ニューラル単語アライメントに基づく言い換え知識獲得

近藤 里咲　　梶原 智之　　二宮 崇
愛媛大学大学院理工学研究科
{kondo@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp

概要

本研究では，ニューラル単語アライメントを用いて言い換え辞書の品質を改善する．大規模な言い換え知識獲得には，対訳コーパス上での単語アライメントに基づく Bilingual Pivoting と呼ばれる手法が用いられてきた．そのため，Bilingual Pivoting で得られる言い換えの品質は単語アライメントの品質に依存する．既存の言い換え辞書は，統計的な単語アライメントに基づく Bilingual Pivoting によって構築されており，単語の意味を考慮せずに言い換えを抽出しているため，獲得できる言い換えの品質に改善の余地がある．本研究では，ニューラル単語アライメントに基づく Bilingual Pivoting を用いて，より高品質な言い換え知識獲得に取り組む．評価実験の結果，本研究で構築した言い換え辞書が適合率と再現率の両方で既存の日本語言い換え辞書を上回った．

1　はじめに

言い換えは，情報検索 [1] や機械翻訳 [2] など，幅広い自然言語処理タスクに活用されている．深層学習が主流の近年においても，単語や文の表現学習 [3–5] をはじめとして，マスク言語モデルの事前訓練 [6] やファインチューニング [7]，語彙平易化 [8–10] などに，言い換え辞書が利用されている．

大規模な言い換え辞書の構築には，対訳コーパスに対して単語アライメントを適用する Bilingual Pivoting [11] という手法が一般的に用いられている．Bilingual Pivoting は，図 1 に示すように，対訳コーパス上での単語アライメントの後，他言語の共通語句 (first author) と対応付けられた語句同士（第一著者-筆頭著者）を言い換えとみなす手法である．そのため，得られる言い換えの品質は単語アライメントの品質に依存する．既存の言い換え辞書 [12–16] では，単語の意味を考慮できない統計的な単語アライメント [17] によって言い換えを獲得しているため，その品質には改善の余地がある．

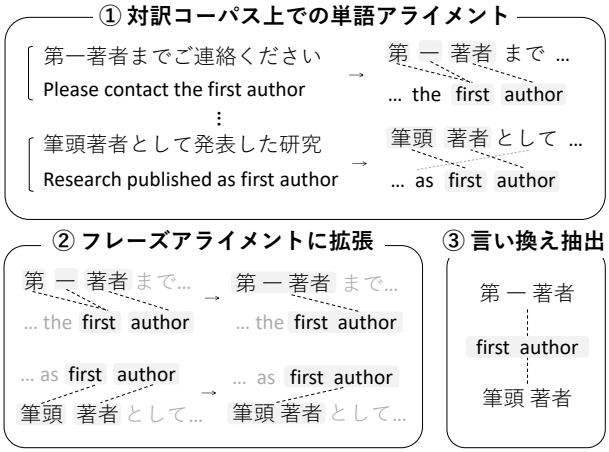

図 1　Bilingual Pivoting [11] による言い換え知識獲得

本研究では，日英対訳コーパスに対してニューラル単語アライメントに基づく Bilingual Pivoting を適用し，日本語言い換え辞書[1)]を構築する．後述のように，ニューラル単語アライメントは従来の統計的単語アライメントよりも高性能であるが，伝統的なフレーズ抽出ヒューリスティックとは相性が悪く，不当に多くのフレーズアライメントを獲得してしまう．この課題に対処するために，本研究ではフレーズやフレーズ対のフィルタリングを行い，高品質な言い換え対を得る．獲得した日本語言い換え辞書の品質を評価した結果，統計的単語アライメントに基づく手法や単純にニューラル単語アライメントを適用する手法と比較して，提案手法が再現率と適合率の両方において高性能であることを確認できた．

2　関連研究

2.1　Bilingual Pivoting と言い換え辞書

言い換え知識獲得の手法 Bilingual Pivoting [11] は，図 1 のように，対訳コーパス上での単語アライメントをフレーズアライメントに拡張し，他言語の共通語句と対応する対象言語の語句を言い換えとする．

