

## ニューラル機械翻訳による公的文書平易化

## Official document simplification using neural machine translation approach

丸山 拓海 \*<sup>1</sup>      山本 和英 \*<sup>1</sup>  
Takumi Maruyama      Kazuhide Yamamoto

\*<sup>1</sup>長岡技術科学大学大学院 電気電子情報工学専攻  
Electrical, Electronics and Information Engineering, Nagaoka University of Technology

Official documents are documents distributed at public facilities such as city halls, hospitals and schools. These documents contain a lot of important information for living. However, they are difficult for non-native speakers because they contain difficult vocabulary and expressions. Therefore, official documents must be simplified. We try to simplify official document using machine translation approach. We use a parallel corpus of the original and three kinds of simplified ones including literal translation, free translation and summary. They are rewritten by 40 Japanese teachers. We adapt several methods for low-resource machine translation such as pre-trained embeddings and sharing encoder, decoder and output embeddings (tied-embeddings). The result shows that Transformer can simplify official document using pre-trained embeddings and tied-embeddings in spite of low resource. Performance improvement using several methods of low resource machine translation shows that Transformer can improve performance more than other methods by extending training data.

## 1. はじめに

近年、たくさんの外国人が日本を訪れている。その数は年間で2,800万人<sup>\*1</sup>にのぼる。また、日本に住む外国人は232万人<sup>\*2</sup>であり、増加の傾向にある。このような状況の一方、日本社会においては、日本語で情報提供されることが多い。情報提供の最も理想的な方法は、すべての外国人に対してそれぞれの母語で情報を伝えることであるが、世界には多くの言語が存在するため、現実的には不可能である。このような言語の選択という問題を解決するための情報提供の手段として、近年、「やさしい日本語」という考え方が注目されている。ここで、「やさしい日本語」とは、簡単な語彙や文構造のみを用いて、日本語に不慣れな外国人にもわかりやすくした日本語のことを指す。また、国立国語研究所の全国調査では、日本に住む外国人のうち、英語を理解できる人の割合よりも、簡単な日本語を理解できる人の割合の方が多くと報告されており[岩田 10]、英語よりも簡単な日本語の方が外国人には伝わりやすいことが知られている。

本研究では、公的文書を対象とする。公的文書とは、市役所や病院、学校等の公共施設で配布される文書であり、これらの文書は生活する上で重要な情報を多く含んでいる。しかし、日本語初学者が学習する文に比べ、難解な語彙や公的文書に出現する固有の表現も含み、理解が困難であるため、平易化が必要な文書である。我々は、日本語教育者が公的文書の日本語を逐語訳、意識、要約の3段階の「やさしい日本語」に書き換えたものをコーパスとし、機械翻訳的なアプローチにより、文単位の平易化を試みる。

## 2. 関連研究

公的文書平易化の試みとして、空らの研究がある。彼らは次のような手順に従って、公的文書の平易化を行った[空 13]。

1. 重要部分の抽出
2. 短文化
3. 表現意図を用いた図示への変換
4. 「やさしい日本語」への変換

ステップ1からステップ3までの処理によって、文の構造を読みやすくしたのちに、彼らが構築した換言辞書を用いて、語彙的な平易化を行っている。平易化システムの出力を「日本語としての正しさ」と「やさしさ」という観点で人手により評価した結果、公的文書をやさしく書き換えることには成功したが、抽出した換言ルールを直接適応させているため、文法的な誤りや意味を保持できない場合が多いことが報告されている。また、直接的表現へ言い換えを行った研究[Moku 12]では、変換ルールの規模が小さく、平易化としての効果があり得られなかったことが報告されている。

一方で、松田らは、統計的機械翻訳を用いて公的文書を「やさしい日本語」へ変換する研究を行っている[松田 09]。BLEUによる精度評価では、20ポイントと低い値に留まった結果となった。また、原文と原文から語彙的なレベルで平易化を行った逐語訳、そして機械翻訳による翻訳結果の3つを人手で、1:比較的良質、2:解読不能・翻訳誤り、3:変化なし、の3段階に評価したところ、比較的良質な変換とされた翻訳結果は7.5%と少なく、そのうちのほとんどが見出しのような短い句であったことが報告されている。この理由として、対訳コーパスが少ないことや、公的文書特有の記号や括弧表記がノイズになっていることを指摘している。また、適切にドメインを設定することによって、性能改善ができる可能性についても述べられている。

