SMT reranked NMT (2)

Terumasa Ehara
Ehara NLP Research Laboratory
Seijo, Setagaya, Tokyo Japan
http://www.ne.jp/asahi/eharate/eharate/

Abstract

System architecture, experimental settings and experimental results of the EHR team for the WAT2018 tasks are described. System architecture is same as the WAT2017 submission. We participate four tasks this time. We can improve 1.5 to 3.0 BLEU score compared with the baseline scores. On the other hand, the pairwise evaluation score of our system to the baseline is negative in ASPEC en-ja task. Our technique also decreases the number of under-translation that is frequently appears in a NMT output.

1 Introduction

Although the NMT provides high quality and fluent translations, it has several drawbacks. One of them is under- and over-translation which is infrequent in a SMT output. We have proposed a reranking method for n-best NMT outputs using a SMT output (Ehara, 2017). Ehara (2017) shows under-translation can be reduced by the method. This time, we use the same technique. Figure 1 shows our system architecture.

![System architecture](image)

Figure 1: System architecture

An input sentence is fed to NMT part and to SMT part. N-best outputs of NMT are reranked by the SMT output as a reference. Reranking measure is BLEU, this time, instead of IMPACT in WAT2017. NMT translation i having highest BLEU score is output as a system output.

2 Experimental setting

In WAT2018 (Nakazawa et al., 2018), we participate four tasks: ASPEC en-ja, JPC zh-ja, JPC ko-ja and JPC en-ja tasks. We use OpenNMT (Minh-Thang Luong et al., 2015) in NMT part. Our SMT part is phrase-based SMT by Moses v.3 (Koehn et al., 2003) with default option settings. We adopt preordering in SMT for en-ja and zh-ja tasks. Option settings for NMT part is described in Ehara (2017). Segmentation policies for NMT are as follows. In JPC zh-ja task, zh part is segmented by a character and ja part is segmented by a word except for alpha-numeric and symbol character sequences. Alpha-numeric and symbol character sequences are segmented by a character. In JPC ko-ja task, ko part and ja part are both segmented by a character. In JPC en-ja task and ASPEC en-ja task, en parts and ja parts are both segmented by a sub word. For JPC en-ja task, vocabulary size of en part is 60,217 and vocabulary size of ja part is 94,542. For ASPEC en-ja task, vocabulary size of en part is 62,203 and vocabulary size of ja part is 107,145.

3 Experimental results

Experimental results are shown in Table 1 with training data size (number of sentences) and final epoch number of NMT training. Training data for JPC en-ja task include WAT2018’s training data and NTCIR-10’s training data (Goto et al., 2013). The differences of three data ID in four subtasks of
JPC en-ja task are epoch number of NMT training. They are 13, 18 and 20. We can improve 1.5 to 3.0 BLEU score compared with the baseline scores. JPO adequacy scores are similar to the last workshop’s scores of EHR team (Adeq/2017).

Table 1: Experimental results

| Data point | Sentence BLEU | Pairwise score |
|------------|---------------|----------------|
| N1         | 37.97         | -0.5           |
| N2         | 34.23         | -0.5           |
| N3         | 34.23         | -0.5           |
| N4         | 34.23         | -0.5           |

Figure 2: Scattering graph of diff BLEU and pairwise evaluation score (ASPEC en-ja task)

Table 2: Contingency table of diff BLEU and pairwise evaluation score.

| Diff BLEU | Pairwise | < 0 | ≥ 0 |
|-----------|----------|-----|-----|
| < 0       | 94       |     | 84  |
| ≥ 0       | 85       |     | 137 |

Table 3: Number of under- and over-translation in unreranked translation and reranked translation for ASPEC en-ja task test data

| normal | under | over | under & over | total |
|--------|-------|------|--------------|-------|
| 10     | 10    | 6    | 1            | 1516  |
| 1      | 1     | 1    | 1            | 1516  |
| 1516   |       |      |              |       |
|        |       |      |              |       |

From Table 3, we can see that in 1812 test sentences, 286 unreranked outputs include under-translation and only 85 reranked outputs include under-translation. On the other hand, in 1812 test sentences, 14 unreranked outputs include over-translation and 23 reranked outputs include over-translation. We can say our reranking technique decreases the number of under-translation, but cannot decrease the number of over-translation. Several examples of these data are described in the next section.

mteval-v13a.pl (ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl) is used to calculate sentence level BLEU.
4 Analysis of translation results

4.1 Comparison of translation results with reranking and without reranking

We compare translation results with reranking and without reranking in this subsection for ASPEC en-ja task. We compute sentence level BLEU of both translation results. Table 4 shows the cases that the BLEU score of a translation with reranking is higher than the BLEU score of a translation without reranking. Table 5 shows the opposite cases.

In these tables, “snt. #” means sentence number of the test set, “src” means source sentence, “ref” means reference sentence, “smt” means translation result by the SMT part, “unreranked” means translation result without reranking, “reranked” means translation result with reranking by our method, “BLEU_f” means BLEU score of the translation result with reranking, “BLEU_r” means BLEU score of the translation result without reranking. Table 4 shows the cases where the BLEU score of a translation with reranking is higher than the BLEU score of a translation without reranking. Table 5 shows the cases where the BLEU score of a translation without reranking is higher than the BLEU score of a translation with reranking.

