6S-08

事前訓練済み系列変換モデルに基づくやさしい日本語への平易化

中町 礼文† 梶原 智之‡

†大阪大学大学院情報科学研究科 ‡大阪大学データビリティフロンティア機構

1 はじめに

日本に定住する外国人は 200 万人を超えており、そのうち日本語を理解できる人数は 62% と、英語の 44% や中国語の 38% を大きく上回る[1]。そこで、災害情報や行政情報の提供、日々のニュース発信など、様々な場面で「やさしい日本語」での情報提供が広がっている。本研究では、所与の日本語文を「やさしい日本語」へと自動変換する日本語のテキスト平易化に取り組む。

テキスト平易化は、同一言語内の翻訳問題として考えられており、難解文と平易な同義文の文対（パラレルコーパス）を用いて系列変換モデルを訓練するのが一般的である。この 10 年間、英語における Wikipedia やニュースの平易化を中心に、統計的機械翻訳[2]やニューラル機械翻訳[3]のモデルが使用されてきた。しかし、数百万から数千万文対のパラレルコーパスを使用可能な機械翻訳とは異なり、数万文対のパラレルコーパスしか使用できないテキスト平易化では充分な性能が得られていない。

本研究では、テキスト平易化における少資源問題に対処するために、Web から収集した大規模な生コーパス上で事前訓練した系列変換モデル BART[4] を用いる。やさしい日本語コーパス[5,6]を用いた実験の結果、提案手法は日本語のテキスト平易化において最高性能を達成した。

2 提案手法

本研究では、事前訓練済み系列変換モデルである BART[4]の日本語モデルを用いて、日本語のテキスト平易化を行う。日本語 BART は、自己注意機構[7]に基づく符号化器および復号器を日本語 Wikipedia の約 1,800 万文で事前訓練したものである。事前訓練には、ノイズ除去自己符号化の手法が採用されており、図 1 の上段に示すように、語句の削除などのノイズを復元する訓練を通して、言語に関する意味的および文法的な知識を獲得する。このような事前訓練を経た BART の系列変換モデルを目的タスクのデータで再訓練することによって、先行研究[4]では特に少資源設定の機械翻訳において、大幅な性能改善が確認されている。

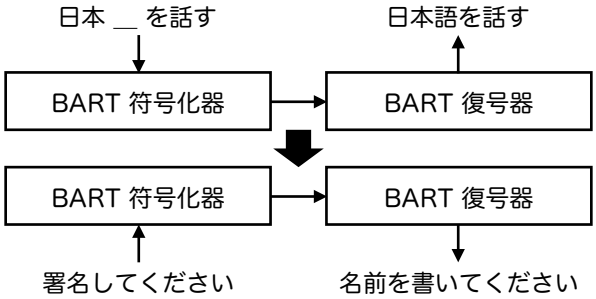

図 1：BART に基づく日本語のテキスト平易化

本研究では、日本語 BART をやさしい日本語コーパス[5,6]上で再訓練し、日本語のテキスト平易化における少資源問題に対処する。つまり、難解な日本語文とは平易な日本語文に「ノイズ」を加えたものであると考え、図 1 の下段に示すように、BART を用いて難解な日本語文に対する「ノイズ除去」を行い、平易な日本語文へ変換するテキスト平易化モデルを構築する。

3 実験設定

やさしい日本語コーパス[1][5,6]を、4 万文対の訓練用データ、1 千文対の検証用データ、100 文対の評価用データに分割して用いた。評価用には、7 種類の平易化が付与されたマルチリファレンス[6]を使用した。訓練用と検証用には、やさしい日本語コーパスのうち、難解文と平易文が完全一致する文対や、無理な平易化によって平易文の流暢性が損なわれている文対を除いた 4.1 万文対を無作為に分割して使用した。

日本語 BART には、京都大学から公開[2]されている Base モデルと Large モデルを使用した。BART に基づくテキスト平易化の有効性を検証するために、以下の 3 つの手法と比較した。

- Baseline：入力文をそのまま出力する。
- RNN：再帰型ニューラルネットワークを用いる手法。先行研究[3]に相当する。
- SAN：自己注意機構[7]を用いる手法。BART から事前訓練を除いたものに相当する。

Japanese Simplification with Pre-trained seq2seq Model
Akifumi Nakamachi† (nakamachi.akifumi@ist.osaka-u.ac.jp)
Tomoyuki Kajiwara‡ (kajiwara@ids.osaka-u.ac.jp)
†Graduate School of Information Science and Technology, Osaka University
‡Institute for Datability Science, Osaka University

[1] http://www.jnlp.org/research/Japanese_simplification
[2] http://nlp.ist.i.kyoto-u.ac.jp/?BART 日本語 Pretrained モデル

