擬似データを用いた教師あり学習による語彙平易化

野口 夏希 [1]　梶原 智之 [1]　荒瀬 由紀 [2]　内田 諭 [3]　二宮 崇 [1]
[1] 愛媛大学　[2] 大阪大学　[3] 九州大学
n_noguchi@ai.cs.ehime-u.ac.jp　{kajiwara, ninomiya}@cs.ehime-u.ac.jp
arase@ist.osaka-u.ac.jp　uchida@flc.kyushu-u.ac.jp

概要

本研究では，文平易化のための大規模なパラレルコーパスから語彙平易化のための擬似的なパラレルコーパスを自動抽出し，マスク言語モデルを再訓練する教師あり学習の枠組みで語彙平易化を行う．語彙平易化のためのパラレルコーパスはアノテーションコストが高いため，千件未満の小規模なものしか存在しない．そのため先行研究では，パラレルコーパスを用いた語彙平易化モデルの教師あり学習は検討されていない．本研究では，擬似的なパラレルコーパスを用いた教師あり学習によって，3 つの英語の語彙平易化タスクで最高性能を達成した．

1　はじめに

子どもや英語非母語者などの読解支援のために，テキスト平易化 [1] の研究が行われている．テキスト平易化は，入力テキストの意味を保持しつつ，語句や構造を変換して読者にとって理解しやすいテキストを生成する技術である．単語 [2]・文 [3]・文書 [4] の各単位を対象にテキスト平易化の技術が研究されているが，本研究では最も基礎的な単語単位の平易化（語彙平易化）[5] に取り組む．

テキスト平易化モデルを訓練する際には，難解なテキストと平易なテキストの対であるパラレルコーパスを使用する．文単位では約 50 万文対 [6]，文書単位では約 15 万文書対 [4] の大規模なパラレルコーパスを利用できる一方で，語彙平易化のためのパラレルコーパスは千件未満の小規模なもの [7–9] しか存在しない．これは，語彙平易化のためのパラレルコーパスの構築が，入力文からの難解語の検出，難解語に対する文脈を考慮した言い換え集合の生成，言い換え候補のリランキングなどの複数回のアノテーションを必要とする高コストなタスクのためと考えられる．そのため，語彙平易化の先行研究 [10–16] では，パラレルコーパスを必要とする教師あり学習の手法は提案されていない．

Much of the water carried by these streams is **diverted**.
⇩ 言い換え候補生成
used, redirected, diverted, channel, poll, less, …
⇩ リランキング
1. redirected, 2. diverted, 3. poll, 4. used, …
⇩
Much of the water carried by these streams is **redirected**.

図 1　語彙平易化

本研究では，文単位のテキスト平易化のためのパラレルコーパス [6, 17, 18] から少数の単語の置換のみによって平易化されている文対を抽出することで，語彙平易化のための擬似的なパラレルコーパスを構築する．そして，1 万件規模のパラレルコーパスを用いてマスク言語モデル [19] を再訓練し，教師あり学習のアプローチで語彙平易化を行う．3 種類の英語の評価用データセット [7–9] を用いた評価実験の結果，教師あり学習によって語彙平易化の性能を改善でき，最高性能を達成した．

2　関連研究

語彙平易化は，図 1 に示すように，1 単語が難解語として注釈付けされた入力文が与えられ，文脈を考慮して難解語の平易な言い換え候補を生成するステップと，言い換え候補を難易度に基づきリランキングするステップの 2 つのサブタスクからなる．本研究では，前者の言い換え候補の生成に取り組む．

2.1　言い換え候補の生成

言い換え候補の生成では，辞書に基づく手法と分散表現に基づく手法が提案されている．初期の研究 [10, 20] では，人手で整備された WordNet [21] などの辞書を用いて難解語の同義語を収集していた．その後，パラレルコーパス上での単語アライメント [22] によって大規模な言い換え辞書を自動構築

表 1　文単位のテキスト平易化のためのパラレルコーパスに含まれる文対の例

難解文	平易文	備考
The seat of the municipality was in Stavros.	The seat of the community was in Stavros.	語彙平易化の例（1 単語の変換）
Relax the current standards ?	Lower the current limits ?	語彙平易化の例（2 単語の変換）
A member of the Democratic Party, she is the first woman to serve as Governor of Rhode Island.	She is a member of the Democratic Party.	文長が異なる例
The underside of the wings is also black.	The underside of the wings is also black.	単語の置換がない例