From Table 4, we can see several under-translations in the unreranked are corrected in reranked (see snt. # 1269 and 44).
### Table 4: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 233    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 233    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 5: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 609    | テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 609    | テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| テロメラーゼ活性は生殖細胞と癌細胞| 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 6: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 1230   | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 1230   | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 7: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 1146   | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 1146   | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 8: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 544    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 544    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 9: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 817    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 817    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.

### Table 10: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is higher than the BLEU score of translation without reranking)

| Item # | src | ref | smt | unre reranked | BLEU f | BLEU r | remarks |
|--------|-----|-----|-----|---------------|--------|--------|---------|
| 817    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |
| 817    | 本稿 | 本稿 | 本稿 | 本稿 | 0.8731 | 0.6921 | 填充物を必要とするメカニカルシステム研究 (著者による補充) |

Remarks: This paper is not translated in ref and unre reranked.
4.2 Comparison of evaluation results by BLEU and pairwise evaluation

In this subsection, we examine the contradiction between BLEU score and pairwise evaluation score in ASPEC en-ja task. Several examples that differ BLEU $\geq 0$ and pairwise score $<0$ are listed in Table 6. Several examples that differ BLEU $<0$ and pairwise score $\leq 0$ are listed in Table 7.

In these tables, “baseline” means baseline translation result, “BLEU $b$” means BLEU score of baseline and “pairwise” means pairwise evaluation score (sum of five evaluator’s score).

Table 5: Comparison of translation results with reranking and without reranking (case that the BLEU score of translation with reranking is lower than the BLEU score of translation without reranking)

Table 6: Comparison of baseline translation and reranked translation (case that diff BLEU $\geq 0$ and pairwise score $<0$)

From Table 6, we can see two reasons why diff BLEU $\geq 0$ and pairwise score $<0$. First is that important information is differ in ref and reranked and not so important information that is expressed
in long phrase is differ in ref and baseline and is same in ref and reranked. Because of it, diff BLEU $\geq 0$ and pairwise score $< 0$. Second reason is that literal translation in baseline and non-literal translation in ref and reranked.

Table 7: Comparison of baseline translation and reranked translation (case that diff BLEU $< 0$ and pairwise score $\geq 0$)

| item | value |
|------|-------|
| src | No side effect was noted during treatment. |
| ref | 治療中。副作用は認めなかった。 |
| baseline | 副作用は認めなかった。 |
| reranked | 副作用は認められなかった。 |
| BLEU_b | 0.6065 |
| BLEU_r | 0.4033 |
| pairwise | 5 |
| remarks | Important information "during treatment" is not translated in baseline. Long phrase "副作用は認めなかった。" is same in ref and baseline, and "副作用は認められなかった。" in reranked has similar meaning. |

Table 7: Comparison of baseline translation and reranked translation (case that diff BLEU $< 0$ and pairwise score $\geq 0$)

| item | value |
|------|-------|
| src | Cutting, patterning polishing, and metalizing work of diamond films with laser beams are reviewed. |
| ref | レーザによるダイヤモンド膜の切断、パターンニング、研磨、金属 化加工について総報した。 |
| baseline | レーザビームを用いたダイヤモンド膜の切断、パターンニング、および金属化作業についてレビューした。 |
| reranked | レーザビームを用いたダイヤモンド膜の切断、パターンニング、研磨、金属化についてレビューした。 |
| BLEU_b | 0.3412 |
| BLEU_r | 0.1592 |
| pairwise | 3 |
| remarks | Important information "metalizing work" is literally translated to "メタライゼーション作業" in baseline and non-literally translated to "金属化" in ref and reranked. Long phrase "ダイヤモンド膜の切断、パターンニング" is same in ref and baseline and "ダイヤモンド膜の切断、パターンニング" in reranked has similar meaning. |

Table 7: Comparison of baseline translation and reranked translation (case that diff BLEU $< 0$ and pairwise score $\geq 0$)

| item | value |
|------|-------|
| src | Through the detailed analysis of the intensity dependence of excited light, this is judged to be due to photoionization. |
| ref | 青と先の強度依存性の詳細な解析により、光イオン化によるものと判断した。 |
| baseline | 青と先の強度依存性の詳細な解析により、光イオン化によるものと判断した。 |
| reranked | 青と先の強度依存性を詳細に解析して、これは光イオン化によるものと判断した。 |
| BLEU_b | 0.562 |
| BLEU_r | 0.4535 |
| pairwise | 3 |
| remarks | Ref, baseline and reranked almost have similar meaning. Ref and baseline are literally near but less fluent than reranked. |

5 Conclusion

We participate four tasks of WAT2018 with the technique of SMT reranked NMT. This technique can reduce under-translation that is frequently found in NMT. For over-translation, the technique, however, cannot reduce it.

References

Terumasa Ehara. 2017. SMT reranked NMT. Proceedings of the 4th Workshop on Asian Translation (WAT2017), pages 119–126.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita and Benjamin K. Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. Proceedings of the 10th NTCIR Conference, pages 260-286.

Toshiaki Nakazawa, Shohei Higashiyama, Chench, Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh and Sadao Kurohashi. 2018. Overview of the 5th Workshop on Asian Translation, Proceedings of the 5th Workshop on Asian Translation (WAT2018).

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

Minh-Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421.