入力文と自然言語処理モデルの相性判定

野口 夏希
愛媛大学工学部
n_noguchi@ai.cs.ehime-u.ac.jp

梶原 智之
愛媛大学大学院理工学研究科
kajiwara@cs.ehime-u.ac.jp

概要

本研究では，入力文のみから自然言語処理モデルの出力品質を推定する手法を提案する．機械翻訳を主な対象とする既存の品質推定は，入力文と出力文の組から出力文の品質を推定する技術である．モデルによる出力文を必要とするため，既存の品質推定の手法はテキスト分類のタスクには適用できない．我々が提案する新たな品質推定の技術は入力文のみを用いるため，文の生成と分類の両タスクに対して適用できる．また，対象タスクにおける自然言語処理モデルの実行が不要なため，大規模なコーパスに対しても高速に文単位の品質推定を実施できる．

1 はじめに

深層学習の技術の発展に伴い，機械翻訳 [1] や感情極性分類 [2] など，多くの自然言語処理タスクの性能が向上している．しかし，全体的な性能が向上しているとはいえ，機械翻訳における流暢な誤訳 [3, 4] のように，個々の事例の中には依然として大きな課題が残されている．そのため，文単位での品質推定の技術に大きな期待が寄せられている．

機械翻訳 [5] やテキスト平易化 [6] などのテキスト生成タスクにおいて研究されている文単位の品質推定は，入力文および出力文の文対が与えられ，出力文の品質を推定する．入力文とともに出力文を用いるため，既存の品質推定の手法 [7, 8] は感情極性分類などのテキスト分類タスクには適用できない．

本研究では，任意の自然言語処理タスクに対して共通に適用可能な文単位の品質推定のフレームワークを提案する．提案手法は，入力文のみから自然言語処理モデルの出力品質を推定する．つまり，自然言語処理モデルを評価するのではなく，入力文が自然言語処理に適しているか否かを評価する．入力文のみを用いるため，提案手法は機械翻訳などのテキスト生成タスクだけでなく，感情極性分類などのテキスト分類タスクにも適用できる．また，対象タスクにおける自然言語処理モデルの実行が不要なため，大規模なコーパスに対しても高速に文単位の品質推定を実施できるという利点を持つ．

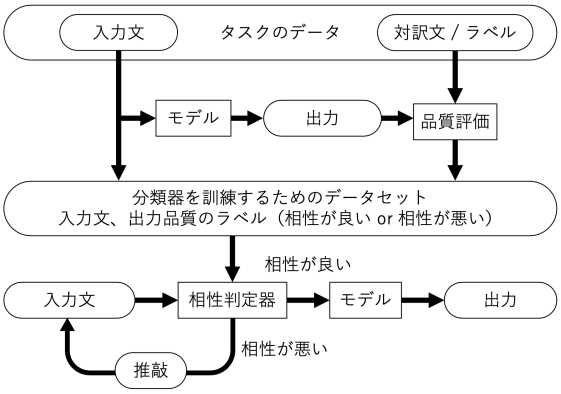

図 1 提案手法の概要

機械翻訳および感情極性分類のタスクを対象に，高品質な出力が期待できる入力文か否かの 2 値分類として文単位の品質推定の評価実験を行った．実験の結果，タスクやドメインに依存せず，7 割以上の正解率で入力文のみから文単位の品質推定を実現できた．また，高品質な出力が期待できると判定された文集合と低品質な出力が予想されると判定された文集合の間で，Transformer [1] または BERT [9] を用いてタスクを解いた際の実際の出力品質を評価したところ，機械翻訳においては最大で 14.6 ポイントの BLEU [10]，感情極性分類においては最大で 27.8 ポイントの正解率という大きな差が見られ，提案手法の有効性を確認できた．

2 提案手法

本研究では，所与の入力文に対して，高品質な出力が期待できるか否かの 2 値分類を行う品質推定器を構築する．本研究で提案する品質推定器は，入力文と出力文の対から出力文の品質を推定する一般的な品質推定器と区別するために，入力文と自然言語処理の相性を判定するものとして以降では相性判定器と呼ぶ．提案する相性判定の概要を図 1 に示す．