する手法が考案され，文単位のテキスト平易化のためのパラレルコーパスを用いる手法 [7] や対訳コーパスを用いる手法 [2] が提案された．しかし，辞書に基づく手法は，自動構築であっても網羅性に限界があり，言い換えの文脈依存性にも対処が難しいため，現在では分散表現に基づく手法が主流である．

分散表現に基づく手法としては，GloVe [23] や fastText [24] などの単語分散表現の余弦類似度を用いて難解語の類義語を収集する先行研究 [13, 15] が多い．近年の手法 [16, 25] は，単語穴埋めの事前訓練を行ったマスク言語モデル BERT [19] を用い，文脈を考慮して言い換え候補を生成している．BERT は，文脈を考慮して言い換え候補を生成できる点と，単語穴埋めの事前訓練が語彙平易化の問題設定と近い点の 2 つの利点を持つ．そのため，本研究でも BERT を用いて言い換え候補の生成に取り組む．

2.2　言い換え候補のリランキング

このステップでは，難解語との同義性，文脈の中での流暢性，候補単語の平易性などの観点から，複数の言い換え候補をリランキングし，言い換え単語を選択する．初期の研究 [10, 20] では，候補単語の 1-gram 頻度を用いて，平易性の観点からリランキングを行っていた．その後，文脈に応じたランキングのために，難解語の周辺単語も考慮する 5-gram 頻度 [9] も使用されるようになった．さらに，より多様な観点からのリランキングのために，1-gram 頻度および 5-gram 頻度に加えて，難解語と候補単語の間の単語分散表現の余弦類似度（同義性）および候補単語と周辺単語の間の単語分散表現の余弦類似度（流暢性）も考慮する平均ランキング [13] が提案されている．近年の手法 [16, 25] では，文脈を考慮する 5-gram 頻度および周辺単語との類似度計算を，BERT の単語穴埋め確率に基づいて改良している．本研究では，言い換え候補の生成に取り組み，リランキングの適用やリランキングの改良による語彙平易化の性能改善については今後の課題とする．

3　提案手法

本研究では，文単位のテキスト平易化のための大規模なパラレルコーパス [6, 17, 18] の中から，単語単位の変換のみが行われている文対を抽出することで，語彙平易化のための擬似的なパラレルコーパスを構築する．まず，3.1 節では，語彙平易化のための擬似的なパラレルコーパスの構築手順について説明する．次に，3.2 節では，BERT [19] に基づく語彙平易化モデルの教師あり学習の方法について述べる．

3.1　擬似的なパラレルコーパスの構築

表 1 に例示するように，文単位のテキスト平易化のためのパラレルコーパスには，語句を置換する平易化，節を省略する平易化，文の構造を変更する平易化など，様々な変換をともなう文対が含まれる．その中で，少数の単語の置換のみが行われている文対は語彙平易化の事例だと考えられるため，このような文対を以下の手順で抽出する．ただし，難解文を C 個の単語からなる $X = x_1, .., x_C$，平易文を S 個の単語からなる $Y = y_1, ..., y_S$ と表記する．

1. 難解文と平易文の文長に差異がある場合，単語の置換以外の変換が行われている．そこで，文長が異なる（$C \neq S$ である）文対を除外する．
2. 文長が同じでも，多くの単語が異なる文対は語彙平易化の事例とは考えにくい．そこで，難解文と平易文の間で i 番目の単語同士を比較し，異なる単語（$x_i \neq y_i$）の個数を数える．これがゼロまたは k よりも大きい文対は除外する．ただし，i は $0 < i \leq C$ の整数である．

3.2　語彙平易化モデルの教師あり学習

先行研究 [16, 25] と同様に，入力文および難解語をマスクした入力文の 2 文を BERT [19] に入力し，マスク部分に対する単語穴埋めの要領で言い換え候補を生成する．2 単語を置換する表 1 の 2 文目の例では，具体的には “[CLS] relax the current standards

? [SEP] [MASK] the current [MASK] ? [SEP]” というトークン列を BERT へ入力する．ここで，[CLS] および [SEP] は BERT における特殊トークンであり，それぞれ文頭および文末を意味する．