機械翻訳の分野では、統計的機械翻訳の性能をはるかに上回るニューラル機械翻訳が提案されている[Bahdanau 14, Luong 15, Gehring 17, Vaswani 17]。近年、テキスト平易化においても、ニューラル機械翻訳を用いるアプローチが盛んに研究されており[Maruyama 17, Nisioi 17, Zhang 17, Surya 18, Zhao 18]、既存の統計的機械翻訳をベースとしたモデルより

連絡先: 丸山拓海, 長岡技術科学大学大学院電気電子情報工学専攻, 新潟県長岡市上富岡町 1603-1, maruyama@jnlp.org

\*1 [https://www.jnto.go.jp/jpn/statistics/visitor\\_trends](https://www.jnto.go.jp/jpn/statistics/visitor_trends)

\*2 <https://www.e-stat.go.jp>

も、高い平易化性能を記録している。ニューラル翻訳をベースとしたモデルでは、学習に大量のデータが必要となるが、テキスト平易化のための大規模な対訳コーパスを用意することは容易ではない。

本研究では、ニューラル機械翻訳のモデルが公的文書の平易化という言語資源が少ないタスクにおいて、どの程度効果を発揮できるのかを検証する。

3. データセット

本研究では、空らが利用した平易化コーパス [空 13] と同じもの (以下、「公的文書書き換えコーパス」と呼ぶ) を利用する。これは「やさしい日本語」のプロジェクトで作成されたものであり、約 40 名の日本語教師が、市役所や病院、学校等の公共施設で配布される公的文書を「やさしい日本語」に書き換えたものである。このコーパスは原文である公的文書 1,101 文書と共にその逐語訳、意識、要約という 3 段階の翻訳を含む対訳コーパスである。それぞれの翻訳の位置付けは下記の通りである。

- 逐語訳: 日本語文の難解な語彙や句をやさしい表現に、書き換えたもの。
- 意識: 文意等を損なわないように可能な限り、やさしい表現に書き換えたもの。
- 要約: 可能な限り文を平易化したもの。

これらは一定の文法基準 [庵 08] と旧日本語能力試験 2 級 (現試験における N2) レベルの語彙のみに制限されている。コーパスにおける「やさしい」の基準は日本語教師の主観である。各翻訳の例を以下に示す。

- 原文: ニュース等で報道されておりますように、世界的に新型 (豚) インフルエンザの流行が危惧されています。
- 逐語訳: ニュースなどにもあるように、世界中で新型インフルエンザの流行が心配されています。
- 意識: さて、ニュースでもありますが、世界中で新型インフルエンザが増えています。
- 要約: さて、世界中で新型インフルエンザが増えています。

上記の例では、原文に対して 3 種類の翻訳文が存在するが、場合によっては、翻訳するのではなく原文全体を削除する事例も存在する。そのため、原文と逐語訳・意識・要約のそれぞれの文数は必ずしも一致しない。本コーパスの文数、平均文長、語彙数を表 1 に示す。

| 表 1: 公的文書書き換えコーパスの統計量 |        |        |        |        |
|-----------------------|--------|--------|--------|--------|
|                       | 原文     | 逐語訳    | 意識     | 要約     |
| 文数                    | 35,861 | 35,809 | 32,841 | 27,588 |
| 平均文長                  | 25.32  | 28.55  | 26.20  | 24.74  |
| 語彙数                   | 14,848 | 11,337 | 10,243 | 9,259  |

