JParaCrawl からの大規模日本語言い換え辞書の構築

近藤 里咲 [1]　梶原 智之 [2]　二宮 崇 [2]
[1] 愛媛大学工学部　[2] 愛媛大学大学院理工学研究科
kondo@ai.cs.ehime-u.ac.jp　{kajiwara, ninomiya}@cs.ehime-u.ac.jp

概要

本研究では，先行研究よりも 25 倍大きい 3.8 億対の日本語言い換え辞書を構築し，公開する．言い換え知識獲得は，Bilingual Pivoting と呼ばれる対訳コーパス上での単語アライメントによって行われてきた．本手法で獲得できる表現の多様性は対訳コーパスの規模に依存するが，既存の日本語言い換え辞書は 200 万文対の日英対訳コーパスから得られた 1,500 万言い換え対が最大である．我々は，10 倍大きい 2,000 万文対の日英対訳コーパス JParaCrawl を用いて，より多様な言い換え知識獲得に取り組む．評価実験の結果，我々の言い換え辞書は再現率と適合率の両方で先行研究を上回り，文類似度推定の外的評価においてもより高い性能を達成した．

1　はじめに

言い換えは情報検索 [1] や機械翻訳 [2] など，多くの自然言語処理タスクの性能改善に貢献する．近年では，単語や文の表現学習 [3–5] や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（PPDB: **P**ara**p**hrase **D**atab**a**se） [11–13] が用いられている．

PPDB などの大規模な言い換え知識獲得には，Bilingual Pivoting [14] と呼ばれる対訳コーパス上での単語アライメントの手法が用いられる．これは，図 1 に示すように，他言語の共通の語句（author）と対応付けられる対象言語の語句同士（執筆者-著者）を言い換えと見なす，言い換え知識獲得の手法のひとつである．そのため，Bilingual Pivoting によって獲得できる言い換え表現の多様性は，対訳コーパスの規模に依存する．既存の日本語 PPDB [15] は，合計約 200 万文対の日英対訳コーパスから Bilingual Pivoting によって約 1,500 万件の言い換え対を収集している．しかし，3.1 節で示すように，既存の日本語 PPDB では実世界の言い換えの約 18%しか網羅できず，言い換え辞書の規模に課題が残っている．

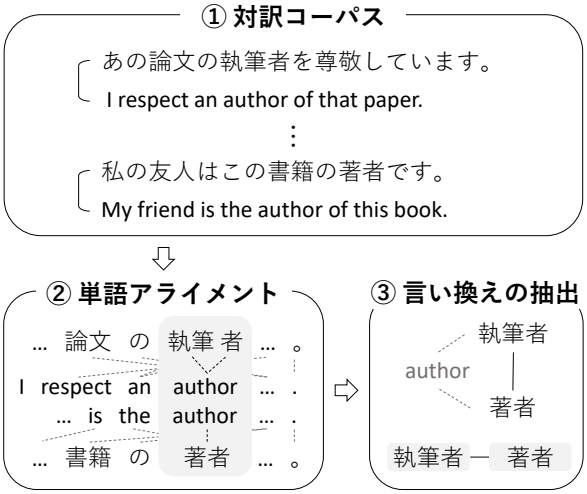

図 1　Bilingual Pivoting [14] による言い換え知識獲得の例

本研究では，2,000 万文対を超える最大規模の日英対訳コーパスである JParaCrawl [16, 17] に対して Bilingual Pivoting を適用し，より大規模な約 3.8 億件の言い換え対からなる日本語 PPDB を構築および公開[1)]する．言い換え辞書の品質評価の結果，再現率に関する自動評価および適合率に関する人手評価の両方で，既存の日本語 PPDB [15] を上回り，大規模かつ高品質な言い換え辞書を構築できたことを確認した．さらに，文類似度推定に関する外的評価においても，既存の日本語 PPDB を用いるよりも高い性能を達成し，我々の辞書の有用性を確認できた．

