多段階難易度制御ニューラル機械翻訳のための ベンチマーク評価データセットの開発

谷和樹 [1]　湯浅亮也 [1]　滝川一毅 [2]　田村晃裕 [1]　梶原智之 [2]　二宮崇 [2]　加藤恒夫 [1]
[1] 同志社大学　[2] 愛媛大学
{cguc1070@mail4,cguc0095@mail4,aktamura@mail,tsukato@mail}.doshisha.ac.jp
{takikawa@ai.,kajiwara@,ninomiya@}cs.ehime-u.ac.jp

概要

本研究では，目的言語文の難易度を多段階で制御するニューラル機械翻訳（多段階難易度制御 NMT）のためのベンチマーク評価データセットを構築する．従来の多段階難易度制御 NMT の評価データは文単位での対応付けや難易度付与がされていないコーパスから自動構築されており，ノイズを含むため評価データとして不適切である．本研究では，人手による翻訳，不適切なデータの自動フィルタリングと最終的な人手確認を行うことで，Newsela コーパスから日英多段階難易度制御 NMT のためのベンチマーク評価データセットを構築する．さらに，従来研究で提案された多段階難易度制御 NMT（パイプラインモデルとマルチタスクモデル）を Transformer モデルで実装し，構築した評価データにおける性能を報告する．

1　はじめに

近年，ニューラル機械翻訳（NMT）はますます発展・普及してきており，利用者層が幅広くなっている．従来の一般的な NMT は利用者や状況に依らない一律な翻訳を行うが，近年では，目的言語文の表現を制御するための研究が盛んになっている [1, 2]．そのひとつに，ユーザの読解レベルにあわせた翻訳を行うため，難易度を入力として受け付け，指定された難易度の目的言語文を生成する難易度制御 NMT がある．これまでは，「平易」と「難解」のような 2 段階で難易度を制御するモデル [3] が中心だったが，近年では，3 段階以上の難易度を制御可能な多段階難易度制御 NMT モデルが提案され [4]，より柔軟に（例えば小学生，高校生，一般，専門家向けなどのように）難易度を制御することを目指す研究がされている．

従来の多段階難易度制御 NMT の研究で用いられている評価データ [4] は，Newsela コーパス[1)]から自動構築されている．そのため，従来の評価データには次の問題がある．(1) 不適切な対訳文対を含む．(2) 難易度が変わると情報が保たれない場合がある．(3) 目的言語文の難易度が適切でない場合がある．

そこで本研究では，人手による翻訳，不適切なデータの自動フィルタリングと最終的な人手確認を行うことで，Newsela コーパスから，日英多段階難易度制御 NMT を適切に評価するための評価データセットを構築する．提案手法により，日本語文と複数の難易度による英語文の組，1,014 組からなる評価データセットを構築した[2)]．さらに，先行研究 [4] で提案されている，パイプラインモデルとマルチタスクモデルを Transformer モデル [5] で実装し，構築した我々の評価データでの性能を報告する．

2　従来の評価データ

多段階難易度制御 NMT の唯一の先行研究は Agrawal and Carpuat の研究 [4] である．先行研究では，Newsela コーパスから自動作成したデータを用いて，英語（英）とスペイン語（西）の間の多段階難易度制御 NMT モデルの学習と評価を行っている．

Newsela コーパスは，複数の難易度の英語記事と，その一部に対応するスペイン語記事で構成されている．西英間は記事単位で対応付けられているが，文単位の対応付けは行われていない．各記事には記事単位で grade level という難易度が付与されている．grade level の値は 2 から 12 であり，値が高いほど記事が難しいことを示す．評価データは，この Newsela コーパスから次の通り作成されている．

1） https://Newsela.com/data/
2） https://github.com/K-T4N1/A-BenchmarkDataset-for-ComplexityControllableNMT.git で公開予定である．

日本語文	難易度	英語文
欧米の優柔不断な行動がプーチンを勢いづかせている。	12	Weak Western action is emboldening Putin.
	9	Weak Western action is strengthening Putin.
	7	Weak Western action is just encouraging Putin.
公衆衛生についての入門書が必要だ。	12	A primer about public health is in order.
	9	A short explanation about public health is in order.
	6	A short explanation about public health is needed.

