事前学習モデルと固有名詞の事後置き換えを用いた日本語から手話への機械翻訳

宮崎太郎　中谷真規　内田翼　金子浩之　佐野雅規
NHK 放送技術研究所
{miyazaki.t-jw, nakatani.n-ge, uchida.t-fi, kaneko.h-dk, sano.m-fo}@nhk.or.jp

概要

手話は先天的あるいは幼少期に聴覚を失った人にとって第一言語であり，それらの人に対する情報提供には手話を用いるのが望ましい．我々は手話での情報提供をより拡充するために，手話 CG の自動生成システムの開発を進めている．本稿では，手話の自動生成に用いる日本語-手話の機械翻訳手法について述べる．

手話は対訳コーパスの作成が難しいため，既存の比較的小規模のコーパスからより良い翻訳精度を得るために，提案手法では翻訳モデルの Encoder に事前学習モデルを用いる．また，手話では固有名詞の表現方法が複雑なため，辞書や固有名詞に特化した翻訳手法を用いて事後的に固有名詞部分の置き換えをするのが望ましい．そこで，翻訳モデルを用いて翻訳した後に，得られた翻訳結果と入力の日本語文から，日本語の固有表現が手話のどの部分に翻訳されたかを推定し，固有表現を後から置き換える手法を提案する．

1　はじめに

先天的あるいは幼少期に聴覚を失った人にとって，手話は第一言語であり，日本語などの音声言語やその書き起こしよりも理解がしやすい重要なコミュニケーション手段である．そのような人たちに対して重要な情報を提供する際には，手話を用いることが望ましい．総務省が週に 15 分以上の手話放送の実施を普及目標に定めるなど，放送分野においても手話による情報提供の拡充が求められている [1]．

手話番組を実施するためには番組の音声を手話に翻訳する手話通訳者の確保などが必要であるが，放送をリアルタイムに翻訳できる手話通訳者の数は限られており，現状では手話番組の大幅な拡充は難しい．そこで，NHK では日本語の文章を自動で手話に翻訳する手話 CG アニメーション（以下，手話 CG）の自動生成技術の研究を進めている．これまでに気象情報やスポーツ情報などを対象として，事前に翻訳した定型的な手話表現をテンプレート化し，データと組み合わせて手話 CG を自動生成する技術を開発してきた [2, 3]．これらのサービスは定型的なデータを基に正確な手話 CG を生成できる反面，変換できる情報が限定されており，日々のニュースなどの情報を翻訳することはできなかった．我々はより幅広い情報を手話 CG に変換するために，機械翻訳を用いた手話 CG 生成技術の研究を進めている．

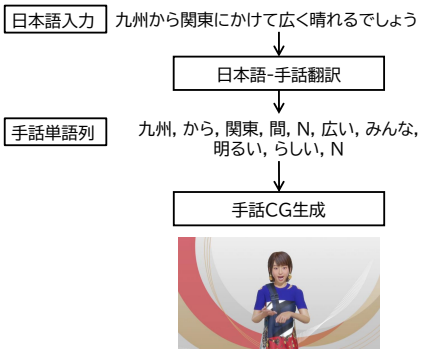

図 1　手話 CG 生成システム概要図

機械翻訳を用いた手話 CG 生成システムの概要図を図 1 に示す．入力の日本語は機械翻訳技術により手話単語列に翻訳される．手話の表現に必要な各単語の動作データはあらかじめモーションキャプチャにより作成してあり，翻訳結果にあわせて動作データをつなぎ合わせることで手話 CG が生成できる．

ここで用いる日本語から手話への機械翻訳の課題として，まず学習データの少なさ，ドメインの偏りが挙げられる．手話はそもそも書き言葉がないため，機械翻訳の学習データとして使えるような対訳コーパスが少なく，手話による情報提供がなされている分野が限られているため，ドメインには偏りが生じる．さらに，手話書き起こしの表記方法が用途