2　対訳コーパスからの言い換え獲得

2.1　Bilingual Pivoting

Bilingual Pivoting [14] は言い換え知識獲得の手法のひとつであり，PPDB [11–13] のような大規模な言い換え辞書を構築する際に採用されてきた．基本的なアイデア（図 1）は，他言語の共通の語句（author）と対応付けられる対象言語の語句同士（執筆者-著者）を言い換えと考えるというものである．

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

表 1 「執筆 者」に対する言い換えの例

言い換え	言い換え確率	ピボット単語
著者	0.398	author, authors, authorship, the author, the authors, the writer, writer, writers, ...
作者	0.082	author, authors who, authors, authorship, the author, the authors, the writer, writer, ...
作家	0.070	an author, author whose, author, authors who, authors, authorship, the author, writer, ...
著者 は	0.028	author, authors, of the authors, the author, the authors, the writer, writers, ...
著者 ら	0.019	authors, authorship, the authors
筆者	0.017	author, authors, writer, writers
ライター	0.012	author, authors, writer, writers
筆頭 著者	0.011	author, lead author

具体的にはまず，対象言語の語句 e およびピボット言語の語句 f の系列からなる文対で構成される対訳コーパスがあるときに，単語アライメントの技術を用いて対象言語とピボット言語の間で語句の対応をとる．そして，単語アライメント確率 $p(f|e)$ および $p(e|f)$ を計算し，対象言語の 2 つの語句 e_1 および e_2 が共有する全てのピボット言語の語句で単語アライメント確率を周辺化し，以下のように e_1 から e_2 への言い換え確率を近似して求める．

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

2.2 先行研究：既存の日本語 PPDB

既存の日本語 PPDB[2] [15] は，田中コーパス[3]や KFTT[4]など，数十万文対の規模の日英対訳コーパス 5 つ（合計約 200 万文対）に対して Bilingual Pivoting を適用し，約 1,500 万対の日本語の言い換えを得ている．単語分割などの前処理の後で GIZA++ [18] による単語アライメントを実施し，最大 7 単語の長さのフレーズ単位で式 (1) を計算し，言い換え確率が 1%以上の言い換え対を抽出している．

なお，Bilingual Pivoting 以外の方法で構築された日本語の言い換え辞書も存在するが，本研究では触れない．例えば，シソーラスの兄弟関係にある単語対から人手で同義語対を抽出した日本語 WordNet 同義対データベース[5]や，人手で内容語に言い換えを付与した SNOW D2: 内容語換言辞書[6]などがあるが，いずれも数万語の規模の辞書であり，Bilingual Pivoting に基づく PPDB と比べて遥かに小さい．

2) https://ahcweb01.naist.jp/old/resource/jppdb/
3) https://www.edrdg.org/wiki/index.php/Tanaka_Corpus
4) https://www.phontron.com/kftt/
5) https://bond-lab.github.io/wnja/jpn/index.html
6) https://www.jnlp.org/GengoHouse/snow/d2

表 2 言い換え確率の閾値ごとの言い換え対数

	S	M	L	XL	ALL
閾値	0.2	0.1	0.05	0.01	10^{-5}
先行研究	2.7M	4.3M	6.0M	11.7M	15.0M
本研究	4.9M	8.6M	13.4M	30.4M	386.9M

2.3 より大規模な言い換え辞書の構築

本研究では，より大規模な日本語の言い換え辞書を構築するために，最大規模の日英対訳コーパスである JParaCrawl[7] [16, 17] に対して Bilingual Pivoting を適用する．なお，ピボット言語は英語である．

前処理 対訳コーパスの英語側は MosesTokenizer [19]，日本語側は MeCab（IPADIC） [20] で単語分割した．そして，空行や空白文字，100 単語を超える長文などを削除する前処理[8]を実施した．