表 1　日英多段階難易度制御 NMT 用評価データの実例

Step 1　西英の Google 翻訳[3] を用いて，スペイン語記事の各文を英語に翻訳する．

Step 2　同一言語で書かれた同一内容の文を対応付ける MASSAlign [6] により，英語記事内の文と，英語に翻訳されたスペイン語文を対応付ける．

Step 3　対応付いた文の組を 1 事例とする．その際，翻訳された文は元のスペイン語文に戻す．また文の難易度は，属する記事の grade level にする．

この自動作成の評価データには次の問題がある．

問題点 1　西英文対は自動で対応付けているため，正しい翻訳文対になっているとは限らない．

問題点 2　同一言語の文間の対応付けを自動で行っているため，対応付いている文の情報の抽象度が同一になるとは限らない．特に，固有名詞など，より具体的な情報が新たに湧き出す場合，難易度制御で情報を増やすことは難しいため問題となる．

問題点 3　記事の難易度を文の難易度に転用しているため，付与されている文の難易度が正しいとは限らない．例えば，記事単位の grade level が異なっていたとしても，完全に同じ文となる場合や記号のみが異なる場合がある．

このような問題がある従来の評価データでは，モデルの正確な性能評価ができなくなりうるため，評価データとして不適切である．

3　評価データセットの作成

本節では，本研究で提案する多段階難易度制御 NMT 用の評価データセットの作成手順と，作成した評価データセットの詳細を示す．本研究では，Newsela コーパス中の英語文に日本語訳を付与することで，日本語文（原言語文）1 文と複数（3 段階以上）の難易度による英語文（目的言語文）の組で構成される日英多段階難易度制御 NMT 用の評価データセットを作成する．Newsela コーパスは，他の単言語平易化コーパス（ [7] や [8] など）とは異なり，ニュース記事を作成するプロにより平易化されていること，規模が大きいこと，また，多段階の難易度が付与されていることから，本研究の評価データセットの作成元にした．

3.1　提案の作成手順

評価データセットは次の 2 ステップで構築する．

Step 1　複数段階の難易度で記述される英語文集合の作成

1-1: Newsela-auto から自動抽出

1-2: 自動フィルタリングでの不適切データ除去

1-3: 人手による最終確認

Step 2　人手翻訳による日本語訳の付与

Step1 では，Newsela コーパスの英語記事集合から複数段階の難易度で記述される同一内容の英語文集合を抽出する．具体的には，Newsela-auto [9][4] 内で自動で対応付いている文対に基づき，3 つ以上の文が同一内容となる英語文集合を抽出する（Step 1-1）．その結果，98,500 組の英語文集合が得られた．

ただし，本研究でも先行研究と同様，文の難易度として属する記事の難易度を設定する．したがって，2 節の問題点 3「文の難易度が正しいとは限らない」が生じる．そこで本研究では，自動フィルタリング（Step 1-2）と人手チェック（Step 1-3）により，確実に難易度が異なる 3 文以上から成る同一内容の英語文集合を作成する．自動フィルタリングでは，記号を除いた状態で全く同じになる文対と grade level の差が 1 以下になる文対を除いた．

また，Newsela-auto は同一内容の文を自動で対応付けているため，2 節の問題点 2「難易度が異なると情報が保たれない場合がある」が生じる．そこで本研究では，Step 1 の最終確認として，固有名詞などの情報が湧き出している事例を人手で除いた．

Step2 では，Step1 で作成した各英語文集合に対して，最も grade level が大きい文（最も難しい文）を人手で日本語に翻訳し，日本語訳を付与する．翻訳