1）https://github.com/EhimeNLP/EhiMerPPDB

言い換え抽出の際には，式 (1) のように，対象言語の語句 e_1 から他言語の語句 f への翻訳確率 $p(f|e_1)$ と f から対象言語の語句 e_2 への翻訳確率 $p(e_2|f)$ を f について周辺化し，言い換え確率 $p(e_2|e_1)$ を得る．

$$p(e_2|e_1) = \sum_f p(e_2|f)p(f|e_1) \quad (1)$$

Bilingual Pivoting によって構築された日本語の言い換え辞書として，PPDB: Japanese[2] [15] および EhiMerPPDB[1] [16] がある．前者は，5 つの日英対訳コーパス（合計約 200 万文対）に対して，統計的単語アライメントの GIZA++ [17] に基づく Bilingual Pivoting を適用し，最大 7 単語の長さの言い換え約 1,500 万対を収集している．後者は，約 2,000 万文対からなる日英対訳コーパス JParaCrawl[3] [18,19] に対して，同じく GIZA++を用いて最大 7 単語の長さの言い換え約 3.8 億対を獲得している．

他言語の先行研究 [12–14] も含めて，これらの言い換え辞書は，単語の共起頻度や相対位置に基づく統計的な情報のみを考慮する単語アライメント手法 [17] に依存するため，単語の意味情報や文脈情報が反映されていない．本研究では，マスク言語モデルに基づくニューラル単語アライメント手法を用いることで，高品質な言い換え知識獲得を目指す．

2.2 ニューラル単語アライメント

BERT [20] をはじめとする事前訓練済みマスク言語モデルの登場により，文脈化単語埋め込みから単語アライメントを得る教師なしニューラル単語アライメント [21,22] が研究されている．これらは教師アライメントやパラレルコーパスを必要とせずに優れた性能を示すため，近年注目を集めている．

最も単純なニューラル単語アライメント手法である SimAlign [21] は，mBERT[4] や XLM-R[5] [23] の多言語マスク言語モデルを用いて，文対の文脈化単語埋め込みから余弦類似度行列を導出し，類似度が最大となる単語対を抽出する．また，PMIAlign [22] は，ある単語が他の多くの単語と高い類似度を持つというハブネス問題を緩和するために，自己相互情報量を用いて SimAlign の類似度行列を改良している．本研究では，これらのニューラル単語アライメント手法を用いて，言い換え知識獲得を改善する．

2) https://ahcweb01.naist.jp/old/resource/jppdb/
3) https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
4) https://huggingface.co/google-bert/bert-base-multilingual-cased
5) https://huggingface.co/FacebookAI/xlm-roberta-base

表 1 単語アライメント性能（#Align は対応付けの総数）

	#Align	再現率	適合率	F 値
GIZA++	19,425	0.502	0.367	0.424
SimAlign (mBERT)	9,671	**0.553**	0.669	0.606
SimAlign (XLM-R)	8,930	0.475	0.614	0.535
PMIAlign (mBERT)	8,827	0.547	**0.721**	**0.622**
PMIAlign (XLM-R)	9,025	0.508	0.636	0.565

3 言い換え辞書の構築

本研究では，Bilingual Pivoting [11] に基づく日本語の言い換え辞書を高品質化するために，大規模な日英対訳コーパス JParaCrawl[3] [18,19] に対してニューラル単語アライメント [21,22] を適用する．

対訳コーパスの前処理 先行研究 [16] と同様に，英語は Moses Tokenizer [24]，日本語は MeCab (IPADIC) [25] を用いて単語分割した．さらに，空行や不要な空白，100 単語を超える文対を削除[6] した．

単語アライメント ニューラル単語アライメントに使用するマスク言語モデルを選択するために，予備実験として，日英対訳コーパス上での単語アライメント性能を評価した．本研究では，マスク言語モデルとして mBERT[4] または XLM-R[5] を使用した．表 1 に，KFTT[7] の評価用データにおける単語アライメントの性能を示す．どちらの手法においても mBERT の性能が高かったため，Bilingual Pivoting には mBERT ベースのニューラル単語アライメントを使用することとした．なお，GIZA++ [17] は，先行研究 [16] と同様に IBM model2 によって日英および英日の各方向の単語アライメントを得た後，grow-diag-final ヒューリスティックで双方向化した．