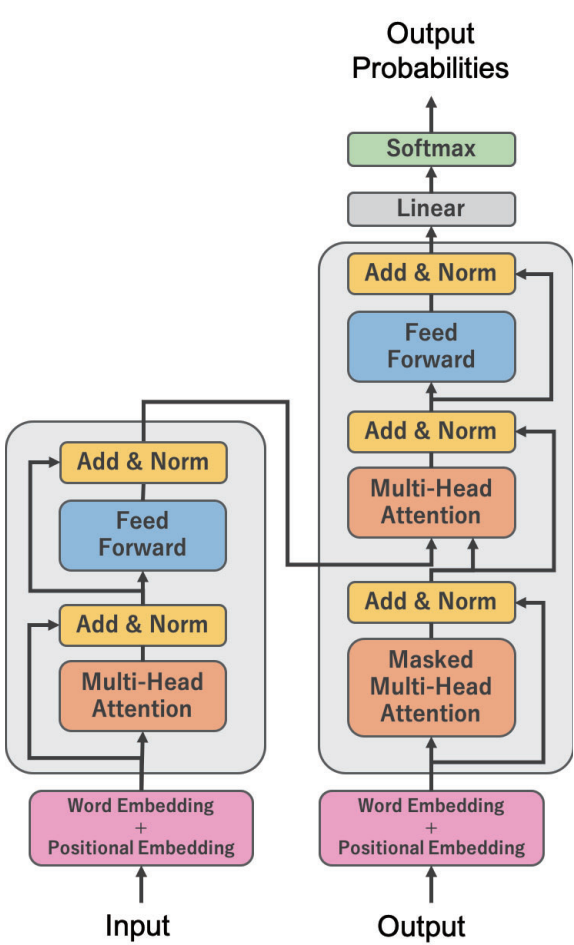


図 1: Transformer

4. 実験方法

今回は、機械翻訳モデルを利用し、原文から逐語訳・意識への文単位の変換を試みる。我々は公的文書書き換えコーパスを、学習データ 34,000 文、開発データ 1,000 文、評価データ 861 文のように分割した。これらのデータを用いて、翻訳モデル (4.1 節) をトレーニングし、モデルの出力結果を BLEU, SARI(4.2 節) により評価する。

4.1 Transformer

本研究では、翻訳モデルとして、Vaswani らが提案している Transformer を用いる [Vaswani 17]。モデルの概略図を図 1 に示す。Transformer の encoder (図 1 左側) は、multi-head self-attention と Feed Forward Neural Network で構成される層を L 段積み上げた構成となっている。multi-head self-attention は、次式に従って、前段の隠れ状態  $e_{(s',l)}$  から隠れ状態  $e_{(s,l)}$  を計算する。

$$e_{(s,l)} = \sum_{s'} \alpha_{(s',l)}^{enc} e_{(s',l-1)} \tag{1}$$

$$\alpha_{(s',l)}^{enc} = a(e_{(s',l)}, e_{(s',l-1)}, H) \tag{2}$$

ここで、 $\alpha_{(s',l)}^{enc}$  は、l 層目、s' ステップの attention distribution を意味する。a(・) を multi-head self-attention 表す関数、H はヘッド数を意味する。一方、decoder (図 1 右側) は、

encoder と同様な機構に加え、encoder の出力に対する attention 機構が存在する。1 段目の multi-head attention では、encoder と同様に次式に従って隠れ状態  $d_{(s,l)}$  を計算する。

$$d_{(s,l)} = \sum_{s'} \alpha_{(s',l)}^{dec} d_{(s',l-1)} \quad (3)$$

$$\alpha_{(s',l)}^{enc} = a(d_{(s',l)}, d_{(s',l-1)}, H) \quad (4)$$

2 段目の multi-head attention では、encoder の出力と前段の multi-head attention による隠れ状態から、次式により、文脈ベクトル  $c_{(s,l)}$  を計算する。

$$c_{(s,l)} = \sum_{s'} \alpha_{(s',l)}^{dec2} e_{(s',l-1)} \quad (5)$$

$$\alpha_{(s',l)}^{enc} = a(d_{(s',l)}, e_{(s',l)}, H) \quad (6)$$

このモデルは、対数尤度  $P = \log P(O|I, \theta)$  を最大化するようにトレーニングされる。ここで、 $O$  は平易文、 $I$  は原文、 $\theta$  はモデルのパラメータを意味する。

