

漫画の機械翻訳における吹き出しよりも長い文脈情報の活用 Using Long Context for Manga Machine Translation

戒能 大翔[†] 梶原 智之[†] 二宮 崇[†] ジョシュア タナー[‡] 石渡 祥之佑[‡]
Hiroto Kaino Tomoyuki Kajiwara Takashi Ninomiya Joshua Tanner Shonosuke Ishiwatari

1. はじめに

近年、非公式に翻訳された海賊版の漫画が海外で流通しており、正規版の日本の漫画を海外展開する上での大きな課題となっている。2021 年の小学館による調査¹では、正規版の約 5 倍の規模の漫画作品が海賊版として海外で流通しており、海賊版の閲覧回数が 1 億回を越えるような作品も存在する。この要因のひとつとして、リアルタイムな正規の翻訳出版が少ないことが考えられる。そのため、高速な翻訳出版を増やしていくために、自然言語処理の技術を用いた漫画の機械翻訳 [1] に大きな期待が寄せられている。

先行研究 [1] でも言及されているように、漫画の機械翻訳というタスクには、処理の単位である吹き出しがそれ単体では翻訳のために十分な情報を持たないという難しさがある。例えば、図 1 の左下のコマでは“「あさがお」はお父さんの夢だからな”という 1 つのセリフが“「あさがお」は”と“お父さんの夢だからな”の 2 つの吹き出しに分割されている。そのため、図 1 の最後の吹き出しはこれ単体では適切に翻訳することが困難であり、文脈情報を参照しつつ翻訳する必要がある。文脈情報として、図 1 の例では直前の吹き出しに含まれるテキストを扱うが、図 2 の例のように前のコマの情報を参照したり、作品や作者の特徴（例えば、図 1 の「あさがお」は花の種類ではなく、ロケットの名前である）を反映したりと、より長いレベルの文脈情報が有用な場合もある。このような文脈情報の活用は、通常の機械翻訳タスクにおいても盛んに研究されている [2] が、漫画の機械翻訳タスクにおいて一層重要な課題である。

この課題に対処するために、先行研究 [1] では前の吹き出しを考慮して 2 件ごとに翻訳する手法およびコマ単位で翻訳する手法が提案され、吹き出しごとに翻訳するベースラインと比べて翻訳品質の向上が確認された。しかし、既存研究が扱う文脈情報は翻訳対象の吹き出しが含まれるコマ内に制限されており、コマを超える範囲の文脈情報は考慮されていない。

本研究では、より長い文脈情報を考慮することを目的として、前のコマを考慮して 2 コマごとに翻訳する手法および作品や作者などの特徴を考慮して翻訳する手法の 2 つの手法を提案する。日本語から英語への漫画翻訳における評価実験の結果、提案手法はそれぞれ翻訳品質の改善に貢献することを確認できた。さらに、両手法を組み合わせることによって、吹き出しごとに翻訳するベースラインと比較して BLEU [3] を 4.24 ポイント改善し、最高性能を達成した。また、詳細な分析の結果、コマ内で言及されていない主語を補うなどのゼロ照応に関する改善の事例が見られ、漫画の機械翻訳における長い文脈情報の有用性が明らかになった。



図 1²: 漫画の機械翻訳において文脈情報の考慮が有用であると考えられるコマの例 ©朽鷹みつき

2. 関連研究

本節では、文脈を考慮する機械翻訳および機械翻訳のスタイル制御の関連研究について概説し、漫画の機械翻訳における先行研究 [1] について説明する。

2.1 文脈を考慮する機械翻訳

従来の文単位の機械翻訳には、語彙選択や照応など、文生成の一貫性に課題がある。これらの曖昧性を解消するために、先行詞を含む文を参照するなどの文脈を考慮する機械翻訳の手法 [2, 4] が研究されている。文脈を考慮する機械翻訳の手法は、結合ベースの手法 [5-8] とマルチエンコーダを用いる手法 [9-18] に大別できる。前者のアプローチは、翻訳対象の文と文脈情報を特殊トークンによって結合

[†] 愛媛大学 Ehime University

[‡] Mantra 株式会社 Mantra Inc.