2.1 データセットの作成

対象タスクのラベル付きコーパスを 2 つ[1)]用意し，相性判定器を訓練するためのデータセットを構築する．はじめに，テキスト生成タスクであれば Transformer [1]，テキスト分類タスクであれば BERT [9] など，対象タスクを解くための任意の自然言語処理モデルを用意し，一方のラベル付きコーパスを用いて訓練する．次に，他方のラベル付きコーパスを入力として，訓練済みモデルを用いて出力を得る．そして，テキスト生成タスクであれば BLEU [10]，テキスト分類タスクであれば正解率など，対象タスクに応じた任意の評価指標を用いて，出力の品質を自動評価する．この自動評価の結果に基づき，各入力文に対して自然言語処理モデルとの相性の良し悪しを 2 値でラベル付けし，相性判定器を訓練するためのデータセットを構築する．

2.2 相性推定器の訓練

2.1 節で構築したデータセットを用いて，入力文と自然言語処理の相性を判定する 2 値分類器を訓練する．相性判定器を使用する際には，相性が悪いと判定された入力文は誤った出力につながる可能性が高いため，人手で推敲し，再入力することを想定している．この推敲の作業も自然言語処理の技術を用いて自動化したいが，これは今後の課題である．

3 実験設定

機械翻訳および感情極性分類の評価実験を行う．

3.1 データ

機械翻訳 日本語およびドイツ語から英語への機械翻訳を行う．翻訳器を訓練するために，日英の言語対では JParaCrawl[2)] [11] の約 1,000 万文対，独英の言語対では WMT[3)] [12] の約 500 万文対を用いた．

相性判定器を訓練および評価するために，ドメインの異なる複数の対訳データを用いた．日英の言語対では，講演字幕の IWSLT [13] および Wikipedia の KFTT[4)]を用いた．独英の言語対では，講演字幕の IWSLT および Wikipedia の WikiMatrix [14] を用いた．相性判定のためのラベル付きコーパスを作成するために，翻訳器の出力文を SentenceBLEU [10] によって文単位で自動評価し，閾値 θ_H を上回る入力文を正例（入力文と自然言語処理の相性が良い），閾値 θ_L を下回る入力文を負例（入力文と自然言語処理の相性が悪い）としてアノテーションした．本実験では，$\theta_H = 30$，$\theta_L = 10$ として，表 1 の件数を無作為抽出した．

表 1 相性判定器の内的評価に用いたデータの件数

	訓練用	検証用	評価用
正例	10,000	150	150
負例	10,000	150	150

感情極性分類 日本語の感情極性 2 値分類を行う．感情極性分類器を訓練するために，ACP コーパス[5)] [15] の中から無作為抽出した 1 万文を用いる．

相性判定器を訓練および評価するために，同じく ACP コーパスを用いる．ただし，分類器の訓練に用いた文は相性判定器の訓練や評価には使用しない．相性判定のためのラベル付きコーパスは，分類器の出力ラベルが正解ラベルと一致する入力文を正例，一致しない入力文を負例としてアノテーションし，表 1 の件数を無作為抽出した．

3.2 モデル

機械翻訳 JoeyNMT[6)] [16] 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[7)]の 2 種類の日英翻訳器を使用した．前者の翻訳器は，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 [1] を訓練した．バッチサイズを 4,096 トークンとし，最適化手法には Adam [17] を使用した．前処理には SentencePiece[8)] [18] 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．

感情極性分類 Wikipedia[9)]および SNS テキスト[10)]を用いて事前訓練された 2 種類の日本語 BERT [9] を再訓練した．バッチサイズを 32 文とし，最適化手法には Adam を使用した．

1) 2 種類のラベル付きコーパスを用意しても良いが，1 種類のラベル付きコーパスを 2 分割しても良い．
2) https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
3) https://www.statmt.org/wmt17/translation-task.html
4) http://www.phontron.com/kftt/
5) https://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/
6) https://github.com/joeynmt/joeynmt
7) https://cloud.google.com/translate
8) https://github.com/google/sentencepiece
9) https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
10) https://github.com/hottolink/hottoSNS-bert

表 2　相性判定器の外的評価に用いた評価データの件数

IWSLT（日英）	3,833
KFTT	1,160
IWSLT（独英）	3,378
WikiMatrix	5,000
ACP コーパス	2,400

相性判定　機械翻訳では，日本語とドイツ語の両方に対応するために，相性判定器として事前訓練された多言語符号化器 XLM-RoBERTa[11] [19] を再訓練した．感情極性分類では，相性判定器として事前訓練された日本語 BERT[9] を再訓練した．バッチサイズを 32 文とし，最適化手法には Adam を使用した．