フレーズアライメントへの拡張 JParaCrawl に対して上記のモデルで単語アライメントをとり，先行研究 [16] と同様に Moses [24] のフレーズ抽出ヒューリスティックを適用して，単語アライメントをフレーズアライメントに拡張した．ここで，フレーズの最大長は，先行研究 [15,16] に従い 7 単語とした．

フレーズのフィルタリング ここまでに得た対訳フレーズ対のうち，記号を含むまたは英数字のみで構成される日本語フレーズはノイズとなるため，それらのフレーズを含む対訳フレーズ対を除去した．その後，日英・英日の両方向の翻訳確率を算出した．

6) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl
7) https://www.phontron.com/kftt/

表 2　言い換え辞書の概要と品質評価

	データ	アライメント手法	言い換え数	全体評価			ドメイン別の再現率		
				再現率	適合率	F 値	ニュース	観光	医療
[15]	200 万文対	GIZA++	15M	0.172	0.379	0.236	0.183	0.224	0.119
[16]	JParaCrawl	GIZA++	387M	**0.209**	0.439	0.283	0.229	**0.252**	**0.147**
本研究	JParaCrawl	SimAlign	37,231M	0.186	0.344	0.241	0.204	0.215	0.138
本研究	JParaCrawl	PMIAlign	40,210M	0.191	0.380	0.255	0.212	0.220	0.140
本研究	JParaCrawl	GIZA++× SimAlign	69M	0.207	0.450	0.284	0.227	0.246	**0.147**
本研究	JParaCrawl	GIZA++× PMIAlign	71M	**0.209**	**0.466**	**0.288**	**0.233**	0.243	0.142

さらに，連接確率の低いフレーズはノイズになりやすいため，word2vec [26] に含まれるフレーズ抽出ツール[8]を用いて，このようなフレーズを除去した．具体的には，JParaCrawl の日本語側と英語側のそれぞれに対して，フレーズ化とスコア計算を 4 回実施し，それぞれ 200，100，50，25 を下回るスコアを持つフレーズをノイズと定義して，それらのフレーズを含む対訳フレーズ対を除去した．

言い換えの獲得　対訳フレーズ対とその翻訳確率が得られたため，式 (1) に従い，言い換え対とその言い換え確率を求めた．SimAlign では約 370 億対，PMIAlign では約 400 億対の言い換えが得られた．

言い換え確率のリランキング　単語単位の言い換え知識獲得 [27] において，統計ベース手法と埋め込みベース手法の組み合わせの有効性が報告されている．統計ベース手法と埋め込みベース手法は異なる観点から言い換えを評価するため，これらの組み合わせによって互いの欠点を補い合い、より高品質な言い換えが得られると期待できる．そこで本研究でも，GIZA++の単語アライメントに基づく言い換え確率とニューラル単語アライメントに基づく言い換え確率を掛け合わせ，言い換え候補をリランキングした．このリランキングは両手法で得られる言い換え候補の積集合に対して実施されるため，この後処理によって言い換えは約 7,000 万対に制限された．

4　評価実験

再現率の自動評価と適合率の人手評価によって言い換え辞書の品質を評価する．既存の日本語言い換え辞書として，PPDB: Japanese[2] [15] および EhiMerPPDB[1] [16] を本研究と比較する．

4.1　実験設定

データセット　先行研究 [16] と同じ評価データを用いた．この評価データは，ニュースドメインの JADES[9] [28]，観光ドメインの MATCHA[10] [29]，医療ドメインの JASMINE[11] [30] の 3 つの言い換えパラレルコーパスから抽出されたものである．JADES から 895 件，MATCHA から 452 件，JASMINE から 429 件の計 1,776 件の語句の言い換えが含まれる．

再現率の評価　言い換え辞書が実世界の言い換えをどれだけ網羅できているかを自動評価した．各語句に対しての 10best 言い換えが集められた PPDB: Japanese と比較するために，提案手法においても各語句に対して言い換え確率の上位 10 件を用いた．