先行研究 [16, 25] における言い換え候補生成では，生コーパス上で事前訓練のみを行った BERT を用いるが，本研究では事前訓練済みの BERT を 3.1 節の擬似的なパラレルコーパス上で再訓練してから用いる．先ほどの例において，我々の再訓練では，1 つめの [MASK] に対しては “lower” を，2 つめの [MASK] に対しては “limits” を，それぞれ正解単語として，事前訓練と同様に単語穴埋めのクロスエントロピー損失最小化の訓練を行う．このような再訓練を経た BERT は，語彙平易化における言い換え候補の生成に特化したモデルになると期待できる．

4 評価実験

提案手法の有効性および訓練ドメインの影響を検証するために，3 種類の英語の語彙平易化タスクにおいて評価実験を行う．

4.1 実験設定

データ 語彙平易化のための擬似的なパラレルコーパスは，文単位のテキスト平易化のための最大規模のパラレルコーパス[1]である Wiki-Auto [6, 17] および Newsela-Auto [6, 18] から抽出した．Wiki-Auto は Wikipedia ドメインの 488,332 文対であり，Newsela-Auto はニュースドメインの 394,300 文対である．前処理として，Moses Tokenizer[2] [26] を用いて単語分割および小文字化を行った．置換された単語数を表すハイパーパラメタ k は，1 から 5 までの 5 種類について実験した．各ドメインで抽出された擬似的なパラレルコーパスの規模を表 2 に示す．なお，$k = 1$ の文対中から 500 件ずつ合計 1,000 文対を無作為抽出して検証用セットとし，その他を訓練に用いた．

評価用には，英語の語彙平易化における代表的なコーパスである LexMTurk[3] [7]・BenchLS[4] [8]・NNSeval[5] [9] の 3 種類を使用した．LexMTurk は Wikipedia を平易化した 500 件，BenchLS は Wikipedia および英語の均衡コーパス [27–29] を平易化した 929 件，NNSeval は BenchLS から高品質なサブセットを抽出した 239 件である．

1) https://github.com/chaojiang06/wiki-auto
2) https://github.com/moses-smt/mosesdecoder/
3) https://cs.pomona.edu/~dkauchak/simplification/
4) http://ghpaetzold.github.io/data/BenchLS.zip
5) http://ghpaetzold.github.io/data/NNSeval.zip

表 2 語彙平易化の擬似的なパラレルコーパスの文対数

	Wiki-Auto (Wikipedia)	Newsela-Auto (News)
$k = 1$	10,260	7,681
$k = 2$	14,453	10,273
$k = 3$	16,446	11,614
$k = 4$	17,961	12,561
$k = 5$	18,995	11,379

モデル 公平な比較のために，モデルは先行研究 [16, 25] に合わせて bert-large-uncased-whole-word-masking[6] [19] を使用した．これは，1,024 次元の埋め込み層および隠れ層を持つ 24 層 12 注意ヘッドの事前訓練済み Transformer 符号化器 [30] である．再訓練には，バッチサイズを 16 文，学習率を $1e-6$，最適化手法を Adam [31] として，検証用データにおけるクロスエントロピー損失が 3 エポック改善されない場合に訓練を停止した．なお，表 2 における各ドメインのパラレルコーパスを用いて，訓練ドメインの異なる 2 種類のモデルを構築した．

4.2 評価方法

単語分散表現に基づく Light-LS [13] と，マスク言語モデルに基づく BERT-LS [16] および SimpleBERT [25] を提案手法と比較する．Light-LS は，単語分散表現 GloVe [23] の余弦類似度を用いて，難解語の類義語を言い換え候補として出力する．BERT-LS は，マスク言語モデル BERT [19] の単語穴埋め確率を用いて，文脈中で難解語の代替となる語を候補として出力する．SimpleBERT は，語彙平易化の仕組みは BERT-LS と同じだが，Simple English Wikipedia および Newsela の平易な英文 [32] を用いて BERT の追加事前訓練を行うことで，性能を改善している．

先行研究の評価の設定に従い，語彙平易化モデルによって出力された上位 10 件の候補単語を正解の言い換え集合と比較する．評価指標には，適合率および再現率と，その調和平均である F 値を用いる．

4.3 実験結果

表 3 に実験結果を，表 4 に各モデルの出力例を，それぞれ示す．擬似的なパラレルコーパスを用いた教師あり学習によって，提案手法は全てのタスクにおいて最高の F 値を達成した．