3）https://translate.google.com/

4）Newsela コーパス中の難易度が異なる英語記事を対象に，同一内容の文を自動で対応付けした結果のコーパスである．

段階数	3	4	5
組数	906	97	11
割合	89.34	9.56	1.08

表 2　評価データセット中の英語文の段階数

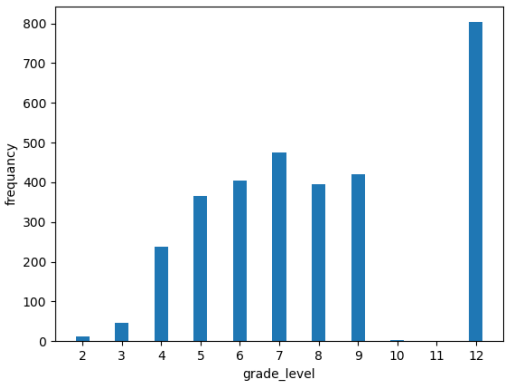

図 1　評価データセット中の英語文の grade level の内訳

は翻訳会社に依頼した．この人手翻訳により，2 節の問題点 1「不適切な対訳文対を含む」を解決する．

以上の作成手順により，1,014 組の評価データからなる，日英多段階難易度制御 NMT 用の評価データセットを作成した．表 1 に作成した評価データの実例を示す．また，表 2 と図 1 に，作成した評価データセット中の英語文の難易度の段階数の内訳と英語文の grade level の内訳をそれぞれ示す．

自動フィルタリングの必要性の考察

本研究では，Step 1 において，記事単位の難易度から転用された文の難易度が不適切なデータを除く目的で自動フィルタリングを導入した．ここでは，その必要性を考察する．

Step 1-1 の結果からランダムに抽出した 100 組の文集合に対して，文に付与された難易度が異なるにもかかわらず，実際には難易度が変わらないデータがどの程度含まれるかを人手で調査した．その結果，実際に難易度が変わっている文で構成されている組は 37 組しか存在せず，残りの 63 組は，grade level が異なるにもかかわらず，全く同じ文対や，記号のみが変化するなどの難易度が変わらない文対を含んでいた．このような事例は評価データとして不適切であるため，評価データから除く必要がある．しかし全体の 60%以上あり人手で除くのは現実的ではない．したがって，本研究で導入した自動フィルタリングが有用であると考えられる．

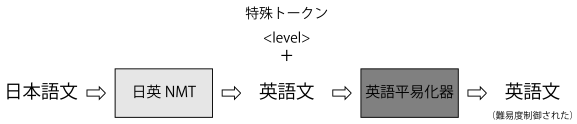

図 2　パイプラインモデルの概要図

4　ベンチマーク実験

本節では，今後の多段階難易度制御 NMT に関する研究のベンチマークとして，先行研究 [4] で提案されている 2 つの多段階難易度制御 NMT の手法（パイプラインモデルとマルチタスクモデル）を Transofrmer モデルで実装し[5]，構築したテストデータに対する性能を報告する．先行研究では系列変換モデルとして RNN（LSTM）ベースのモデル [11] を用いているが，本研究では，近年様々な NLP タスクでデファクトスタンダードになっている Transformer モデルを用いた．また，先行研究では西英間の翻訳を対象としており，教師データとして Newsela コーパスのスペイン語記事を用いることができるが，本研究では Newsela コーパスに含まれない日本語と英語間の翻訳が対象であることにも注意されたい．

4.1　パイプラインモデル

パイプラインモデルは，機械翻訳と多段階平易化 [12] の 2 つの独立したモデルをつなげたモデルである．「機械翻訳 → 多段階平易化」と「多段階平易化 → 機械翻訳」の 2 通りが考えられるが，先行研究 [4] では「機械翻訳 → 多段階平易化」の方が性能が良かったこと，また，Newsela コーパスからは日本語の多段階平易化のための教師データを作成できないことから，本研究では，「機械翻訳 → 多段階平易化」の性能のみを評価，報告する．図 2 に評価するパイプラインモデルの概要図を示す．

日英 NMT モデルおよび英語の多段階平易化モデルは共に Transformer モデルを用いる．日英 NMT モデルは，Kiyono ら [13] の日英翻訳モデルに倣って作成した．教師データは JParaCrawl [14] と News Commentary を用いた．ただし，langid[6] により原言語文が日本語かつ目的言語文が英語となった，9.7M 文対を教師データとして用いた．英語文にのみ trucasing を行い，Sentencepiece [15] によりサブワードサイズ 32,000 でサブワード化を行った．

英語の多段階平易化モデルは，先行研究 [12] に倣い，入力の英語文の先頭に目的の難易度を表す

5）本実験ではフレームワークとして fairseq [10] を用いた．
6）https://github.com/saffsd/langid.py

特殊トークンを追加した系列を，目的の難易度の英語文に変換する系列変換モデルで実現した．Newsela-auto からランダムに抽出した 150K を教師データとした．ただし，構築した評価データセットに含まれるデータは除いている．教師データからは 100 単語以上の文と入出力文長比が 2 以上の文対は除いた．その後，fastBPE[7] によりサブワードサイズ 8,000 でサブワード化を行った．その他の実験設定は付録に示す．

