疑似データを用いた GPT-2 による日本語文章の多段階平易化

郷原聖士 [1]　綱川隆司 [1]　西田昌史 [1]　西村雅史 [1]
[1] 静岡大学情報学部
gobara.seiji.21@shizuoka.ac.jp,{tuna,nishida,nisimura}@inf.shizuoka.ac.jp

概要

現在，我々は常に新しい情報を取捨選択する情報社会で暮らしているが，それらの情報の内，多くはある程度習熟した成人が対象の文書である．したがって，まだ文書を理解するための知識が不足している子供や留学生などの非母語話者にとって，それらの情報を理解して生活に役立てるのは難しいという問題がある．そこで我々は，一般向けの文書を利用者の日本語理解度に応じて適切な難易度の情報に変換するための日本語の機械学習モデルを疑似データによって作成した．自動評価実験において，目標とする難易度に応じて平易化文に難易度差を付与できていることが評価指標から示唆されたが，人手評価実験においては有意な結果は得られなかった．

1　はじめに

テキスト平易化は，難解な文章から同一の意味を保持しつつも平易な文章へと変換させる，自然言語処理におけるテキスト生成タスクの一つである．

関連研究には，English Wikipedia と Simple English Wikipedia のパラレルコーパス (EW-SEW) を用いた統計的機械翻訳による手法 [1][2] やニューラル機械翻訳による手法 [3][4] がある．

統計的機械翻訳による手法は，分割，削除，並び替え，置換を総合的にカバーする一方で，EW-SEW のような大規模のパラレルコーパスが必要であり，そのような資源のない言語以外にそのまま適用出来ない．ニューラル機械翻訳による手法は，統計的機械翻訳と同様に EW-SEW のような大規模の平易化用データセットを用いて難解な文と平易な文のペアを学習することで系列変換を行う注意機構を組み込んだシステムを構築し，高精度の平易化を実現した．また，近年では平易化用の大規模データセットが存在しない言語を対象に，疑似的な平易化用データセットを構築することで平易化を実現する手法 [5] が提案されているが，いずれも平易化後の文章の難易度は考慮されていないという課題がある．

日本語においては，中町ら [6] が事前学習済み系列変換モデルにやさしい日本語対訳コーパス [7][8] を用いて平易化を実現している．最近では，コーパスベースの手法 [3][6] が多く，ターゲットの難易度は学習データに依存している．一方で平易化のターゲット文は多様であり，ユーザーに合わせて難易度を調節できた方が望ましい．

英語の平易化の研究におけるターゲット文の難易度には，EW-SEW をベースにした難解文から平易文へと変換を行う二段階の平易化が用いられてきたが，Xu ら [9] は，文章の複雑さを測定して，生徒の読解力を評価するために広く用いられている Lexile[1)] の下で 11 段階の難易度付きパラレルコーパス Newsela を構築した．Newsela を用いて難易度ラベルを入力文の文頭に付与した上で学習することで，生成するテキストの多段階な難易度制御を実現する手法 [10]，その手法をベースに目標の難易度に適した単語を出力するため，単語の分散表現を拡張した素性の利用や，ハード・ソフトな語彙制約を課す手法 [11]，語彙レベルや文長などの文章中の特徴量に着目して生成文を制御する手法 [12] がある．

上記のように，英文の多段階平易化に関する研究は盛んだが，和文を対象とする多段階平易化に関する研究は少ない．この背景には，日本語の多段階平易化のためには Newsela のような難易度付き大規模データセットが新たに必要だが，構築には莫大なコストがかかるという問題がある．

そこで本研究では，言語資源の少ない日本語の平易化用のデータセットを補うために，やさしい日本語対訳コーパスを用いてファインチューニングした GPT-2 ベースの平易化モデルと，BERT(Bidirectional Encoder Representations from Transformers）ベースで作成した難易度推定器を組み合わせることで疑似的な難易度付きパラレルコーパスを構築した．その後，Alessio ら [5] や Eriguchi ら [13] と同様にして，