6) https://huggingface.co/bert-large-uncased-whole-word-masking

表 3 実験結果

	訓練ドメイン	LexMTurk			BenchLS			NNSeval		
		適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
Light-LS [13]	-	0.151	0.122	0.135	0.142	0.191	0.163	0.105	0.141	0.121
BERT-LS [16]	-	0.296	0.230	0.259	0.236	0.320	0.272	**0.190**	0.254	0.218
SimpleBERT [25]	-	0.353	0.275	0.309	**0.269**	0.364	0.309	-	-	-
提案手法（$k = 1$）	Wikipedia	**0.358**	**0.357**	**0.358**	0.254	**0.409**	**0.313**	**0.190**	**0.341**	**0.244**
提案手法（$k = 3$）	Wikipedia	0.352	0.353	0.353	0.251	0.403	0.309	0.189	0.326	0.239
提案手法（$k = 5$）	Wikipedia	0.348	0.350	0.349	0.248	0.403	0.307	0.186	0.327	0.237
提案手法（$k = 1$）	Newsela	0.270	0.277	0.274	0.205	0.349	0.259	0.168	0.330	0.223
提案手法（$k = 3$）	Newsela	0.258	0.264	0.261	0.193	0.329	0.243	0.164	0.313	0.216
提案手法（$k = 5$）	Newsela	0.245	0.253	0.249	0.184	0.317	0.233	0.150	0.299	0.200

表 4 出力例（下線は入力文に含まれる難解語，太字は正解と一致する出力）

入力文	The intent is to display the likeness, personality, and even the mood of the person.
正解単語	aim, meaning, idea, plan, goal, attempt, reason, point, purpose, desire, trotting
BERT-LS	intent, intention, **purpose**, objective, intended, **aim**, **desire**, object, emphasis, understanding
提案手法（$k = 1$）	**aim**, **goal**, **purpose**, **idea**, intention, object, objective, job, task, **plan**

訓練ドメインに関しては，Newsela で訓練したモデルよりも Wikipedia で訓練したモデルの方が高い性能を達成した．特に，Wikipedia のみを対象とする LexMTurk においてより大きな性能改善が見られたため，訓練用コーパスと評価用コーパスのドメインが一致することが重要であることがわかる．

4.4 擬似パラレルコーパスの人手評価

訓練データに含まれる置換単語数の最大値 k を増加させるほど訓練データの規模も増加するが，本実験では $k = 1$ のときに全てのタスクにおいて最高性能が得られた．つまり，単純に訓練データ数を増やすことが性能改善に寄与するわけではない．

$k \geq 2$ では，“Linee played predominantly at centre.” → “Linee played as a centre.” のように，複数単語を同時に置換する句の平易化が訓練データに含まれる．しかし，語彙平易化の実際の評価データには，1 単語のみを変換する事例しか含まれないため，これらの事例が訓練に悪影響を与えている可能性がある．

そこで，置換単語数ごとに，Wiki-Auto から 100 件ずつの訓練事例を無作為抽出し，適切な語彙平易化の事例であるか否かの人手評価を実施した．ここで，適切な語彙平易化の事例とは，i 番目の単語同士が全て置換可能な文対である．なお，置換単語が複数含まれる場合は，各単語がそれぞれ独立に置換できる場合を適切な事例とする．複数単語を同時に置換する事例や，並び替えによって i 番目の単語同士が対応しない事例は，不適切とする．

人手評価の結果，置換単語数 1 の事例では 69 件，置換単語数 3 の事例では 11 件，置換単語数 5 の事例では 1 件のみが適切な語彙平易化であった．これは，3.1 節の単純な規則では，置換単語数の増加とともにノイズ事例を多く収集してしまうことを意味する．そのため本実験では，訓練ノイズの少ない $k = 1$ の設定で最高性能が得られたと考えられる．

5 おわりに

本研究では，文単位の大規模なテキスト平易化のためのパラレルコーパスから語彙平易化のための擬似的なパラレルコーパスを自動抽出することで，初めて語彙平易化モデルのための教師あり学習を実現した．3 種類の英語の語彙平易化タスクにおける評価実験の結果，提案手法が全てのタスクにおいて最高の F 値を達成した．詳細な分析の結果，訓練データと評価データのドメインを一致させることの有効性や，ノイズの少ない 1 単語のみの置換の事例を用いて訓練することの有効性が明らかになった．

今後の課題として，単語類似度推定 [33] を用いた擬似データ収集の高品質化や，リランキングのサブタスクへの教師あり学習の適用に取り組みたい．

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP21H03564, JP22H00677）の助成を受けたものです．

参考文献

[1] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-Driven Sentence Simplification: Survey and Benchmark. **Computational Linguistics**, Vol. 46, No. 1, pp. 135–187, 2020.