や制作者によって異なるため，他者が作成したコーパスをそのまま使うことが難しい．その他の課題として，固有表現の翻訳が難しいことが挙げられる．一般の書き言葉であれば固有表現はその発音を基に音訳したものを用いることができる．一方，手話では例えば「佐野」という人名は「サノ」という発音を表す指文字を使用するなど音訳で対応できるものもあるが，「宮崎」という人名はそれぞれの文字の意味を基にして翻訳した「宮（神社）」と「先」の 2 単語で表すなど，さまざまな表現方法を組み合わせて表現される．そのため，固有名詞の翻訳には固有名詞に特化した翻訳手法が必要であり，それを用いるためには，文全体の翻訳結果の固有名詞部分を置き換えることが必要となる．

本稿では，コーパスの少なさを補うための事前学習モデルを用いた機械翻訳手法に，固有名詞をより正確に表現するための事後置き換えを組み合わせた翻訳手法を提案する．

2 関連研究

2.1 プレースホルダーを用いた置き換え

Luong et al.[4] は翻訳が難しい未知語や低頻度語をプレースホルダーに置き換えて翻訳することで，これらの翻訳誤りを防ぐ手法を提案した．プレースホルダーを用いた翻訳は多くの研究で取り入れられており，良好な翻訳性能が報告されている [5, 6]

これらの手法では，我々がターゲットとしているニュースなどに頻出し，かつ重要な語句である固有表現の翻訳誤りを軽減可能であるが，プレースホルダーに置き換えたコーパスから翻訳モデルを学習する必要がある．

2.2 事前学習モデルを用いた翻訳

BERT[7] や RoBERTa[8] などの事前学習モデルを用いた機械翻訳手法も多く報告されている．

Imamura et al.[9] は BERT を翻訳モデルの Encoder として利用することで，特に学習データが少ない言語対の翻訳時に翻訳性能が向上することを示した [9]．本稿では基本となる翻訳モデルにこの Imamura et al. が提案したモデルとほぼ同一のものを用いる．

Clinchant et al.[10] は事前学習モデルを Encoder として用いることで，単純に BLEU スコアなどの翻訳性能が向上するだけでなく，学習データと異なるドメインの文章の翻訳性能を向上する効果があることを示した．また，Zhu et al.[11] は学習データのみから学習した Encoder と事前学習モデルを Attention Mechanism を用いて組み合わせることにより翻訳性能の向上を実現した．

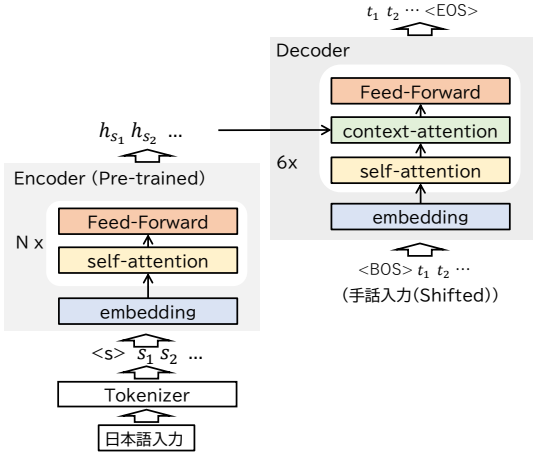

図 2 翻訳モデルの概要

事前学習モデルは機械翻訳の性能向上に大きな効果が期待できるが，一方で扱うことができる語彙が事前学習モデルの語彙に限定されるため，プレースホルダーに置き換えた学習が難しい．

本稿では，事前学習モデルを用いて学習データの少なさを補いつつ，プレースホルダーを用いた固有表現の置き換えと同様の後処理を可能とすることで，翻訳性能の向上とともに，ニュースで現れる新語や専門用語への対応を可能とすることを目指す．

3 翻訳手法

提案手法は，ベースとなる翻訳モデルによる機械翻訳と，翻訳結果の事後置き換えの 2 つのブロックからなる．以下でそれぞれについて説明する．

3.1 ベースとなる翻訳モデル

Transformer を用いた Encoder-Decoder モデル [12] をベースとし，Encoder 部に事前学習モデルを用いたものを利用する．図 2 に翻訳モデルの構成を示す．入力日本語 $S = \{s_1, s_2, \ldots s_{n_s}\}$ を事前学習モデルからなる Encoder に入力し，各単語に対応するベクトル表現 $H = \{h_{s_1}, h_{s_2}, \ldots h_{s_{n_s}}\}$ を得る．このとき，n_s は S の Token 数を表す．Decoder には，最初は文頭を表す記号（図中の BOS）と，Encoder からの H を入力し，一単語ずつ次の単語を予想するようにして翻訳終了を表す記号（図中の EOS）が出力されるまで翻訳する．