¹ <https://prtimes.jp/main/html/rd/p/000000004.000059295.html>

² <https://github.com/mantra-inc/open-mantra-dataset>

し、標準的な機械翻訳モデルに入力する。後者のアプローチは、翻訳対象の文を扱うエンコーダと文脈情報を扱うエンコーダをそれぞれ用意する。結合ベースの手法は、機械翻訳のモデル構造を変更しない単純な手法でありながら、マルチエンコーダを用いる手法と同等以上の翻訳品質を達成できる [19]。そのため、漫画の機械翻訳における先行研究 [1] と同様に、本研究でも結合ベースの手法によって文脈を考慮した機械翻訳を行う。

2.2 機械翻訳のスタイル制御

機械翻訳や自動要約などで用いられる系列変換モデルの出力文のスタイルは、その訓練データに含まれるテキストのスタイルに依存する。しかし、機械翻訳などの実用に際しては、敬語への統一や出力文長の指定など、状況に応じて出力文のスタイルを制御したい場合がある。そこで、単一の系列変換モデルを用いて出力文のスタイルを制御する手法が研究されている。系列変換モデルにおける埋込層やメモリセルを変更して出力スタイルを制御する手法 [20-21] も存在するものの、多くの研究は系列変換モデルの構造を変更することなく特殊トークンを入力して出力スタイルを制御する手法 [22-28] を採用している。

機械翻訳の出力スタイルの制御に取り組む先行研究では、丁寧さ [22-24] や平易さ [25-26] の他、態 [27] や文長 [21]、さらには出力言語 [28] をスタイルとして扱っている。本研究でもこのアプローチを採用し、漫画の機械翻訳における出力スタイルを制御する。

2.3 漫画の機械翻訳

先行研究 [1] は、日本語および英語の漫画画像から、吹き出しの検出およびアライメントやテキストの検出を行い、約 400 万文対の規模の漫画の日英対訳コーパスを構築した。そして、このコーパスに基づき、漫画のある吹き出しを日本語から英語へ機械翻訳する際に、その周辺の吹き出しを文脈情報として活用する 2 つの手法を提案した。以下で概説するこれらの手法は、いずれも吹き出し単位での機械翻訳に比べて、高い翻訳品質を達成している。

2.3.1 直前の吹き出しを考慮する翻訳

本手法では、2.1 節で説明した結合ベース文脈考慮型機械翻訳 [5] のアプローチを採用して吹き出しの翻訳品質を改善している。具体的には、翻訳対象の吹き出しとその直前の吹き出しを、境界を表す特殊トークン“<SEP>”によって結合し、機械翻訳モデルに入力する。例えば図 1 において、最後の“お父さんの夢だからな”という吹き出しを翻訳する場合、機械翻訳モデルには“「あさがお」は <SEP> お父さんの夢だからな”というテキストが入力される。

2.3.2 コマ単位での翻訳

より長い範囲の文脈を考慮するために、本手法では、翻訳対象の吹き出しと同じコマに含まれる全ての吹き出しを用いて、結合ベース文脈考慮型機械翻訳を行う。例えば図 1 の右下のコマの場合、“だってお父さん全然帰って来ないし… <SEP> お仕事頑張りすぎだよ <SEP> そうか？まあ仕方ないな”という 3 つの吹き出しに分かれたテキストを、“<SEP>”の特殊トークンによって結合して用いる。



図 2: 漫画の機械翻訳において前のコマの参照が有用であると考えられる例 ©朽鷹みつき

3. 提案手法

本研究では、先行研究 [1] よりもさらに長い文脈情報を活用して漫画の機械翻訳を改善する。まず 3.1 節では、先行研究 [1] で提案された直前の吹き出しを考慮する手法とコマ単位で翻訳する手法のアイデアを組み合わせることで自然に拡張し、直前のコマを考慮したコマ単位での翻訳を提案する。次に 3.2 節では、スタイル制御の手法を応用し、作品全体の属性情報を考慮する手法を提案する。

3.1 直前のコマを考慮する翻訳

図 2 の例において、2 コマ目の先頭の吹き出しは、前のコマから続くセリフの一部である。そのため、従来の吹き出し単位の翻訳やコマ単位の翻訳では、この吹き出しのガ格は不明であり、適切に翻訳することが困難である。しかし、前のコマを参照することで、ガ格が「お前=相川仁」であることがわかり、翻訳に成功する可能性が高まる。

この課題に対処するために、本研究では、直前のコマを考慮するコマ単位の機械翻訳を提案する。本手法でも、結合ベース文脈考慮型機械翻訳 [5] のアプローチを採用し、翻訳対象のコマとその直前のコマを特殊トークンによって結合する。ただし、コマ内にも複数の吹き出しが存在し得るため、コマ内外の結合を明示するために 2 種類の特殊トークンを用いる。図 2 の例では、“相川仁 <コマ内> お前に