[2] Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A Paraphrase Database for Simplification. In **Proc. of ACL**, pp. 143–148, 2016.

[3] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring Neural Text Simplification Models. In **Proc. of ACL**, pp. 85–91, 2017.

[4] Renliang Sun, Hanqi Jin, and Xiaojun Wan. Document-Level Text Simplification: Dataset, Criteria and Baseline. In **Proc. of EMNLP**, pp. 7997–8013, 2021.

[5] Gustavo Henrique Paetzold and Lucia Specia. A Survey on Lexical Simplification. **Journal of Artificial Intelligence Research**, Vol. 60, No. 1, pp. 549–593, 2017.

[6] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF Model for Sentence Alignment in Text Simplification. In **Proc. of ACL**, pp. 7943–7960, 2020.

[7] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a Lexical Simplifier Using Wikipedia. In **Proc. of ACL**, pp. 458–463, 2014.

[8] Gustavo Henrique Paetzold and Lucia Specia. Benchmarking Lexical Simplification Systems. In **Proc. of LREC**, pp. 3074–3080, 2016.

[9] Gustavo Henrique Paetzold and Lucia Specia. Unsupervised Lexical Simplification for Non-Native Speakers. In **Proc. of AAAI**, pp. 3761–3767, 2016.

[10] Siobhan Devlin and John Tait. The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers. **Linguistic Databases**, pp. 161–173, 1998.

[11] Or Biran, Samuel Brody, and Noemie Elhadad. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In **Proc. of ACL**, pp. 496–501, 2011.

[12] Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. Selecting Proper Lexical Paraphrase for Children. In **Proc. of ROCLING**, pp. 59–73, 2013.

[13] Goran Glavaš and Sanja Štajner. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In **Proc. of ACL**, pp. 63–68, 2015.

[14] Gustavo Henrique Paetzold and Lucia Specia. LEXenstein: A Framework for Lexical Simplification. In **Proc. of ACL**, pp. 85–90, 2015.

[15] Gustavo Henrique Paetzold and Lucia Specia. Lexical Simplification with Neural Ranking. In **Proc. of EACL**, pp. 34–40, 2017.

[16] Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. Lexical Simplification with Pretrained Encoders. In **Proc. of AAAI**, pp. 8649–8656, 2020.

[17] William Coster and David Kauchak. Simple English Wikipedia: A New Text Simplification Task. In **Proc. of ACL**, pp. 665–669, 2011.

[18] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. **TACL**, Vol. 3, pp. 283–297, 2015.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[20] Jan De Belder and Marie-Francine Moens. Text Simplification for Children. In **Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems**, pp. 19–26, 2010.

[21] George A. Miller. WordNet: A Lexical Database for English. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[22] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. **Computational Linguistics**, Vol. 29, No. 1, pp. 19–51, 2003.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In **Proc. of EMNLP**, pp. 1532–1543, 2014.

[24] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. **TACL**, Vol. 5, pp. 135–146, 2017.

[25] Renliang Sun and Xiaojun Wan. SimpleBERT: A Pre-trained Model That Learns to Generate Simple Words. **arXiv:2204.07779**, 2022.

[26] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In **Proc. of ACL**, pp. 177–180, 2007.

[27] Serge Sharoff. Open-source Corpora: Using the Net to Fish for Linguistic Data. **International Journal of Corpus Linguistics**, Vol. 11, No. 4, pp. 435–462, 2006.

[28] Diana McCarthy and Roberto Navigli. SemEval-2007 Task 10: English Lexical Substitution Task. In **Proc. of SemEval**, pp. 48–53, 2007.

[29] Jan De Belder and Marie-Francine Moens. A Dataset for the Evaluation of Lexical Simplification. In **Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing**, pp. 426–437, 2012.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.

[31] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proc. of ICLR**, 2015.

[32] Xingxing Zhang and Mirella Lapata. Sentence Simplification with Deep Reinforcement Learning. In **Proc. of EMNLP**, pp. 584–594, 2017.

[33] Yuki Arase and Tomoyuki Kajiwara. Distilling Word Meaning in Context from Pre-trained Language Models. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 534–546, 2021.