3.2 翻訳結果の事後置き換え

3.1 節で述べた翻訳モデルを用いて，対象とする入力言語側の固有表現が翻訳結果のどの単語に翻訳されたかのスコアを計算する．翻訳結果から，このスコアが最大となる部分を辞書などにより置き換えることで，翻訳結果の事後置き換えを実現する．

入力文 S 中の n から n' 番目の単語からなる $S_n^{n'}$ が置き換えたいフレーズであるとする．この $S_n^{n'}$ を Encoder に入力した場合に，翻訳結果 T 中の m から m' 番目までの単語からなるフレーズ $T_m^{m'}$ に翻訳される尤度 $p(T_m^{m'}|S_n^{n'})$ を翻訳モデルから求める．具体的には，$T_m^{m'}$ 中の単語 $t_{\tilde{m}}$ が出力される尤度 $p(t_{\tilde{m}}|T_m^{\tilde{m}-1}, S_n^{n'})$ は，Encoder に $S_n^{n'}$ を，Decoder に BOS と $T_m^{\tilde{m}-1}$ を入力したときに単語 $t_{\tilde{m}}$ が出力される尤度として計算ができる．$T_m^{m'}$ の各単語と翻訳終了をあらわす EOS が出力される尤度を算出し，相乗平均を取ったものを部分翻訳尤度 $p(T_m^{m'}|S_n^{n'})$ とする．

$$p(T_m^{m'}|S_n^{n'}) = \left((\prod_{k=m}^{m'} p(t_k|T_m^{k-1}, S_n^{n'}) \times p(\mathrm{EOS}|T_m^{m'}, S_n^{n'})\right)^{\frac{1}{m'-m+2}} \quad (1)$$

この部分翻訳尤度を用いて，翻訳元の文中の $S_n^{n'}$ が翻訳結果中の $T_m^{m'}$ に翻訳された確度を表す $\mathrm{score}(T_m^{m'}|S_n^{n'})$ を以下のようにする．

$$\mathrm{score}(T_m^{m'}|S_n^{n'}) = p(T_m^{m'}|S_n^{n'}) - \bar{p}(T_m^{m'}|S_n^{n'}) \quad (2)$$

ここで，$\bar{p}(T_m^{m'}|S_n^{n'})$ は，入力文 S から $S_n^{n'}$ を除去したものを Encoder に入力した場合に，翻訳結果 T の $T_m^{m'}$ を除去した部分が出力される部分翻訳尤度を表し，これも翻訳モデルから算出する．

入力 $S_n^{n'}$ に対して $\mathrm{score}(T_m^{m'}|S_n^{n'})$ が最大となる m, m' について，$S_n^{n'}$ が $T_m^{m'}$ に翻訳されたものもあるとみなし，翻訳結果 T 中の $T_m^{m'}$ を辞書により置き換える．

4 実験条件

4.1 手話ニュースコーパス

評価実験には，NHK で作成している手話ニュースコーパスを用いた．手話ニュースコーパスは NHK で放送している「手話ニュース」を 2009 年 4 月から収集したもので，番組の日本語音声の書き起こしと，手話映像，手話映像の書き起こしを約 160,000 文対が含まれている．手話映像の書き起こしは，手話ニュース中の手話映像を基に，ろう者や CoDA [1] により単語ごとに書き起こしている．書き起こしの例を表 1 に示す．表中の手話書き起こしは，カンマ区切りで手話単語列を表記している．N はうなずきを，ｐｔ３は指差しをそれぞれ表し，［頭を下げる］は NMMs[2] を表す．また，「落ち着く１」などの単語末尾に付与された数字は，同じ意味を持つ複数の手話単語の分別用に付したもので，「新日本語-手話辞典」[13] に準拠している．