Bilingual Pivoting の適用 単語アライメントには GIZA++（IBM model 2） [18] を用い，grow-diag-final のヒューリスティックによって対称化した．そして，Moses [19] のフレーズ抽出のヒューリスティックを用いて，日英翻訳のフレーズテーブルを得た．ここで，先行研究 [15] と同様に，フレーズの最大長は 7 単語とした．このフレーズテーブルを用いて式 (1) を計算し，言い換え対を抽出した．

後処理 不要な言い換え対やノイズとなる表現を除外するための後処理を実施した．$e_1 = e_2$ である言い換え対は不要であるため除外し，記号を含む語句や英数字のみで構成される語句を含む言い換え対はノイズとなるため除外した．最終的に，約 3.8 億件の言い換え対を得た．表 1 に，本研究で収集した言い換えの例を示す．また，表 2 に，言い換え確率

7) https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
8) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

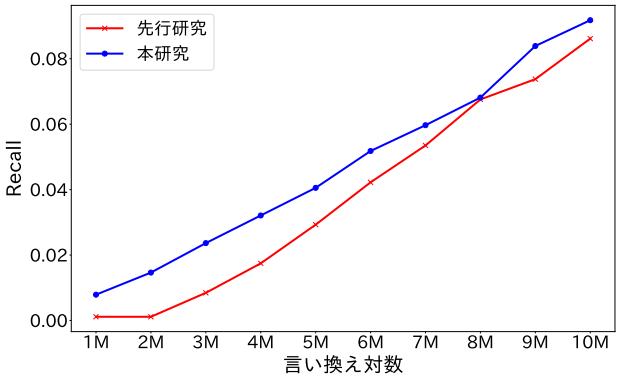

図 2　言い換えの再現率に関する自動評価

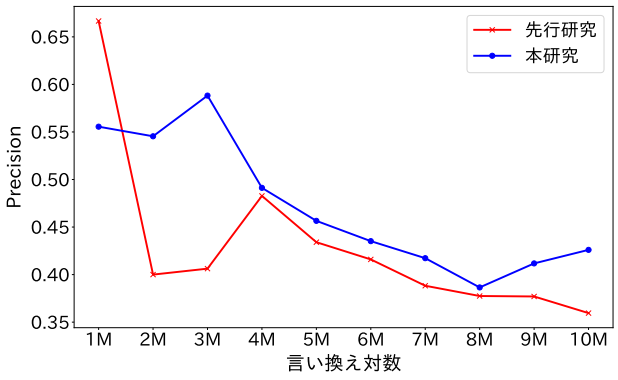

図 3　言い換えの適合率に関する人手評価

の閾値ごとの言い換え対数について，既存の日本語 PPDB [15] と我々の PPDB の比較を示す．[9)]

3　評価実験

言い換え辞書の品質および有用性を評価するために，言い換えの再現率および適合率に関する内的評価および文類似度推定に関する外的評価を行う．

3.1　実験 1：再現率の自動評価

実験設定　後述の評価用データセットを用いて，言い換え辞書が実世界の言い換えをどれだけ網羅できるかを自動評価する．既存の日本語 PPDB [15] と我々が構築した PPDB から，それぞれ言い換え確率の上位 k 件ずつを抽出し，再現率を自動評価する．

データセット　言い換えの再現率を自動評価するために，人間が行う自然な言い換えを収集した評価用データセットを作成する．そこで，専門家によって人手で構築された日本語の文単位の言い換えパラレルコーパスから，語句の言い換えを抽出する．特定ドメインへの偏りを避けるために，本研究では，JADES[10)]（ニュース）[21]，MATCHA[11)]（観光），JASMINE[12)]（医療）の 3 つの言い換えパラレルコーパスを対象にする．[13)]

各コーパスから 200 文対ずつを無作為抽出し，日本語を母語とする大学生 3 名が語句の言い換えを人手で抽出した．なお，2.3 節と同じく MeCab [20] で単語分割し，意味的に対応する最小の語句を抽出した．また，2.3 節と同じ後処理を適用し，7 単語以下の長さの言い換えを抽出した．その結果，JADES から 895 件，MATCHA から 452 件，JASMINE から 429 件の合計 1,776 件の語句の言い換えを得た．