エンコーダとデコーダの単語埋め込み層では、学習済みの単語ベクトル nwjc2vec[浅原 17](pre-emb.) を用いる。また、エンコーダの単語埋め込みとデコーダの入出力の単語埋め込み層を共有する (tied-emb.)。モデルのハイパーパラメータは、Zhao らの研究に倣い、エンコーダとデコーダを 4 層、それぞれのアテンション機構のヘッド数を 5、ドロップアウトを 0.3 に設定している。また、各単語埋め込み層の次元は、nwjc2vec に合わせ、200 次元としている。

## 4.2 評価方法

テキスト平易化では、一般的に流暢性、意味保持性、平易さの 3 つの観点からモデルの出力結果を評価する。ここで、流暢性とは、日本語として正しい文を出力できているかを測る評価尺度であり、意味保持性とは、原文 (モデルに対する入力文) とモデルが出力した文の意味が一致するかどうかを測る評価尺度である。また、平易さとは、出力文が入力文に比べ、どの程度簡単になっているかを示す評価尺度である。これらを自動で評価する方法として、BLEU と SARI[Xu 16] が用いられる。

BLEU とは、機械翻訳の評価で広く用いられる評価尺度であり、テキスト平易化においては、流暢性と意味保持性に関して正の相関があることが知られている [Vu 18]。

SARI は、最近提案された平易化の評価指標であり、平易化で行われる単語の追加、削除、保持の 3 つの操作における n-gram の precision 及び recall である。この尺度は、平易化すべき部分を適切に平易化した際に、高いスコアを与える仕組みとなっている。具体的には、モデルが参照文にのみ存在する単語を追加した場合や参照文に存在する単語をそのまま出力 (保持) した場合、参照文にない単語を削除した場合などに高いスコアを与える。一方、入力文をそのまま出力するようなモデルに対してはペナルティが与えられるように構成されている。実験的評価より、SARI が人間の平易さの判断とよく相関があることを示している [Xu 16, Vu 18]。

## 5. 結果と考察

原文から逐語訳への変換及び原文から意識への変換の結果を表 2 に示す。Origin は、原文をそのまま出力した際の結果である。NTS は、Nisioi らのニューラル機械翻訳によるテキス

ト平易化モデルである。ハイパーパラメータは、先行研究と同様に設定している。ただし、単語埋め込み層は、nwjc2vec の次元数に合わせて、200 次元に設定している。

表 2: 実験結果

|             | 逐語訳          |              | 意識           |              |
|-------------|--------------|--------------|--------------|--------------|
|             | BLEU         | SARI         | BLEU         | SARI         |
| Original    | 34.24        | 16.65        | 28.14        | 14.17        |
| NTS         | 36.49        | 46.47        | 29.46        | <b>41.85</b> |
| + pre-emb.  | 34.60        | 45.17        | 28.87        | 41.31        |
| + tied-emb. | 34.85        | 44.94        | <b>29.69</b> | 41.38        |
| Transformer | 33.88        | 45.94        | 18.61        | 35.56        |
| + pre-emb.  | 34.52        | 46.01        | 24.65        | 39.90        |
| + tied-emb. | <b>41.89</b> | <b>48.43</b> | 29.41        | 41.63        |

実験結果より、NTS, Transformer どちらも、Original を上回る BLEU, SARI を記録しており、流暢性を保ちつつ、平易化できていることが分かる。NTS の結果に注目すると、逐語訳への変換、意識への変換どちらの場合においても、学習済みの単語ベクトルの使用 (NTS + pre-emb.) やエンコーダの単語埋め込みとデコーダの入出力の単語埋め込み層の共有 (NTS + tied-emb.) を行なっても、BLEU や SARI の向上にあまり効果がないことが分かる。一方で、Transformer の結果に着目すると、それらの工夫が大きく効果を示していることが分かる。特に、単語埋め込み層の共有 (Transformer + tied-emb.) の効果は大きく、逐語訳においては、そのままの Transformer に比べ、BLEU を 8.0 ポイント、SARI を 2.5 ポイント向上させている。意識においては、BLEU を 11 ポイント、SARI を 6.1 ポイントを改善させている。これらの結果は Transformer において、学習データを拡張させることでさらなる性能改善が可能であることを示している。