表 1 手話ニュースコーパスの例．

日本語	気象庁は周囲の状況を確認し，できるかぎり安全を確保するよう呼びかけています。
手話	N, 天気, 庁, ｐｔ３, 宣伝, N, みんな, 状態, 調べる, N, できる, だけ, すべて, 落ち着く１, 守る１, 頼む１, ［頭を下げる］, 宣伝, ｐｔ３, N

今回はこの中から固有表現（地名，組織名，人名）が含まれる 69,776 文を実験に用い，学習データが 66,776 文，開発データが 2,000 文，評価データが 1,000 文となるようにランダムに分割した．事前に文中に出現する固有表現を人手により抽出し，日本語-手話の間での固有表現の表現の対応付けをした．

4.2 実験設定

翻訳モデルなどの実装には Pytorch[14] と Transformers[15] を用いた．提案手法の Encoder に用いる事前学習モデルには rinna 株式会社が公開している japanese-roberta-base[16] を，Decoder には 6 層の Transformer Decoder を用いた．モデルの最適化には RAdam[17] を用い，Decoder 部の学習率を 1.0×10^{-3}，事前学習モデルを用いた Encoder 部の学習率を 2.0×10^{-5} とした．学習エポック数を最大 20 として開発データの BLEU 値が最大となるモデルを評価に用いた．学習はそれぞれのモデルについてランダムシードを変えて 3 回行い，その中で最高の性能となるモデルを用いて評価した．デコード時には Beam 幅 10 の Beam Search を用いた．

手話の語彙は，学習データ中に 3 回以上出現する全単語からなる 8,767 とした．これに含まれない単語は，学習時点で OOV に変換して学習した．

手話ではうなずきや指差しなど，機能語として用いられ固有名詞には出現しない単語がある．今回の実験では，固有名詞の事後置き換えの際に，これら

1） Children of Deaf Adults: 聴覚障害者を親に持つ子どもで，聴者でかつ第一言語が手話の人のことを指す．
2） Non-Manual Markers: 手指以外を用いて意味を表現する動作．

の機能語が先頭や末尾につくフレーズは日本語の固有名詞と対応付けないルール処理を加えた．

手話では固有表現の表現揺れが大きい．今回は表現揺れによる影響を排除するために，固有表現の置き換え辞書には，事前に対象となる評価用の文対ごとに人手で日本語-手話の間での固有表現の対応付けをしたものを用いた．

4.3 ベースライン手法

事前学習モデルを用いない No Pre-train，プレースホルダーを用いる Placeholders の 2 種類の翻訳モデルをベースラインとして性能を比較する．なお，提案手法の事前学習を用いる翻訳モデルは以下では With Pre-train と表記する．No Pre-train と With Pre-train では 3.2 節で述べた固有名詞の事後置き換えをした場合，しない場合の 2 パターンについてそれぞれ性能を比較する．

No Pre-train 事前学習モデルを使わず，学習データのみから翻訳モデルを学習する．翻訳モデルには 6 層の Transformer を用いた．学習率は最適な結果が得られた 5.0×10^{-4} とした．

Placeholders No Pre-train と同様のモデルを用い，学習時に固有表現を人名，地名，組織名それぞれ異なるプレースホルダーに置き換えて学習する．デコード時には固有表現をプレースホルダーに置き換えて翻訳した上で，翻訳結果のプレースホルダーを事後処理で置き換える．このモデルのみ，最適化には予備実験で最適な性能を得た AdamW[19] を用い，学習率を 3.0×10^{-4} とした．

5 実験結果

5.1 翻訳性能の評価

表 2 に実験結果を示す．提案手法 (With Pre-train, 固有表現置き換え後) が全体で最もよい性能を示した．これは，事前学習モデルを用いた翻訳性能の向上と固有名詞の事後置き換えの両方の利点が活用できたためと考えられる．

No Pre-train と with Pre-train を比較すると，事前学習モデルの使用と固有表現の置き換えの使用によりそれぞれ翻訳性能が向上したことがわかる．また，Placeholders とその他の手法の比較により，特別にプレースホルダーを用意しなくても，翻訳モデルから得られる翻訳尤度を用いた置き換えで十分に固有表現の置き換えが可能であることが示された．