4.2 マルチタスクモデル

マルチタスクモデルは，翻訳のみの損失，平易化のみの損失，難易度を指定した翻訳の損失の 3 つの損失に基づいた，次の損失関数 $loss$ により，1 つの Transformer モデルを学習する．

$$loss = L_{MT} + L_{Simplify} + L_{CMT} \quad (1)$$

$$L_{MT} = \sum_{(s_i, s_o) \in D_{MT}} \log P(s_o | s_i; \theta) \quad (2)$$

$$L_{Simplify} = \sum_{(s_o, c_{o'}, s_{o'}) \in D_S} \log P(s_{o'} | s_o, c_{o'}; \theta) \quad (3)$$

$$L_{CMT} = \sum_{(s_i, c_o, s_o) \in D_{CMT}} \log P(s_o | s_i, c_o; \theta) \quad (4)$$

ただし D_{MT} は，JparaCrawl からランダムに抽出した 3,000K の日英対訳文対，D_{CMT} は，Newsela-auto に含まれる各英語文集合に対して，最高難易度の英語文を Google 英日翻訳で日本語に翻訳し，翻訳した日本語（原言語）と最高難易度以外の英語文（目的言語）とその難易度の三つ組 (評価データセットに含まれるデータは除く) をランダムに抽出した 200K，D_S は，D_{CMT} の 200K の三つ組に対して，日本語文を元の英語文に置き換えたデータである．また，θ はモデルパラメータ，s_i は原言語文，s_o は目的言語文，$s_{o'}$ は s_o を平易化した文，$c_{o/o'}$ は $s_{o/o'}$ の難易度である．その他の実験設定は付録に示す．

4.3 評価指標

評価指標は，先行研究 [4] に倣って BLEU [16] と SARI [7] [8] を用いる．また，出力文の文長と参照文の文長の絶対平均誤差 MAE$_{\text{LEN}}$ [17] も算出する．ここで，SARI [7] を算出するためには，同一言語の，平易化前の文，平易化後の文，参照文の三つ組が必要になるが，従来の評価データでは原言語文と同じ難易度の目的言語文が必ずしも与えられていない．そこで先行研究 [4] では，原言語文を機械翻訳で翻訳した文を，平易化前の文として用いている．しかし，機械翻訳の翻訳文には誤りが多く含まれるため，先行研究の SARI は，本来評価すべき多段階難易度制御 NMT の純粋な平易化性能を評価できていない．一方で，本研究で作成した評価データセットには，原言語文に対応する最高難易度の英語文が付与されているため，SARI を計算する際に最高難易度の英語文を平易化前の文として与えることで適切な SARI を算出できる．

モデル	BLEU (%)	SARI (%)	MAE$_{\text{LEN}}$
パイプライン	15.11	33.51	5.10
マルチタスク	15.89	33.32	4.15

表 3 実験結果

4.4 実験結果

4.1 節のパイプラインモデルおよび 4.2 節のマルチタスクモデルの性能を表 3 に示す．表 3 より，先行研究同様，マルチタスクモデルの方がパイプラインモデルよりも BLEU が高くなった．一方で，SARI はパイプラインモデルの方が高くなった．

パイプラインモデルについては，機械翻訳部分と多段階平易化部分それぞれ単独の評価データセットにおける性能を評価した．多段階平易化部分の性能評価では，各英語文集合の中の最高難易度の英語文を入力として，その他の難易度の英語文に対する平易化性能を測った．その結果，機械翻訳部分の BLEU は 14.31%となった[9]．また，多段階平易化部分の BLEU，SARI，MAE$_{\text{LEN}}$ は，それぞれ，68.40%，18.99%，3.98 となった．

5 まとめ

本研究では，従来の評価データより適切にモデル性能を評価可能な，多段階難易度制御 NMT のためのベンチマーク評価データセットを構築した．人手による翻訳，不適切なデータの自動フィルタリングと最終的な人手確認を行うことで，評価データの品質を確保した．また，今後の研究のためのベンチマーク結果として，2 つの多段階難易度制御 NMT モデルを実装し，その性能を構築した我々のテストデータで評価した．

7） https://github.com/glample/fastBPE.git
8） https://github.com/cocoxu/simplification
9） 参考のため，Google 日英翻訳の我々の評価データセットにおける翻訳性能を測った結果，BLEU は 10.12%となった．

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

[2] James Kuczmarski and Melvin Johnson. Gender-aware natural language translation. In **Technical Disclosure Commons, (October 08, 2018)**, 2018.