## 6. まとめ

我々は公的文書を対象に平易化を試みた。公的文書とは、市役所や病院、学校等の公共施設で配布される文書であり、生活する上で重要な情報を多く含んでいる。しかし、日本語初学者が学習する文に比べ、難解な語彙や公的文書に出現する固有の表現も含み、理解が困難であるため、平易化が必要な文書である。

本研究では、約 40 名の日本語教師が公的文書の日本語を逐語訳、意識、要約の 3 段階の「やさしい日本語」に書き換えたものをコーパスとし、機械翻訳的なアプローチにより、文単位の平易化を行った。また、学習済みの単語ベクトルの使用やエンコーダの単語埋め込みとデコーダの入出力の単語埋め込み層の共有を行い、モデル側で学習データの不足を補うことを試みた。

結果として、Transformer に学習済みの単語ベクトルやエンコーダの単語埋め込みとデコーダの入出力の単語埋め込み層の共有を利用することにより、小規模な学習データであっても適切に平易化できることを示した。学習済みベクトルや単語埋め込み層の共有による性能改善は、Transformer において、学習データの拡張によってさらなる改善が可能であることを示している。今後は、学習データの擬似的な拡張や転移学習といった方法を検討していきたい。



## 参考文献

- [Bahdanau 14] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473 (2014).
- [Gehring 17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin: Convolutional Sequence to Sequence Learning, arXiv:1705.03122 (2017).
- [Luong 15] Minh-Thang Luong, Hieu Pham, Christopher D. Manning: Effective Approaches to Attention-based Neural Machine Translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1412-1421 (2015).
- [Maruyama 17] Takumi Maruyama, Kazuhide Yamamoto: Sentence Simplification with Core Vocabulary, Proceedings of the International Conference on Asian Language Processing, pp.363-366 (2017).
- [Moku 12] Manami Moku, Kazuhide Yamamoto, Ai Makabi: Automatic Easy Japanese Translation for information accessibility of foreigners, Proceedings of the Workshop on Speech and Language Processing Tools in Education, pp.85-90 (2012).
- [Nisioi 17] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, Liviu P. Dinu: Exploring Neural Text Simplification Models, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.85-91 (2017).
- [Surya 18] Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, Karthik Sankaranarayanan: Unsupervised Neural Text Simplification, arXiv:1810.07931 (2018).
- [Vaswani 17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), pp.5998-6008 (2017).
- [Vu 18] Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, Hong Yu: Sentence Simplification with Memory-Augmented Neural Networks, Proceedings of NAACL-HLT 2018, pp.79-85 (2018).
- [Xu 16] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch: Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics, Vol.4, pp.401-415 (2016).
- [Zhang 17] Xingxing Zhang, Mirella Lapata: Sentence simplification with deep reinforcement learning, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.584-594, (2017).
- [Zhao 18] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, Bambang Parmanto: Integrating Transformer and Paraphrase Rules for Sentence Simplification, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.3164-3173 (2018).
- [浅原 17] 浅原 正幸, 岡 照晃:nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ, 言語処理学会第23回年次大会 発表論文集, (2017).
- [庵 08] 庵功雄: 「やさしい日本語」をめぐる. 多文化共生社会における日本語教育研究会, 第4回研究会, pp.1-12 (2008).
- [岩田 10] 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」結果と広島の実例から (特集:日本社会の変容と言語問題), 社会言語科学会, Vol.13, No.1, pp.81-94 (2010).
- [松田 09] 松田真希子: やさしい日本語への自動言い換えシステムの開発, 日本語教育学会大会 2009(平成 21) 年度春季大会予稿集, pp.91-93 (2010).
- [奎 11] 奎 真奈見, 山本 和英: 公的文書に対する「やさしい日本語」換言辞書作成のための調査, 言語処理学会 第17回年次大会 発表論文集, pp.376-379 (2011).
- [奎 13] 奎 真奈見, 山本 和英: 「やさしい日本語」変換システムの試作, 言語処理学会 第19回年次大会 発表論文集, pp.678-681 (2013).