表1：やさしい日本語への平易化の実験結果
（人手評価において、**はt検定によってp＜0.01で比較手法との有意差が認められたもの）

	BLEU	SARI	文法性	同義性	平易性
Baseline	55.86	21.86	−	−	−
RNN	64.90	61.23	3.72	3.33	3.02
SAN-Base	71.11	61.55	3.88	3.44	3.30
SAN-Large	66.20	61.13	4.02	3.60	3.23
BART-Base	**82.29**	**64.68**	**4.82****	**4.62****	**3.93****
BART-Large	81.18	63.59	**4.82****	**4.62****	3.91**

各モデルの詳細を述べる。RNNには、256次元の埋込層および隠れ層を持つ2層のLSTMを用いた。SANおよびBARTでは、Baseモデルと Largeモデルの2種類を実験した。Baseモデルには、768次元の埋込層および隠れ層を持つ6層のモデルを使用し、注意ヘッドは12とした。Largeモデルには、1,024次元の埋込層および隠れ層を持つ12層のモデルを使用し、注意ヘッドは16とした。正則化には、埋込層および隠れ層にdropoutを適用し、さらにlayer-normalizationおよびlabel-smoothingを使用した。最適化にはadamを使用し、perplexityに基づく32checkpointのearly-stoppingを適用した。

比較手法はSockeye[3]で実装し、提案手法はFairseq[4]で実装した。単語分割にはSentencePiece[5]の1-gram言語モデル（語彙サイズは8,000に制限）を用いた。

各手法の性能は、自動評価と人手評価の両方で評価した。自動評価には、テキスト平易化の先行研究[3]に従い、BLEU[6]およびSARI[6]を用いた。BLEUは出力文と正解文の語句の一致率に基づく自動評価であり、SARIは入力文も用いるテキスト平易化用の改良版である。人手評価には、出力文の文法性、入出力間の同義性、出力文の平易性の3項目について、大学院生3名による5段階評価（1:最低-5:最高）を実施した。

4 実験結果

表1に実験結果を示す。自動評価と人手評価で一貫して、RNNよりもSANの性能が高く、BARTが最高性能を達成した。SAN-BaseとBART-Baseを比較すると、BLEUの自動評価で約10ポイント、文法性や同義性の人手評価で約1ポイントと、大幅な性能改善を確認できた。BaseモデルとLargeモデルの比較からは、4万文対という小規模なデータしか利用できない今回の設定では、Largeモデルの恩恵を得られないことがわかった。

出力例としては、「警察は、逃亡者を追跡している。」という文の平易化において、RNNは「警察は、逃げることを追っている。」、SANは「警察は、死んだ者を追っている。」と出力したが、BARTは「警察は逃げた人を追っている。」と適切な平易化を実現できた。この例では、「逃亡者」という単語がやさしい日本語コーパスに出現しないため、RNNでは「逃」、SANでは「亡」の意味のみを捉えて平易化に失敗したと考えられる。一方で、Wikipediaには「逃亡者」が500回以上出現するため、BARTは事前訓練の恩恵を受け、高品質な平易化を実現できた。

5 おわりに

本研究では、生コーパス上での事前訓練によって、日本語のテキスト平易化の性能を大幅に改善した。今後は、文の平易性報酬に基づく強化学習[8]を導入し、平易性を更に改善したい。

参考文献

[1] 岩田一成: 言語サービスにおける英語志向―「生活のための日本語：全国調査」結果と広島の事例から―, *社会言語科学*, Vol.13, No.1, pp.81-94 (2010).

[2] Lucia Specia: Translating from Complex to Simplified Sentences, *Lecture Notes in Computer Science*, Vol.6001, pp.30-39 (2010).

[3] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, Liviu P. Dinu: Exploring Neural Text Simplification Models, *In Proc. of ACL*, pp.85-91 (2017).

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *In Proc. of ACL*, pp.7871-7880 (2020).

[5] Takumi Maruyama, Kazuhide Yamamoto: Simplified Corpus with Core Vocabulary, *In Proc. of LREC*, pp.1153-1160 (2018).

[6] Akihiro Katsuta, Kazuhide Yamamoto: Crowdsourced Corpus of Sentence Simplification with Core Vocabulary, *In Proc. of LREC*, pp.461-466 (2018).

[7] 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).

[8] Akifumi Nakamachi, Tomoyuki Kajiwara, Yuki Arase: Text Simplification with Reinforcement Learning Using Supervised Rewards on Grammaticality, Meaning Preservation, and Simplicity, *In Proc. of AACL-SRW*, pp.153-159 (2020).

[3] https://github.com/awslabs/sockeye
[4] https://github.com/pytorch/fairseq
[5] https://github.com/google/sentencepiece
[6] https://github.com/feralvam/easse