適合率の評価　言い換え辞書のノイズの少なさを人手評価する．評価データに含まれる語句のうち，日本語母語話者の大学生 3 名で構成される評価者がよく知っている語句のうち，無作為抽出された 100 件を本評価に使用した．これら 100 件の語句に対する 10best 言い換えを，先行研究 [16] と同様に，以下の 4 段階で評価した．

1. 意味的に等価ではない
2. 意味的に等価ではあるものの，置換はできない
3. 文脈によっては置換できる
4. 常に置換できる

このうち，1 または 2 と評価された言い換え対を負例，3 または 4 と評価された言い換え対を正例と定義し，評価者の多数決によって適合率を算出した．なお，評価者間の一致率を重み付き Kappa 係数で評価した結果，0.56〜0.72 と充分な合意が確認できた．

8）https://github.com/tmikolov/word2vec/blob/master/word2phrase.c

9）https://github.com/naist-nlp/jades

10）https://github.com/EhimeNLP/matcha

11）https://github.com/EhimeNLP/JASMINE

表 3　例：「不気味な」に対する言い換え上位 10 件

GIZA++		PMIAlign		GIZA++× PMIAlign	
言い換え	言い換え確率	言い換え	言い換え確率	言い換え	言い換え確率
不気味	0.2903	不気味	0.0584	不気味	0.0169
恐ろしい	0.0196	不吉な	0.0050	恐ろしい	8.0944e-05
キモ	0.0152	恐ろしい	0.0041	な	3.4555e-05
な	0.0150	不思議	0.0033	奇妙な	3.2621e-05
奇妙な	0.0114	ような	0.0031	不吉な	2.8480e-05
creepy 気味悪い	0.0087	奇妙な	0.0029	キモ	2.4925e-05
気味悪い	0.0084	奇妙	0.0028	気味悪い	2.3102e-05
で不気味	0.0080	不思議な	0.0028	奇妙	2.1925e-05
奇妙	0.0079	気味悪い	0.0027	不思議な	1.4978e-05
おかしい	0.0063	な	0.0023	おかしい	1.2504e-05

4.2　実験結果

ニューラル単語アライメントが言い換え獲得の高品質化に貢献　表 2 に，再現率および適合率の評価結果と，それらに基づき算出した F 値を示す．全ての評価指標において，統計ベース手法と埋め込みベース手法の組み合わせ（GIZA++× PMIAlign）が最高性能を達成した．本手法は，統計ベースの既存手法 [16] と比較して言い換え数を 5 分の 1 以下に抑えつつも，再現率を維持して適合率を改善できた．表 3 に示す例からも，統計ベース手法で上位にあった “不気味な-creepy 気味悪い” や埋め込みベース手法で上位にあった “不気味な-ような” などの言い換えノイズが，GIZA++× PMIAlign の組み合わせ手法によって上位から外れていることを確認できた．

SimAlign よりも PMIAlign の方が高性能　ニューラル単語アライメント手法の SimAlign と PMIAlign を比較すると，それぞれ単体においても統計ベース手法との組み合わせにおいても，PMIAlign の方が一貫して高い性能を示した．表 1 でも見たように，単語アライメントの性能は PMIAlign の方が高いため，高性能な単語アライメント手法を用いることで，Bilingual Pivoting による言い換え知識獲得の品質も高くなることが示唆される．

フレーズ抽出には課題が残る　一方で，ニューラル単語アライメントは単体では GIZA++の統計的単語アライメントに再現率と適合率の両方で劣っている．これは，本研究でも使用した伝統的な Moses [24] のフレーズ抽出ヒューリスティックがニューラル単語アライメントとは相性が悪いことが原因であると考えられる．表 1 に示したように，ニューラル手法の単語アライメント性能は高いものの，抽出する単語対の数は統計的単語アライメントの半分以下である．フレーズ抽出ヒューリスティックは，このような大量かつ雑音の多い統計的単語アライメント手法のために設計されていることに注意されたい．そのため，傾向の異なるニューラル単語アライメント手法と組み合わせた際には，表 2 に示したように，不当に多くの言い換えフレーズ対を抽出することにつながってしまう．本研究ではフレーズのフィルタリングに取り組んだが，ニューラル単語アライメントに適したフレーズ抽出手法の改良は，今後の課題として残されている．

