2W-04

日英機械翻訳におけるモデルと入力文の相性判定

野口 夏希† 梶原 智之‡

† 愛媛大学工学部工学科 ‡ 愛媛大学大学院理工学研究科

1 はじめに

深層学習の技術の発展に伴い機械翻訳の性能は向上しているものの、流暢な誤訳[1,2]が依然として大きな課題として残っている。目的言語の知識が少ない使用者は、流暢な誤訳をそのまま使用してしまう恐れがあり、医療現場やビジネスシーンなど様々な場面で問題となっている。

本研究では、機械翻訳による誤訳を未然に防ぐことを目的に、翻訳器と入力文の相性判定を行う。所与の翻訳器において誤訳につながる可能性が高い入力文を「相性が悪い文」と定義し、これを検出する。翻訳器との相性が悪い入力文の検出により、機械翻訳の使用者に入力文の推敲を促すことができ、誤訳の防止を期待できる。

日本語から英語への機械翻訳における実験の結果、我々が構築した翻訳器とオンライン翻訳器の両方で、相性が良い文と悪い文の間に顕著な翻訳品質の差が見られた。この結果から、翻訳器と入力文の相性判定の有効性を確認できた。

2 提案手法

翻訳器と入力文の相性の良し悪しを判定する分類器を訓練するために、まず訓練用のラベル付きコーパスを作成する。図 1 に示すように、相性判定の対象となる翻訳器を用いて、対訳コーパスに含まれるソース言語の文を翻訳し、文単位で翻訳品質を評価する。

翻訳品質が閾値 θ_H を上回るソース文を「相性が良い文」として、閾値 θ_L を下回るソース文を「相性が悪い文」としてそれぞれ抽出し、相性判定器のための訓練用コーパスを作成する。相性判定器として、このラベル付きコーパスを用いて、入力文を 2 値分類する分類器を訓練する。

機械翻訳を行う際には、翻訳の前に入力文と翻訳器の相性を判定する。ここで、相性が悪いと判定された文は、自動または手動で推敲して再入力することを想定している。相性が良いと判定された文のみを翻訳器に入力し、翻訳する。

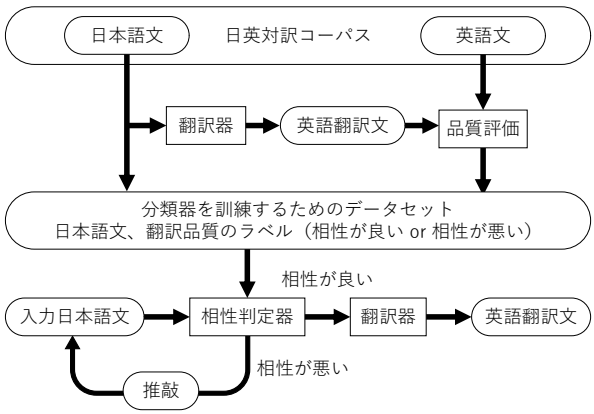

図 1：本研究の概要

3 実験設定

本研究では、Joey NMT[1]を用いて実装した翻訳器およびオンライン翻訳器である Google 翻訳[2]の 2 種類の日英翻訳器を対象に、提案手法の有効性を検証した。前者の翻訳器には、512 次元の埋込層および隠れ層を持つ 6 層 8 注意ヘッドの Transformer モデル[3]を JParaCrawl[4]の約 1,000 万文対の対訳コーパス上で訓練した。バッチサイズを 4,096 トークンとし、最適化手法には adam を使用した。前処理には SentencePiece[3]の 1-gram 言語モデル（語彙サイズは 32,000）によるサブワード分割を行った。

相性判定器を構築するために、京都フリー翻訳タスク（KFTT）[4]の対訳コーパスを用いた。文単位の翻訳品質は、SacreBLEU[5]を用いて Sentence BLEU を評価した。翻訳器と入力文の相性の良し悪しを判定するための閾値は、$\theta_H = 30$、$\theta_L = 10$ とした。そして、表 1 に示すように、該当する日本語文を訓練用コーパスから 3 万文ずつ、検証用コーパスおよび評価用コーパスから 150 文ずつ無作為抽出した。

相性判定のための 2 値分類器は、事前訓練された BERT[6][5]を再訓練して構築した。バッチサ

Suitability Estimation between Models and Input Sentences for Machine Translation
† Natsuki Noguchi (n_noguchi@ai.cs.ehime-u.ac.jp)
‡ Tomoyuki Kajiwara (kajiwara@cs.ehime-u.ac.jp)
Ehime University