3.3　比較手法

相性判定のベースラインとして，Transformer [1] に基づくニューラル言語モデルのパープレキシティを用いる．言語モデルのパープレキシティは，多くの品質推定モデル [5] において採用されている素性のひとつであり，入力文のみから得られる品質推定素性の代表的なものである．本実験では，翻訳器や感情極性分類器の訓練に用いた訓練用コーパス上でニューラル言語モデルを訓練し，閾値を設定して相性判定の 2 値分類に使用した．なお，閾値は表 1 の検証用コーパスにおける 2 値分類の正解率を最大化するように設定した．

3.4　評価

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．内的評価として，表 1 の評価用コーパスを用いて，相性判定の 2 値分類の正解率を評価した．外的評価として，表 2 の評価用コーパスを用いて，タスクごとの性能評価を行った．外的評価に用いる評価用コーパスは，3.1 節においてサンプリングする前の各タスクにおける評価用コーパス全体である．機械翻訳の外的評価は，相性判定器によって分類された各文集合において BLEU [10] を評価した．なお，BLEU の計算には SacreBLEU[12] [20] を用いた．感情極性分類の外的評価は，相性判定器によって分類された各文集合において感情極性分類の正解率を評価した．

11）https://huggingface.co/xlm-roberta-base
12）https://github.com/mjpost/sacrebleu

4　実験結果

4.1　機械翻訳

表 3 に機械翻訳の実験結果を示す．相性判定の正解率に関する内的評価に着目すると，ニューラル言語モデルのパープレキシティに基づくベースライン相性判定器が 6 割弱の正解率である一方で，我々の相性判定器は 7 割を超える正解率を達成した．

BLEU に関する外的評価に着目すると，ニューラル言語モデルのパープレキシティに基づくベースライン相性判定器では相性が良いと判定された文集合と相性が悪いと判定された文集合の間で BLEU の差が最大で 2.7 ポイントしかなく，入力文と翻訳器の相性を充分に推定できていない．一方で提案手法では，一貫して 10 ポイント以上の大きな BLEU の差が見られるため，入力文と翻訳器の相性を良く推定できている．特に，IWSLT における独英翻訳の設定では，相性が良いと判定された文集合と相性が悪いと判定された文集合の間で 14.6 ポイントという顕著な BLEU の差を確認できた．

本実験では，日英および独英の 2 つの言語対を評価したり，言語対ごとに講演字幕および Wikipedia の 2 つのドメインを評価したり，IWSLT における日英翻訳において JoeyNMT で実装した Transformer および Google 翻訳の 2 つの翻訳器を評価したりと，様々な設定で評価を行ったが，提案手法は全ての設定において 7 割強の正解率を達成し，10 ポイント以上の顕著な BLEU の差を示した．これらの実験結果から，本研究で提案する相性判定器は，言語・ドメイン・翻訳器に依存せず，入力文と自然言語処理の相性を高精度に推定できると言える．

4.2　感情極性分類

表 4 に感情極性分類の実験結果を示す．相性判定の正解率に関する内的評価に着目すると，ニューラル言語モデルのパープレキシティに基づくベースライン相性判定器が 6 割程度の正解率である一方で，我々の相性判定器は 8 割程度の正解率を達成した．

感情極性分類の正解率に関する外的評価に着目すると，ニューラル言語モデルのパープレキシティに基づくベースライン相性判定器では相性が良いと判定された文集合と相性が悪いと判定された文集合の間で正解率の差が 4 ポイントしかなく，入力文と分類器の相性を充分に推定できていない．一方で提案

表 3 機械翻訳における実験結果

言語対	データ	翻訳器	相性判定器	正解率	BLEU	
					相性が良い文	相性が悪い文
日英	IWSLT	JoeyNMT	Perplexity	59.3	10.3	10.0
日英	KFTT	JoeyNMT	Perplexity	54.3	12.3	12.0
独英	IWSLT	JoeyNMT	Perplexity	55.3	31.3	28.6
独英	WikiMatrix	JoeyNMT	Perplexity	55.6	14.3	13.7
日英	IWSLT	JoeyNMT	XLM-RoBERTa	75.0	19.9	9.6
日英	IWSLT	Google 翻訳	XLM-RoBERTa	73.0	21.4	9.5
日英	KFTT	JoeyNMT	XLM-RoBERTa	78.0	20.6	10.4
独英	IWSLT	JoeyNMT	XLM-RoBERTa	73.6	31.5	16.9
独英	WikiMatrix	JoeyNMT	XLM-RoBERTa	71.0	19.6	7.5