1） https://lexile.com

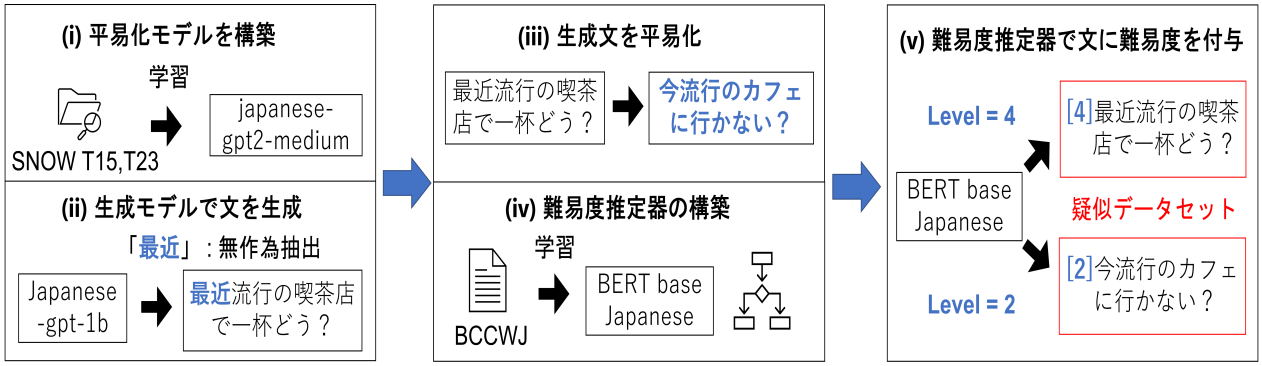

図 1　多段階平易化用疑似データセットの作成

任意の X-Y 方向の多言語機械翻訳を任意の X-Y 方向の多段階平易化と見なし，GPT-2 モデルに疑似データセットでファインチューニングした 5 つのモデルを組み合わせて多段階平易化システムを構築した．

2　手法

2.1　難易度推定器

郷原ら [14] と同様に，東北大学乾研究室が公開している日本語の大規模モデル BERT[2] をベースにファインチューニングして難易度推定器を作成した．なお，文単位の難易度推定を実施するにあたって，先行研究 [10][11] と同様に，文の難易度はその文が含まれる文書の難易度と同一のものであるという仮定の下で正解データの作成を行った．

2.2　疑似データセット構築

疑似データセットを構築する手段を図 1 に記述する．(i) やさしい日本語対訳コーパス [7][8] に含まれている日本語の平易化文対を用いて rinna 社の公開している事前学習済み日本語 GPT-2 モデル[3] をファインチューニングし，平易化モデルを構築する．(ii)rinna 社の事前学習済み日本語 GPT-2 モデル[4] において，日本語辞書が持つ語彙の内，日本語であり文頭に適切と思われる文字や語彙から始まるトークンをプロンプトに入力し，100 文字を上限に文を生成する．(iii) (ii) で生成した各文を (i) で構築したモデルに入力し，疑似的な平易化文対を得る．(iv) 2.1 節で述べた手法で難易度推定器を構築する．(v) (iii) の平易化文対に対して，(iv) で作成した難易度推定器によって文の難易度を疑似的に付与する．

ここで，(iii) での平易化は一段階のため，同じ文に対して複数回平易化を適用しても，明確な差異のある平易化文対を得るのは難しい．そこでランダムに生成された文に対する平易化文を十分な規模で収集し，同じ難易度帯に変換する文対毎にデータセットを構築することで，段階的な平易化を実現する．

2.3　疑似データセットのフィルタリング

2.2 節で作成したデータセットにおいて，平易化前後で文の難易度が変化しない文対や同義性が保たれないノイズとなるような文対が存在していた．そこでノイズとなる文対を除去するために，データセットの中から難易度が変化する文対のみを抽出する．さらに同義性を一定レベルで担保するために，3.3 節で後述する BERTScore を計算した．生成した平易化文全体の内，BERTScore が上位 50%の文対はある程度の同義性を保っている一方，上位 50%-75%の文対は同義性が保たれている文対とやや意味が異なっている文対が含まれていた．また，BERTScore が下位 25%の文対に関しては，多くが意味の異なる文対になっていた．ここで，Newsela[9] などを用いた先行研究 [10][11][12] では，各難易度に対して数万文対のデータセットを用いていたため，本研究でも同程度の文対を確保するために，生成文と BERTScore のバランスを鑑みて，BERTScore が上位 75%の文対を抽出した．なお，BERTScore が上位 75%の文対を抽出するための閾値は 0.710 であった．