[1] https://github.com/joeynmt/joeynmt
[2] https://cloud.google.com/translate
[3] https://github.com/google/sentencepiece
[4] http://www.phontron.com/kftt/
[5] https://github.com/mjpost/sacrebleu
[6] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

表 1：コーパスのサイズ

	訓練用	検証用	評価用
KFTT の文対	440,288	1,166	1,160
相性が良い文	30,000	150	150
相性が悪い文	30,000	150	150

表 2：実験結果（正解率が実験 1、BLEU が実験 2）

	正解率	相性が良い文の BLEU	相性が悪い文の BLEU
Joey NMT	0.777	21.72	9.47
Google 翻訳	0.707	27.00	13.50

表 3：相性判定の混同行列

Joey NMT	相性が良い	相性が悪い
相性が良いと判定	128	45
相性が悪いと判定	22	105
Google 翻訳	相性が良い	相性が悪い
相性が良いと判定	127	62
相性が悪いと判定	23	88

表 4：相性判定器を交換した際の翻訳品質

	相性が良い文の BLEU	相性が悪い文の BLEU
Joey NMT	20.68	8.43
Google 翻訳	29.33	14.09

イズを 32 文、学習率を 1e-5 として、相性判定器の訓練用コーパス 6 万文を用いてクロスエントロピー損失最小化の訓練を 5 エポック行った。

相性判定器の性能を評価するために、2 種類の実験を行った。実験 1 では、相性判定器のための評価用コーパス 300 文を用いて、相性判定の 2 クラス分類の正解率を評価した。実験 2 では、KFTT の評価用コーパス 1,160 文対を用いて、相性が良いと判定された入力文の翻訳品質と相性が悪いと判定された入力文の翻訳品質を比較した。実験 2 におけるコーパス単位の翻訳品質は、SacreBLEU を用いて Corpus BLEU を評価した。

4 実験結果

4.1 実験 1：分類品質による内的評価

表 2 に実験結果を示す。Joey NMT と Google 翻訳の両方で、相性判定は 70%を超える正解率を達成した。全体的な翻訳品質は Google 翻訳の方が高いものの、Joey NMT に基づく相性判定器の方が高品質な相性判定を実現できた。

表 3 に相性判定の混同行列を示す。Joey NMT と Google 翻訳の両方で、相性が悪い文を相性が良いと誤判定する例が比較的多い。最終的な翻訳品質を高めるためには、相性が悪い入力文を漏れなく検出することが重要なため、混同行列の右上を更に改善することが今後の課題である。

4.2 実験 2：翻訳品質による外的評価

表 2 の実験結果から、Joey NMT と Google 翻訳の両方で、相性が良いと判定された入力文と相性が悪いと判定された入力文の間に顕著な翻訳品質の差が見られた。Joey NMT では 12.25 ポイント、Google 翻訳では 13.5 ポイントの BLEU の差が見られるため、提案手法が所与の翻訳器に適さない入力文の検出に成功していると言える。

表 4 に、相性判定器を交換した際の翻訳品質を示す。つまり、Google 翻訳のための相性判定器を用いて Joey NMT による機械翻訳を行い、Joey NMT のための相性判定器を用いて Google 翻訳による機械翻訳を行った。相性判定器を交換した場合でも、Joey NMT において 12.25 ポイント、Google 翻訳において 15.24 ポイントの BLEU の差が見られたことから、本研究で構築した相性判定器の有効性は翻訳器に依存しないことがわかる。つまり、この相性判定器は、入力文と任意の翻訳器の相性を判定できる。

5 おわりに

本研究では、機械翻訳による誤訳を防ぐために、翻訳器と入力文の相性判定に取り組んだ。日英機械翻訳における実験の結果、約 70%の正解率で相性判定を実現でき、相性が良い文と悪い文の間に顕著な翻訳品質の差を確認できた。また、我々の相性判定器が特定の翻訳器に依存せず有効であることが明らかになった。

今後の課題として、他のドメインや言語対における検証や、相性が悪い入力文の検出漏れの改善に取り組みたい。また、翻訳器と相性が悪い入力文の自動的な前編集の技術も検討したい。

参考文献

[1] Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, Hang Li. Context Gates for Neural Machine Translation. TACL, Vol.5, pp.87-99, 2017.
[2] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, Jie Zhou. Prevent the Language Model from being Overconfident in Neural Machine Translation. In Proc. of ACL, pp.3456-3468, 2021.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention Is All You Need. In Proc. of NIPS, pp.5998-6008, 2017.
[4] Makoto Morishita, Jun Suzuki, Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In Proc. of LREC, pp.3603-3609, 2020.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL, pp.4171-4186, 2019.