表 4 感情極性分類における実験結果

BERT	相性推定器	正解率	感情極性分類の正解率	
			相性が良い文	相性が悪い文
Wikipedia	perplexity	60.7	90.7	86.9
SNS	perplexity	61.7	90.9	86.9
Wikipedia	BERT	80.3	97.8	70.0
SNS	BERT	77.3	96.5	69.3

手法では，27 ポイント以上の大きな正解率の差が見られるため，入力文と分類器の相性を良く推定できている．本実験では，Wikipedia および SNS という訓練ドメインの異なる 2 つの BERT を用いて感情極性分類器を構築したが，提案手法は対象モデルに依存せず，入力文と感情極性分類器の相性を高精度に推定できた．

5 おわりに

本研究では，自然言語処理モデルの出力品質を文単位で推定するために，入力文と自然言語処理の相性を推定する 2 値分類のアプローチを提案した．既存の品質推定の手法とは異なり，入力文のみを使用して出力品質を推定するため，提案手法はテキスト生成だけでなくテキスト分類のタスクにも共通に適用できる．また，対象タスクにおける自然言語処理モデルの実行が不要なため，大規模なコーパスに対しても高速に文単位の品質推定を実施できる．

評価実験の結果，機械翻訳においては 7 割強，感情極性分類においては 8 割程度の正解率で，入力文のみから出力品質の高低を推定できた．また，高品質な出力につながると推定した入力文集合と低品質な出力につながると推定した入力文集合の間で，実際に出力品質を比較した結果，機械翻訳においては 10.2 ポイントから 14.6 ポイントという顕著な BLEU の差が確認でき，感情極性分類においては 27.2 ポイントから 27.8 ポイントという顕著な正解率の差が確認できた．さらに，提案手法は入力文の言語やドメイン，対象の自然言語処理モデルに依存せず，入力文と自然言語処理の相性を良く推定できた．

今後の課題として，本研究において低品質な出力につながると推定された入力文に対して，具体的にどの語句を修正すべきなのかを特定する語句単位の品質推定にも取り組みたい．また，入力文の自動的な前編集の技術も検討し，自然言語処理モデルの利用者による推敲コストを削減したい．

謝辞

本研究は JST（ACT-X，課題番号: JPMJAX1907），JSPS 科研費（基盤研究 B，課題番号：JP22H03651）および国立研究開発法人情報通信研究機構の委託研究（課題番号：225）による助成を受けたものです．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, pp. 5998–6008, 2017.

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. ecursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, 2013.

[3] Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. Context Gates for Neural Machine Translation. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 87–99, 2017.

[4] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. Prevent the Language Model from being Overconfident in Neural Machine Translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing**, pp. 3456–3468, 2021.

[5] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. Quality Estimation for Machine Translation. **Synthesis Lectures on Human Language Technologies**, Vol. 11, No. 1, pp. 1–162, 2018.

[6] Sanja Štajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared Task on Quality Assessment for Text Simplification. In **Proceedings of Shared Task on Quality Assessment for Text Simplification**, pp. 22–31, 2016.

[7] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 117–122, 2019.

[8] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 5070–5081, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[11] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 3603–3609, 2020.

[12] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation. In **Proceedings of the Second Conference on Machine Translation**, pp. 169–214, 2017.

[13] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 Evaluation Campaign. In **Proceedings of the 14th International Conference on Spoken Language Translation**, pp. 2–14, 2017.

[14] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1351–1361, 2021.

[15] Nobuhiro Kaji and Masaru Kitsuregawa. Automatic Construction of Polarity-Tagged Corpus from HTML Documents. In **Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics**, pp. 452–459, 2006.

[16] Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A Minimalist NMT Toolkit for Novices. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations**, pp. 109–114, 2019.

[17] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.

[18] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, 2018.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, 2020.

[20] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, 2018.