は～学力がある <コマ外> なのになぜ～狙う? <コマ内> 実に夢がない” というテキストを機械翻訳モデルに入力する。

3.2 漫画の属性を考慮する翻訳

漫画のジャンル (スポーツ、ラブコメ、ミステリ…) ごとに使用される表現に偏りがあったり、作者や作品ごとに表現の特徴があったりということが考えられる。このような特徴を捉えるために、スタイル制御のアプローチ [22-28] を応用して、漫画の属性を考慮する機械翻訳を提案する。例えば図 1 の作品において作者の属性を考慮する場合は、“<朽鷹みつき> おかえりお父さん!” のように、入力テキストの先頭に属性を示す特殊トークンを付与する。

本研究では、作品・作者・出版社・掲載誌・ジャンルの 5 種類の漫画の属性を扱う。これらの属性情報は Wikipedia の Infobox や漫画に関するオンライン百科事典であるマンガペディア³から収集した。

4. 評価実験

先行研究 [1] において構築された Manga Corpus を用いて日英翻訳の評価実験を行い、提案手法の有効性を検証する。

4.1 実験設定

Manga Corpus から無作為に 50 作品ずつを抽出し、各作品の最新巻から 10 ページずつを無作為抽出して検証用データおよび評価用データを構築した。その結果、検証用データおよび評価用データとして、50 作品、500 ページ、1,000 コマ、2,000 吹き出し程度の規模のデータセットを得た。その他のページを訓練用データとして使用した。

前処理として、日本語には KyTea⁴ [29] を用いた単語分割、英語には Moses⁵ [30] を用いた単語分割を行った後、Byte Pair Encoding⁶ [31] によるサブワード分割を適用した。サブワード分割の語彙サイズは、日英あわせて 32,000 とした。

機械翻訳モデルには、fairseq⁷ [32] の実装を用いて Transformer [33] を訓練した。モデルの構成は、層数を 6、注意機構のヘッド数を 16、埋め込み次元数を 1,024、全結合層の次元数を 4,096 とした。最適化関数には Adam [34] を使用し、その他の訓練の設定はバッチサイズを 1,024、ラベルスムージングを 0.1、Dropout 率を 0.3、学習率を 1e-7、Warmup ステップを 4,000 とした。訓練は、検証用データにおけるクロスエントロピー損失が 5 エポック改善されなくなる時点で停止した。

翻訳品質を自動評価するために、SacreBLEU⁸ [35] を用いて BLEU [3] を求めた。BLEU はコマ単位で計算し、異なる乱数シードを用いて訓練した 3 モデルの平均値を報告する。

比較手法には、吹き出し単位で翻訳するベースライン、直前の吹き出しを考慮する既存手法 [1]、コマ単位で翻訳する既存手法 [1] を用いる。提案手法である直前のコマを考慮する手法 (3.1 節) をこれらと比較するとともに、漫画の属性を考慮する手法 (3.2 節) をベースラインおよび提案手法に適用し、提案手法の有効性を評価する。

なお、本実験ではページをまたぐ処理は行っていない。そのため、直前の吹き出しやコマを考慮する手法において、翻訳対象がページの先頭に位置する場合には、直前の吹き

表 1: 文脈情報を考慮する機械翻訳の実験結果

		BLEU
ベースライン	吹き出し単位での翻訳	18.93
既存手法 [1]	直前の吹き出しを考慮	21.14
既存手法 [1]	コマ単位での翻訳	21.02
提案手法	直前のコマを考慮	21.46

表 2: 漫画の属性情報を考慮する機械翻訳の実験結果

		BLEU
吹き出し単位での翻訳		18.93
+ 作品		20.87
+ 作者		20.46
+ 出版社		18.91
+ 掲載誌		19.49
+ ジャンル		20.58
+ 全ての属性情報		21.08
直前のコマを考慮する翻訳		21.46
+ 作品		22.51
+ 作者		23.02
+ 出版社		22.14
+ 掲載誌		22.65
+ ジャンル		22.61
+ 全ての属性情報		23.17

出しやコマを取得できず、入力テキストに文脈情報を含められていない。

4.2 実験結果

機械翻訳モデルに入力する文脈情報を変化させた際の BLEU の評価結果を表 1 に示す。文脈情報を考慮しないベースラインモデルと比較して、既存手法および提案手法はそれぞれ 2 ポイント以上の改善が見られることから、漫画の機械翻訳において文脈情報の活用が有用であることがわかる。その中でも、最も長い文脈情報を考慮する提案手法が最高性能を達成したことから、コマを超えて長い範囲での文脈情報を活用することが重要であることがわかる。