表 2 実験結果 (BLEU)

翻訳モデル	固有表現 置き換えなし	固有表現 置き換えあり
With Pre-train	24.5	**28.8**
No Pre-train	22.9	27.2
Placeholders	—	26.7

5.2 固有表現の日本語-手話対応付け

3.2 節で述べた固有表現の対応付けの性能評価を行った．提案手法で得られた対応付けが正解データと完全一致した場合を正解とした場合の正解率は 71.7% (正解: 1,029, 誤り: 407) であった．誤りで最も多かったのが，手話で固有名詞の直後につける固有表現の種類を表す単語を出力に含める誤りであった．例えば日本語の「宮崎」は，手話では「宮（神社)」「先」という 2 つの手話単語で表現するが，このあとに人名か地名かを区別するために，地名であれば「場所」，人名であれば「彼」「彼女」という単語を後ろにつけて表現する場合がある．この，「場所」や「彼」「彼女」を固有名詞に含めて出力してしまう誤りである．この種類の誤りが 88 回発生した．ついで多かったものが，「周辺」や一帯をあらわす「みんな」など，地名と組み合わせて範囲を表現する単語を固有表現の一部として出力する場合であった．これは 59 回発生した．これらはパターン化可能なため，ルール等で対応が可能である．

これ以外の誤りも含め，誤りの多くは 1 単語の過不足によるもので，翻訳結果に大きく影響するほどの誤りは少なかった．

6 おわりに

本稿では日本語から手話単語列に翻訳する機械翻訳手法について述べた．翻訳モデルの Encoder に事前学習モデルを用い，また，プレースホルダーを使わずに事後的に固有表現を置き換える翻訳手法を提案した．提案手法はベースライン手法と比較して良好な翻訳性能が得られることを確認した．

今後の課題として，今回の実験では固有表現に対する手話の表現揺れの影響を排除するために，評価データの文対ごとにそのデータから作成した誤りのない辞書を使用したが，これを一般に使える辞書に変更した場合の性能を確認する必要がある．また，翻訳性能にも直結する翻訳元と翻訳結果の間での固有名詞の対応付け部分について，専用のモデルを用意するなど，性能向上を図りたい．

参考文献

[1] 総務省. 放送分野における情報アクセシビリティに関する指針, 2018.

[2] Tsubasa Uchida, Taro Miyazaki, Makiko Azuma, Shuichi Umeda, Naoto Kato, Hideki Sumiyoshi, Yuko Yamanouchi, and Nobuyuki Hiruma. Sign language support system for viewing sports programs. In **Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility**, pp. 339–340, 2017.

[3] Makiko Azuma, Nobuyuki Hiruma, Hideki Sumiyoshi, Tsubasa Uchida, Taro Miyazaki, Shuichi Umeda, Naoto Kato, and Yuko Yamanouchi. Development and evaluation of system for automatically generating sign-language cg animation using meteorological information. In **International Conference on Computers Helping People with Special Needs**, pp. 233–238. Springer, 2018.

[4] Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 11–19, Beijing, China, July 2015. Association for Computational Linguistics.

[5] Soichiro Murakami, Makoto Morishita, Tsutomu Hirao, and Masaaki Nagata. NTT's machine translation systems for WMT19 robustness task. In **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 544–551, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1639–1657, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[9] Kenji Imamura and Eiichiro Sumita. Recycling a pretrained BERT encoder for neural machine translation. In **Proceedings of the 3rd Workshop on Neural Generation and Translation**, pp. 23–31, Hong Kong, November 2019. Association for Computational Linguistics.

[10] Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. On the use of BERT for neural machine translation. In **Proceedings of the 3rd Workshop on Neural Generation and Translation**, pp. 108–117, Hong Kong, November 2019. Association for Computational Linguistics.

[11] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. Incorporating bert into neural machine translation. In **International Conference on Learning Representations**, 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[13] 日本手話研究所（編）米川明彦（監修）. 新日本語-手話辞典. 全日本聾唖連盟出版局, 2006.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[16] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 169–170, 2021.

[17] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In **International Conference on Learning Representations**, 2019.

[18] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **arXiv preprint arXiv:1808.06226**, 2018.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.