実験結果　図 2 に，上位 100 万件から 1,000 万件まで言い換え対の数を変化させた際の再現率の評価結果を示す．両辞書とも上位 1,000 万件まで見ても実世界の言い換えの 10%も網羅できないものの，我々の PPDB は既存の日本語 PPDB よりも一貫して高い再現率を示した．また，辞書全体を用いた場合，先行研究の再現率が 18%であるのに対して，本研究では 51%と顕著に高い再現率を達成できた．実世界の多様な言い換えを網羅することは依然として難しいものの，本研究では実世界の言い換えのうち半数を収録した言い換え辞書を構築でき，先行研究を網羅性の点で大きく上回ることができた．

3.2　実験 2：適合率の人手評価

実験設定　言い換え辞書に含まれるノイズ（非言い換え対）の少なさを人手評価する．既存の日本語 PPDB [15] と我々が構築した PPDB から，それぞれ言い換え確率の上位 k 件ずつを抽出し，適合率を人手評価する．ただし，人手評価のコスト軽減のために，両辞書（表 2 の XL サイズ）に共通して掲載されている語句の中から 100 種類を無作為抽出[14)]し，それらの語句に対する言い換え対を全て評価する．なお，上位 1,000 万件の設定で，先行研究では 331 対，本研究では 223 対が評価対象となった．

アノテーション　日本語を母語とする大学生 3 名が，以下の 4 段階で言い換え対を評価した．

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，置換できない

9)　先行研究には，$e_1 = e_2$ の対や英数字のみで構成される語句を含む対など，我々の後処理で除外した言い換え対も含まれることに注意されたい．また，先行研究の ALL は，各語句に対して言い換え確率の上位 10 件ずつを抽出したものであり，閾値によって調整されていないことにも注意されたい．

10)　https://github.com/naist-nlp/jades

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

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

13)　クラウドソーシングの SNOW [22] や自動構築された TMUP [23]，英語から機械翻訳された PAWS-X [24] は対象外

14)　評価者が理解できない語句は対象外として選び直した

表 3　文類似度推定データセットの文対数

	訓練用	検証用	評価用
JSTS	12,451	1,457	-
JSICK	4,500	-	4,927

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 が付与された言い換え対を負例，3 または 4 が付与された言い換え対を正例と定義する．評価者の多数決によって評価対象の各言い換え対を評価し，適合率を求める．なお，評価者間の 4 段階評価の一致率は，重み付き Kaapa 係数 [25] で 0.62 から 0.71 であり，充分な合意が見られた．

実験結果　図 3 に，上位 100 万件から 1,000 万件まで言い換え対の数を変化させた際の適合率の評価結果を示す．上位 100 万件の設定を除いて，我々の PPDB が既存の日本語 PPDB よりも高い適合率を示した．なお，上位 100 万件の設定では，両辞書とも評価対象の言い換え対が 9 件のみであり，正例数の差は 1 件であった．また，平均適合率を求めると，先行研究は 43%，本研究は 47%であり，より高い適合率を達成できた．これらの結果から，本研究では言い換え対の規模だけでなく，高品質な言い換えの割合の点でも先行研究を改善できたと言える．

3.3　実験 3：文類似度推定の外的評価

実験設定　言い換え辞書の有用性を検証するために，英語の先行研究 [26] と同様に文類似度推定タスクにおける外的評価を行う．本タスクは 2 文間の意味的な類似度を推定するタスクであり，本研究では JSTS[15] [27] および JSICK[16] [28] のデータセットを用いて人手評価とのピアソン相関を評価する．なお，JSTS は $[0, 5]$，JSICK は $[1, 5]$ の範囲で類似度の人手評価が付与されている．表 3 に各データセットの規模を示す．JSTS は評価用データが公開されていないため，本実験では検証用データで評価する．