ドメインごとに性能差がある　表 2 の右に，評価データのドメイン別の再現率を示す．全ての言い換え辞書が，観光ドメインに対する性能が最も高く，医療ドメインに対する性能が最も低かった．ニュースおよび観光のドメインにおいては 20〜25%の再現率を達成する一方，医療ドメインにおいては 15%以下の再現率に留まった．特定のドメインに特化した言い換え知識獲得への拡張も，今後の課題である．

5　おわりに

本研究では，大規模な日英対訳コーパスに対してニューラル単語アライメントに基づく Bilingual Pivoting を適用し，約 7,000 万対からなる日本語の言い換え辞書[1)]を構築した．実験の結果，統計的手法とニューラル手法を組み合わせて言い換え確率を求める提案手法が，先行研究よりも高品質な言い換え知識獲得を実現できた．今後の課題には，ニューラル単語アライメントに適したフレーズ抽出手法の改良および言い換え知識獲得のドメイン適応がある．

謝辞

本研究は，株式会社メルカリ R4D の支援を受けて実施した．

参考文献

[1] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In **Proc. of SIGIR**, p. 266–272, 2004.

[2] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved Statistical Machine Translation Using Paraphrases. In **Proc. of NAACL**, pp. 17–24, 2006.

[3] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. In **Proc. of NAACL**, pp. 1606–1615, 2015.

[4] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Charagram: Embedding Words and Sentences via Character n-grams. In **Proc. of EMNLP**, pp. 1504–1515, 2016.

[5] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards Universal Paraphrastic Sentence Embeddings. In **Proc. of ICLR**, 2016.

[6] Renliang Sun, Wei Xu, and Xiaojun Wan. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In **Findings of ACL**, pp. 9345–9355, 2023.

[7] Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification. In **Proc. of TSAR**, pp. 147–153, 2022.

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

[9] Daiki Nishihara and Tomoyuki Kajiwara. Word Complexity Estimation for Japanese Lexical Simplification. In **Proc. of LREC**, pp. 3114–3120, 2020.

[10] Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. LSBert: Lexical Simplification Based on BERT. **TASLP**, Vol. 29, pp. 3064–3076, 2021.

[11] Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In **Proc. of ACL**, pp. 597–604, 2005.

[12] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In **Proc. of NAACL**, pp. 758–764, 2013.

[13] Juri Ganitkevitch and Chris Callison-Burch. The Multilingual Paraphrase Database. In **Proc. of LREC**, pp. 4276–4283, 2014.

[14] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In **Proc. of ACL**, pp. 425–430, 2015.

[15] Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Building a Free, General-domain Paraphrase Database for Japanese. In **Proc. of COCOSDA**, pp. 1–4, 2014.

[16] 近藤里咲, 梶原智之, 二宮崇. JParaCrawl からの大規模日本語言い換え辞書の構築. 言語処理学会第 30 回年次大会, pp. 1736–1740, 2024.

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

[18] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In **Proc. of LREC**, pp. 3603–3609, 2020.

[19] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In **Proc. of LREC**, pp. 6704–6710, 2022.

[20] 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.

[21] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In **Findings of ACL**, pp. 1627–1643, 2020.

[22] Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. PMI-Align: Word Alignment With Point-Wise Mutual Information Without Requiring Parallel Training Data. In **Findings of ACL**, pp. 12366–12377, 2023.

[23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **Proc. of ACL**, pp. 8440–8451, 2020.

[24] 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.

[25] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proc. of EMNLP**, pp. 230–237, 2004.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In **Proc. of ICLR Workshop**, 2013.

[27] Tomoyuki Kajiwara, Mamoru Komachi, and Daichi Mochihashi. MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting. In **Proc. of IJCNLP**, pp. 80–89, 2017.

[28] Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In **Proc. of TSAR**, pp. 179–187, 2022.

[29] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. MATCHA：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. 自然言語処理, Vol. 31, No. 2, pp. 590–609, 2024.

[30] Koki Horiguchi, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Evaluation Dataset for Japanese Medical Text Simplification. In **Proc. of NAACL-SRW**, pp. 219–225, 2024.