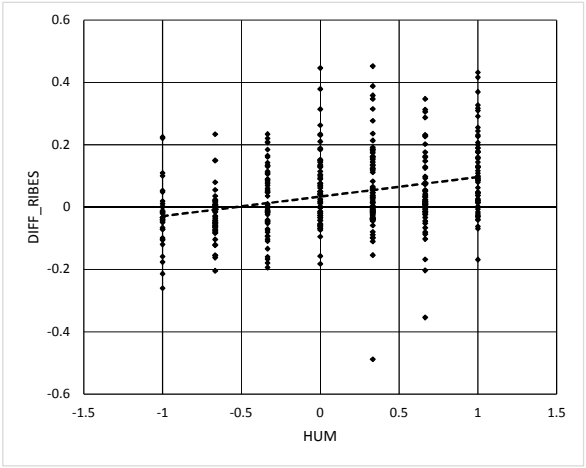

Figure 2: Correlation between HUM and DIFF_RIBES

Table 4 shows examples having big difference between HUM and DIFF_RIBES. At Table 4 (a), lexical error "暗号陽２３号" may affect HUM score to be negative, while DIFF_RIBES is positive because sentence structures of EIWA and reference is more similar than the sentence structures of baseline and reference.

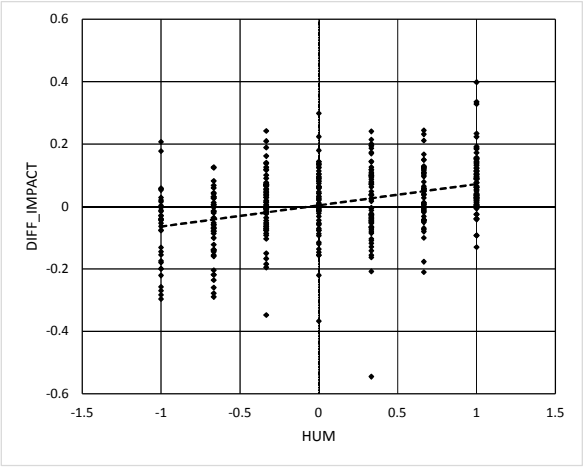

Figure 3: Correlation between HUM and DIFF_IMPACT

At Table 4 (b), the main part of the sentence, "Ｅｍｍａ／ＱｏＳを提案した" in the reference, is correctly translated in EIWA. So, HUM may be 1. It is not clear that DIFF_RIBES is negative.

At Table 4 (c), baseline translation is perfectly equal to the reference translation. So, RIBES score of baseline is 1 then DIFF_RIBES is negative. However, two annotators evaluated EIWA's win and one annotator evaluated baseline's win. The reason may be that EIWA's translation is more literal than baseline translation. Another reason may be that the references were not shown to annotators when they made evaluation. Multiple references by additional translation or scrambling of reference (Isozaki et al., 2014) may make DIFF_RIBES be smaller.

6 Conclusion

System architecture, experimental setting and evaluation results of EIWA are described. We are in the middle position in human and automatic evaluation. One of the future issues is to improve parsing accuracy in the RBMT part. Syntactically collapsed RBMT outputs cannot be recovered by the SPE part. Other future issue is to combine rule-based technique and statistical technique more tightly beyond RBMT plus SPE method.

EIWA	水稲のカドミウムの高吸収特性を示す品種「暗号陽23号」を用いた。
base line	カドミウム高吸収特性の品種「密阳23号」を示すイネを採用した.
reference	イネはカドミウム高吸収特性を示す品種「密陽23号」を用いた。
source	稻子采用了显示镉高吸收特性的品种“密阳23号”。

(a) HUM= -1, DIFF_RIBES=0.222

EIWA	本稿では，ALMプランEmma／QoSを提案し，オーバレイネットワークでは，同時に複数のビデオを伝送を実現する場合，分散制御のQoSである.
base line	Emma／QoSはオーバレイネットワークでは，同時に複数のビデオを伝送を実現する方案を提案した分散制御のQoSをALM
reference	本論文では，オーバレイネットワークにおける複数ビデオの同時配信時にQoSを分散制御で実現するALMプロトコルEmma／QoSを提案した.
source	本文提出了ALM方案Emma／QoS，在Overlay网络中，实现同时传输复数视频时分散控制QoS。

(b) HUM=1, DIFF_RIBES= -0.1678

EIWA	表3に人工株式市場の各エージェントの予測の木で用いる記号を構成することを示した。
base line	人工株式市場を構成する各エージェントの予測の木に用いる記号を表3に示す.
reference	人工株式市場を構成する各エージェントの予測の木に用いる記号を表3に示す.
source	表3显示构成人工股份市场的各代理商的预测树中使用的记号。

(c) HUM=0.3333, DIFF_RIBES=-0.4873

Table 4: Examples having big difference between HUM and DIFF_RIBES

Reference

Loïc Dugast, Jean Senellart and Philipp Koehn. 2007. Statistical Post-Editing on SYSTRAN’s Rule-Based Translation System. *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220-223.

Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, *Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI)*, pages 151-158.

Terumasa Ehara. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. *Proceedings of Machine Translation Summit XI, Workshop on Patent Translation*, pages13-18.

Terumasa Ehara. 2010. Machine translation for patent documents combining rule-based translation and statistical post-editing. *Proceedings of NTCIR-8 Workshop Meeting*, pages 384-386.

Terumasa Ehara. 2011. Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the PatentMT Task. *Proceedings of NTCIR-9 Workshop Meeting*, pages 623-628.

Terumasa Ehara. 2013. Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the NTCIR-10 PatentMT Task. *Proceedings of the 10th NTCIR Conference*, pages 335-338.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada and Kevin Duh. 2010. Head finalization: A simple reordering rule for SOV languages. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 244-251.

Hideki Isozaki, Natsume Kouchi and Tsutomu Hirao. 2014. Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 287–292.

Philipp Koehn, Franz J. Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation, *Proceedings of HLT-NAACL 2003*, pages 48-54.

Michel Simard, Nicola Ueffing, Pierre Isabelle and Roland Kuhn. 2007. Rule-based Translation With Statistical Phrase-based Post-editing. *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203-206.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. *Proceedings of the International Conference on Spoken Language Processing*.