3　実験設定

3.1　データセット

難易度推定器作成にあたって，データセットには現代日本語書き言葉均衡コーパス (BCCWJ)[15] に含

2)　https://github.com/cl-tohoku/bert-japanese
3)　https://huggingface.co/rinna/japanese-gpt2-medium
4)　https://huggingface.co/rinna/japanese-gpt-1b

まれている日本語教科書コーパス及び図書館サブコーパスの文書中から無作為に文を抽出し，訓練，開発，評価データをそれぞれ 236,773 件，28,921 件，29,381 件に分割した．なお，正解値は，日本語教科書コーパスの対象学年である小学校低学年，中学年，高学年，中学校，高校及び図書館サブコーパス中の「やや専門的な一般向き」に相当する一般の 6 段階に設定した．

疑似データセットの作成にあたって，rinna 社の公開している GPT-2 モデルにおいて，max_length と min_length は 100 に設定し，top_k=500，top_p=0.95 にした上で文章生成を行った．なお，文章生成は日本語辞書に登録されている単語から，絵文字などの文頭に付与する単語には不適切であると考えられる単語をノイズとして除去した上でサンプリング抽出した 1 単語をプロンプトに与え，続く単語列を推測させる形で行った．その後，生成した文章から句点区切りで文を抽出し，平易化する前の文とした．平易化前後の文の一部には，文法的に正しくても意味的には不自然な文が存在していたが，本研究では，語彙的換言や複雑な文構造の簡単化などを段階的に行うことが目的であるため，正常な文として許容した．上記生成文を 100 万文，200 万文ずつ作成し，難易度推定器の推定値に差がある文対のみをそれぞれ抽出した (以下，本設定を model-100 及び model-200 と呼ぶ．以降も同様にする)．また，開発データにおいて性能がより高かった model-200 の学習データに対して，BERTScore によるフィルタリングを行ったモデルも作成した (model-200-bert)．ここで，得られた難易度差のある文対の訓練データは model-100，model-200，model-200-bert それぞれ 337,413 件，674,699 件，487,980 件であり，開発データ及び評価データは 1,000 件とした．

3.2 モデル

rinna 社の公開している GPT-2 モデルを，3.1 で作成した多段階平易化のための疑似データセットを用いてファインチューニングを行った．なお，モデルは huggingface のライブラリ[5)]を用いて実装を行った．また，本研究では，想定する難易度を 6 段階 (0-5) に設定したため，最高難易度 (5) からそれ以下の 5 段階の難易度 (0-4) への平易化モデルを構築した．

3.3 評価指標

実装したモデルが段階的に文の難易度を易しく変換出来ているかを評価するために，2 つの自動評価指標と人手評価を用いた．

自動評価のための評価指標には，**BERTScore**[16] 及び**平均推定難易度**を用いた．BERTScore は，翻訳や文書生成のタスクにおいてしばしば用いられる指標であり，BERT を用いて文書と参照文の類似性を評価することが出来る．平均推定難易度は，文の難易度を制御して平易化する場面において用いられる指標であり [11]，英語においては，FKGL[17] や平均 PMI(Pointwise Mutual Information) が用いられている．一方で，本研究では日本語を対象であり，人手で作成された多段階平易化のための評価用データセットはほとんど存在せず，単に英語と同様の手法を用いても言語の不一致によって正しく性能を図ることは困難である．そこで本研究では，文単位の難易度推定器を作成し，難易度推定器の推定した平均難易度値を評価指標に用いた．ここで，作成した難易度推定器の妥当性を検証するため，既存の日本語用の文章難易度システム (JReadability.net[6)]，帯 [18]) を用いて評価データからサンプリングした 200 文の推定難易度を求め，BERT ベースの難易度推定器の推定値との相関係数を求めた．