既存の日本語 PPDB [15] と我々が構築した PPDB から，それぞれ言い換え確率の上位 k 件ずつを抽出し，文類似度推定に用いる．ただし，応用タスクの性能を最大化するためには，再現率と適合率のバランスが重要である．そこで，各辞書（表 2 の XL サイズ）においてハイパーパラメータ k を 100 万ずつ変化させ，訓練用データ上での最適値を探索する．

15）https://github.com/yahoojapan/JGLUE
16）https://github.com/verypluming/JSICK

表 4　外的評価：文類似度推定のピアソン相関

	JSTS	JSICK
先行研究	0.651	0.768
本研究	**0.699**	**0.770**

言い換え辞書に基づく文類似度推定　辞書ベースの教師なし文類似度推定手法である DLS@CU [29] に従い，言い換え辞書に基づく単語アライメントを行い，対応付けられた単語の割合に応じて式 (2) のように文類似度を推定する．

$$\mathrm{sim}(s_1, s_2) = \frac{a(s_1) + a(s_2)}{n(s_1) + n(s_2)} \tag{2}$$

ここで，$n(s)$ は文 s の単語数，$a(s)$ は文 s において単語アライメントがとられた単語数を表している．なお，言い換え辞書とは独立に，両文に共通して出現する単語は強制的に対応付けた．

実験結果　まず，訓練用データ上でのハイパーパラメータ探索の結果，JSTS においては先行研究で上位 1,000 万件，本研究で上位 3,000 万件の言い換えを用いた場合に最高性能を示した．一方，JSICK においては先行研究で上位 500 万件，本研究で上位 1,000 万件が最適であった．これらの結果から，相対的な特性として，JSTS は言い換えの再現率，JSICK は言い換えの適合率を重視することが示唆される．

表 4 に，JSTS の検証用データおよび JSICK の評価用データにおける文類似度推定の実験結果を示す．両データセットにおいて，既存の日本語 PPDB を用いるよりも我々の PPDB を用いる方が人手評価との高い相関を達成した．これらの結果から，本研究では文類似度推定タスクのために有用な言い換えを収集できていることが確認できた．

4　おわりに

本研究では，現時点で最大規模の日英対訳コーパスである JParaCrawl [16, 17] に対して Bilingual Pivoting [14] を適用し，現時点で最大規模の約 3.8 億件の日本語言い換え辞書[1]を構築および公開した．評価の結果，我々の辞書は再現率と適合率の両方で既存の言い換え辞書 [15] を上回り，高品質かつ網羅性の高い言い換え知識獲得ができたことを確認した．また，文類似度推定の外的評価からも，既存の辞書に対する我々の辞書の有用性を確認できた．

今後の課題として，表 1 の「執筆者-著者は」などの文法的に不適切な言い換え対の除去やニューラル単語アライメント手法 [30] の適用に取り組みたい．

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 the Association for Computational Linguistics: ACL 2023**, 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] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In **Proc. of NAACL**, pp. 758–764, 2013.

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

[13] 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.

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

[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] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

[22] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In **Proc. of LREC**, pp. 461–466, 2018.

[23] Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. Building a Non-Trivial Paraphrase Corpus Using Multiple Machine Translation Systems. In **Proc. of ACL-SRW**, pp. 36–42, 2017.

[24] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In **Proc. of EMNLP**, pp. 3687–3692, 2019.

[25] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. **Psychological Bulletin**, Vol. 70, No. 4, pp. 213–220, 1968.

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

[27] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In **Proc. of LREC**, pp. 2957–2966, 2022.

[28] Hitomi Yanaka and Koji Mineshima. Compositional Evaluation on Japanese Textual Entailment and Similarity. **TACL**, Vol. 10, pp. 1266–1284, 2022.

[29] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In **Proc. of SemEval**, pp. 148–153, 2015.

[30] Zi-Yi Dou and Graham Neubig. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In **Proc. of EACL**, pp. 2112–2128, 2021.