漫画の属性情報を考慮する機械翻訳の実験結果を表 2 に示す。多くの場合において、文脈を考慮しないベースラインモデルと文脈を考慮する提案モデルの両方に対して、属性情報を用いて翻訳品質を改善できた。ただし、ベースラインモデルに出版社の情報を与えた場合のみ、BLEU の改善を確認できなかった。これは、ひとつの出版社から様々なジャンルや様々な作者にわたる多様な作品が出版されるため、出版社の情報からは翻訳スタイルを特徴づけることが難しいということだと考えられる。その他の属性情報は有効に機能しており、提案モデルの入力文に対して作品・作者・掲載誌・ジャンルの各属性情報を付与することで、BLEU を 1 ポイントから 1.5 ポイントずつ改善できた。さらに、5 種類の全ての属性情報を併用したところ、ベースラインモデルと提案モデルの両方において、それぞれ最高性能を達成した。これらの結果から、漫画の属性情報、つまり、吹き出しやコマよりもさらに長い範囲での文脈情報が、

³ <https://mangapedia.com/>

⁴ <http://www.phontron.com/kytea/>

⁵ <https://github.com/moses-smt/mosesdecoder>

⁶ <https://github.com/glample/fastBPE>

⁷ <https://github.com/facebookresearch/fairseq>

⁸ <https://github.com/mjpost/sacrebleu>

表 3: 図 2 の“実に夢がない”という吹き出しに対する機械翻訳の例

	入力	出力
吹き出し単位での翻訳	実に夢がない	I don't have a dream.
直前の吹き出しを考慮	なのになぜ～狙う? <SEP> 実に夢がない	I don't have any dreams.
コマ単位での翻訳	なのになぜ～狙う? <SEP> 実に夢がない	I don't have any dreams.
直前のコマを考慮	十年後—夏 <コマ内> 相川仁 <コマ内> お前には ～学力がある <コマ外> なのになぜ～狙う? <コマ内> 実に夢がない	You really don't have a dream.
正解文	実に夢がない	Don't you have aspirations?

表 4: ジャンルの指定によって翻訳結果が変化する例

入力	ジャンル	出力
最初から翔ばないほうがいい	<少年漫画>	You shouldn't have jumped from the start.
	<野球>	You shouldn't have hit from the start.
	<ロボット>	You shouldn't have taken flight in the first place.
	<青年漫画>	It's better if you didn't fly from the start.

漫画の機械翻訳のために有用であることがわかる。表 1 の結果と合わせて、長い文脈情報の有効性が明らかになった。

4.3 分析

直前のコマを考慮する提案手法によって翻訳品質が改善した事例を表 3 に示す。図 2 の 2 コマ目にある“実に夢がない”という吹き出しは、このコマ内にはこの吹き出しのガ格が存在しないため、コマ外の文脈を考慮しない既存手法では正確な翻訳が困難である。前のコマを参照すると、“お前には一流の大学を目指す学力がある”という吹き出しが存在し、この吹き出しのガ格が“お前 (you)”であることがわかる。表 3 に示すように、前のコマを考慮しないベースラインモデルや既存モデルは主語として誤って“**I**”を出力してしまっているが、提案モデルでは前のコマを参照できるため、適切に“**you**”を出力できた。この事例からも、長い文脈を考慮することが漫画の機械翻訳にとって有効であることが確認できる。

また、漫画の属性を考慮することによる翻訳結果の変化の事例を表 4 に示す。“最初から翔ばないほうがいい”という入力文に対して、提案手法によって異なるジャンルを指定した際の翻訳結果を観察した。<少年漫画>のジャンルを指定した際には“**jumped**”という表現が使われている部分が、<野球>のジャンルを指定した際にはボールを打つという意味の“**hit**”という表現に変化し、<ロボット>のジャンルを指定した際には飛行機に乗るという意味の“**flight**”という表現に変化した。また、<青年漫画>などのジャンルを指定した際には、単語単位の変化ではなく文の構造が大きく変化するような例も見られた。この事例から、漫画の属性を考慮することで適切な訳し分けができることが確認できる。

5. おわりに

本研究では、漫画の機械翻訳の品質を改善するために、長い文脈情報を扱う手法を提案した。具体的には、結合ベース文脈考慮型機械翻訳の技術を応用した (1) 前のコマを考慮するコマ単位の翻訳手法および機械翻訳のスタイル制御の技術を応用した (2) 漫画の属性情報を考慮する翻訳手法の 2 つの手法を提案した。日本語から英語への漫画の機械翻訳に関する評価実験の結果、2 つの提案手法はそれぞれ翻訳品質の改善に貢献し、両手法の組み合わせによ