人手評価では，先行研究 [10][11] を元にして，各難易度に相当する文及び各段階に相当する流暢性や同義性の破綻具合を表す文を示した上で，文の同義性 (synonymity)・平易化前後の難解性 (difficulty_o, difficulty_s)・流暢性 (fluency) の 4 項目の評価をクラウドワーカーに委託した．なお，各項目は難易度推定器の段階に合わせて 0 から 5 までの 6 段階 (5 が最良) に設定し，作成した平易化モデル毎に 30 問ずつ評価用データセットから無作為に抽出し，各段階で異なる 100 文ずつの回答が得られるようにして平均難易度を求めた．ここで，少しでも多くの文に対するアノテーションを行うべく 20 人のクラウドワーカーに依頼して，一文あたり少なくとも 6 回答が得られるようにそれぞれのワーカーに異なるデータセットを作成した．その後，得られたデータセットの内，20 人のワーカー間においてそれぞれ共通の回答が得られた文に対する順序尺度における重み付けの指標 QWK(Quadratic Weighted Kappa) を求めた．その後，平均 QWK の中でも難解性につい

5) https://github.com/huggingface/transformers

6) https://jreadability.net/

表 1　難易度推定器の評価結果

	pred-bert				
	MAE	Pearson	Spearman	Acc.	F1
pred-BERT	0.166	0.892	0.893	0.872	0.791

表 2　既存難易度推定システムとの相関

	pred-bert	jread-score	pred-B9	pred-T13
level	0.866	0.182	0.501	0.640
pred-bert	-	0.246	0.562	0.640
jread-score	-	-	0.238	0.147
pred-B9	-	-	-	0.650

ては，0.039-0.372 程度と極めて低い数値であったため，ワーカー間の一致度を向上させるべく，各ワーカーにおける QWK の平均値が 0.16 以下であるワーカーのデータを除去した．また，アノテーションミスなどの実験不備のあるデータを除去することで，全 3,000 件のデータの内 2,249 件を人手評価の対象とした．

4　実験結果と考察

4.1　難易度推定

表 1 及び表 2 に難易度推定器の評価実験結果を示す．表 1 では，それぞれ平均絶対値誤差 (MAE)，ピアソンの相関係数 (Pearson)，スピアマンの相関係数 (Spearman)，正解率 (Acc.)，F1-Score(F1) を示しており，郷原ら [14] の文章単位の評価結果とも遜色ない数値であることが分かる．また，表 2 では，それぞれ正解値 (level)，提案手法の推定値 (pred-bert)，Jreadability.net の出力値 (jread-score)，帯の出力値 (pred-B9，pred-T13) を示している．pred-bert と level 及び帯の出力値との相関関係は高く，推定された難易度が妥当であることを示唆していると考えられる．

4.2　平易化の自動評価

表 3 に平均推定難易度による平易化モデルの評価結果を示す．表中の mean は平均推定難易度，BScore は平易化前後の文の BERTScore の値の平均値を示す．表 3 において，0 を除いた各目標難易度の上昇に合わせて平均推定難易度の数値が上昇しているため，疑似的なデータセットの作成によって段階的な平易化が実現出来たと考えられる．ここで，目標難易度 0 に変換する平易化文対は他と比べて著しく少なかったため，データ不足の影響を受けたと推測される．

表 3　平易化モデルの平均推定難易度による評価結果

	model-100		model-200		model-200-bert	
目標難易度	mean	BScore	mean	BScore	mean	BScore
0	4.025	0.732	3.792	0.763	4.173	0.747
1	3.490	0.749	3.395	0.753	3.564	0.762
2	3.648	0.752	3.632	0.758	3.711	0.773
3	3.837	0.761	3.883	0.765	3.960	0.775
4	4.155	0.761	4.169	0.765	4.167	0.786

表 4　model-200-bert の人手評価の平均値による評価結果

目標難易度	synonymity	difficulty_o	difficulty_s	fluency
0	1.862	3.411	2.556	3.740
1	2.347	3.313	2.488	3.331
2	2.430	3.324	2.577	3.604
3	2.344	3.431	2.535	3.400
4	2.562	3.192	2.595	3.582

4.3　平易化の人手評価

表 4 に人手評価による評価結果を示す．表 4 では，各目標難易度による difficulty_s (平易化文の難解性) に有意な差が見られなかった．ここで，評価実験には生成文を元にした平易化文を用いており，意味的，文法的に不自然な文が含まれている．したがって，単語単位では平易化が実現出来ていても，文単位では意味が破綻している文を許容している．加えて，本タスクにおける同義性・難解性・流暢性評価は，各段階に相応の例文を提示しているものの明確な判断基準を与えていないため，人手ではやや難易度が高く，QWK の値も低いことからラベリング結果の個人差が大きいことが示唆される．人手評価上では段階的な平易化を十分に確認出来なかったと考えられる．