[3] Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. Controlling the reading level of machine translation output. In **Proceedings of Machine Translation Summit XVII: Research Track**, pp. 193–203, 2019.

[4] 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 (EMNLP-IJCNLP)**, pp. 1549–1564, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[6] Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. MASSAlign: Alignment and annotation of comparable documents. In **Proceedings of the IJCNLP 2017, System Demonstrations**, pp. 1–4, 2017.

[7] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[8] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4668–4679, 2020.

[9] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF model for sentence alignment in text simplification. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7943–7960, 2020.

[10] 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, 2019.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In **Proceedings International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[12] Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 712–718, 2018.

[13] Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. Tohoku-AIP-NTT at WMT 2020 news translation task. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 145–155, 2020.

[14] 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, Marseille, France, May 2020. European Language Resources Association.

[15] 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.

[16] 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.

[17] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable text simplification with lexical constraint loss. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 260–266, 2019.

A 付録

表 4 に 2 節で記述した問題点 2「難易度が異なると情報が保たれない場合がある」の具体例を示す．また，表 5 に 2 節で記述した問題点 3「文の難易度が正しいとは限らない」の具体例を示す．そして，表 6 に本研究で評価したパイプラインモデルとマルチタスクモデルの実験設定を示す．

level	英語文
8	However, she says that there are times when "you just need to get away."
5	Yet she says that there are times when "you just need to get away."
3	Still, **Bopp** says that there are times when "you just need to get away."
12	The White House said the U.S. will suspend participation in preparatory meetings for the G-8 economic summit planned.
7	The White House said the U.S. will stop participating in planning meetings for the G-8 economic summit.
5	The White House said the U.S. will stop participating in meetings about the G-8 summit **in Russia**.

表 4 固有名詞の湧き出しを含むデータ例

level	英語文
12	**A company called AquaBounty has been seeking for more than 20 years to win FDA approval to bring a genetically modified fast-growing salmon to supermarkets.**
9	A company called AquaBounty has been seeking for more than 20 years to win Food and Drug Administration (FDA) approval to bring a genetically modified fast-growing salmon to supermarkets.
7	**A company called AquaBounty has been seeking for more than 20 years to win FDA approval to bring a genetically modified fast-growing salmon to supermarkets.**
12	So few Indians drink brewed coffee that virtually all its best crop is exported to countries such as Italy, where the beans are used in name-brand espresso blends and sold at a huge markup.
9	**There the beans are used in name-brand espresso blends and sold at a huge price increase.**
7	**There, the beans are used in name-brand espresso blends and sold for a huge price increase.**

表 5 grade level は異なるが難易度が変わらない文対を含むデータ例

	Pipeline Model		Multi-Task Model
	NMT Model	Simplification Model	
arch	transformer	transformer	transformer
share-decoder-input-output-embed	True	True	True
activation-fn	relu	relu	relu
optimizer	adam	adam	adam
adam-betas	'(0.9, 0.98)'	'(0.9, 0.98)'	'(0.9, 0.98)'
clip-norm	1.0	0.0	1.0
lr	7e-4	7e-4	5e-4
lr-scheduler	inverse_sqrt	inverse_sqrt	inverse_sqrt
warmup-updates	4000	4000	4000
warmup-init-lr	1e-7	1e-7	1e-7
weight-decay	0.0001	0.0001	1e-5
dropout	0.3	0.1	0.1
criterion	label_smoothed_cross_entropy	label_smoothed_cross_entropy	label_smoothed_cross_entropy
label-smoothing	0.1	0.1	0.1
max-tokens	40000	80000	4096
patience	-	-	5
fp 16	True	True	True
max-epoch	100	100	100,000,000

表 6 fairseq を用いた実験設定