って最高性能を達成した。これらの実験結果から、コマを超える長い文脈情報の有効性が明らかになった。

今後の展望として、大規模コーパスで事前訓練された多言語モデルや機械翻訳モデルを活用し、翻訳品質の更なる改善に取り組みたい。また、画像からの情報抽出に取り組み、マルチモーダル機械翻訳の枠組みで漫画の機械翻訳を改善したい。

謝辞

本研究は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) の助成を受けたものです。

参考文献

- [1] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards Fully Automated Manga Translation. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, pp.12998-13008, 2021.
- [2] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. ACM Computing Surveys, Vol.54, No.2, pp.1-36, 2021.
- [3] Kishore Papineni, Salim Rukos, 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.
- [4] Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. Effective Use of Target-side Context for Neural Machine Translation. In Proceedings of the 28th International Conference on Computational Linguistics, pp.4483-4494, 2020.
- [5] Jörg Tiedemann and Yves Scherrer. Neural Machine Translation with Extended Context. In Proceedings of the Third Workshop on Discourse in Machine Translation, pp.82-92, 2017.
- [6] Rachel Bawden, Rico Sennrich, Alexandra Blich, and Barry Haddow. Evaluating Discourse Phenomena in Neural Machine Translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.1304-1313, 2018.
- [7] Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. Analysing Concatenation Approaches to Document-level NMT. In Proceedings of the Third Workshop on Discourse in Machine Translation, pp.51-61, 2019.
- [8] Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. Focused Concatenation for Context-Aware Neural Machine Translation. In Proceedings of the Seventh Conference on Machine Translation, pp.830-842, 2022.

- [9] Sebastien Jean, Stanislas Lauly, Orhan First, and Kyunghyun Cho. Does Neural Machine Translation Benefit from Larger Context? arXiv:1704.05135, 2017.
- [10] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploring Cross-Sentence Context for Neural Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.2826-2831, 2017.
- [11] Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to Remember Translation History with a Continuous Cache. Transactions of the Association for Computational Linguistics, Vol.6, pp.407-420, 2018.
- [12] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.2947-2954, 2018.
- [13] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.1264-1274, 2018.
- [14] Sammen Maruf and Gholamreza Haffari. Document Context Neural Machine Translation with Memory Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.1275-1284, 2018.
- [15] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, and Yang Liu. Improving the Transformer Translation Model with Document-Level Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.533-542, 2018.
- [16] Sameen Maruf, André F.T. Martins, and Gholamreza Haffari. Selective Attention for Context-aware Neural Machine Translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.3092-3102, 2019.
- [17] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling Coherence for Discourse Neural Machine Translation. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, pp.7338-7345, 2019.
- [18] Elena Voita, Rico Sennrich, and Ivan Titov. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.1198-1212, 2019.
- [19] António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level Neural MT: A Systematic Comparison. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp.225-234, 2020.
- [20] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling Output Length in Neural Encoder-Decoders. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.1328-1338, 2016.
- [21] Sho Takase and Naoaki Okazaki. Multi-Task Learning for Cross-Lingual Abstractive Summarization. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp.3008-3016, 2022.
- [22] Rico Sennrich, Barry Haddow, and Alexandra Brich. Controlling Politeness in Neural Machine Translation via Side Constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.35-40, 2016.
- [23] Xing Niu, Marianna Martindale, and Marine Carpuat. A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.2814-2819, 2017.
- [24] Xing Niu and Marine Carpuat. Controlling Neural Machine Translation Formality with Synthetic Supervision. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, pp.8568-8575, 2020.
- [25] Sweta Agrawal and Marine Carpuat. Controlling Text Complexity in Neural Machine Translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp.1549-1564, 2019.
- [26] Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation. In Proceedings of the 13th Conference on Language Resources and Evaluation, pp.6744-6752, 2022.
- [27] Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. Improving Japanese-to-English Neural Machine Translation by Voice Prediction. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp.277-282, 2017.
- [28] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, Vol.5, pp.339-351, 2017.
- [29] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In Proceedings of the 49th Annual Meeting of the Association for Conference Linguistics: Human Language Technologies, pp.529-533, 2011.
- [30] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp.177-180, 2007.
- [31] Rico Sennrich, Barry Haddow, and Alexandra Brich. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1715-1725, 2016.
- [32] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the North American Chapter of the Association for Computational Linguistics, pp.48-53, 2019.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.5998-6008, 2017.
- [34] Diederik P. Kingma, and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, 2015.
- [35] Matt Post. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation, pp.186-191, 2018.