5　おわりに

本研究では，多段階平易化のための言語資源が不足している日本語において，既存のデータセットをベースに疑似的な多段階平易化のためのデータセットを作成することで，多段階平易化モデルの構築を行った．その結果，自動評価の平均難易度推定値によって生成文が段階的に平易化出来ることを示した．一方で，人手評価に不備があり，かつ生成した平易化文は疑似データセットの影響を受けて，無理な変換や文法的な破綻をしてしまうという課題がある．そこで今後は，人手評価の再実施に加えて，同義性の担保や流暢な多段階平易化を実現するために，多段階平易化のための質の良い半自動データセットの作成を行いたいと考えている．

参考文献

[1] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In **Proceedings of the 23rd International Conference on Computational Linguistics**, 2010.

[2] 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.

[3] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, 2017.

[4] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. Integrating transformer and paraphrase rules for sentence simplification. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, 2018.

[5] Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and A Di Gangi Mattia. Neural text simplification in low-resource conditions using weak supervision. In **Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)**, 2019.

[6] 中町礼文, 梶原智之. 事前学習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. 情報処理学会第 83 回全国大会, 2021.

[7] Takumi Maruyama and Kazuhide Yamamoto. Simplified corpus with core vocabulary. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[8] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[9] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 283–297, 2015.

[10] 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)**, 2018.

[11] 西原大貴, 梶原智之, 荒瀬由紀. テキスト平易化における語彙制約に基づく難易度制御. 自然言語処理, Vol. 27, No. 2, pp. 189–210, 2020.

[12] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. Controllable sentence simplification. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020.

[13] Akiko Eriguchi, Shufang Xie, Tao Qin, and Hany Hassan. Building multilingual machine translation systems that serve arbitrary XY translations. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2022.

[14] 郷原聖士, 綱川隆司, 西田昌史, 西村雅史. BERT による日本語文章の難易度推定. 第 21 回情報科学技術フォーラム, 2022.

[15] 言語資源開発センター. 現代日本語書き言葉均衡コーパス (BCCWJ). 国立国語研究所, 2014.

[16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[17] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[18] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

A 付録

A.1 平易化文の生成例

表 5 平易化の例文

目標難易度	例文
0	..... 作業中の雑談は作業の邪魔となりますので慎んで下さい。
1	会議中の会話は仕事の**進め方に問題を与えます**ので、気をつけましょう。
2	会議をしている時などの話は作業の**進め方に関係があります**ので、気をつけましょう。
3	仕事中の**会話**は仕事の邪魔となりますので、静かにしてください。
4	**会議**中の雑談は仕事の**邪魔**となりますので**注意して**ください。
original(5)	業務中の雑談は作業の妨げとなりますので慎んで下さい。

平易化文の生成例を表 5 に示す．表 5 において，目標難易度 4 では元の文から「業務」，「妨げ」，「慎んで」がそれぞれ平易な単語に変換されている．また，目標難易度 3 では，元の文の「雑談」，目標難易度 2 及び 1 では，目標難易度 3，4 において「妨げ」を変換した「邪魔」から「進め方に関係があります」というように，さらに平易な語彙に変換していることが分かる．

A.2 疑似データセットの詳細な内訳

表 6 疑似データセットの目標難易度別内訳

	model-100		model-200		model-200-bert	
目標難易度	orig	simp	orig	simp	orig	simp
0	0	770	0	1,549	0	938
1	38	43,357	81	86,943	71	68,312
2	4,763	114,609	9,570	229,326	8,491	166,453
3	22,140	94,005	44,211	188,026	35,567	137,094
4	43,698	84,672	87,464	168,855	61,414	115,183
5	266,774	0	533,373	0	382,437	0

表 6 に作成した疑似データセットの目標難易度別内訳を示す．表中の orig は生成文のデータ数，simp は平易化文のデータ数を表している．表 6 において，特に難易度 0 への変換を行う文対が少ないが，これは元となったやさしい日本語対訳コーパス [7][8] での文対中で平易化文が目標難易度 0 になる言い換えが少ないことが原因であると考えられる．