Paraphrase Acquisition via Bilingual Pivoting Based on Neural Word Alignment

Risa Kondo[†] **Seiji Sugiyama**[†] **Tomoyuki Kajiwara**[†‡] **Takashi Ninomiya**[†]
[†] Graduate School of Science and Engineering, Ehime University, Japan
[‡] D3 Center, The University of Osaka, Japan
{kondo@ai.cs., sugiyama@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp

Abstract

We utilize neural word alignment to improve the quality of paraphrase databases in English and Japanese. For large-scale paraphrase acquisition, previous studies have employed a framework of bilingual pivoting based on word alignment on bilingual parallel corpora. Naturally, the quality of paraphrases acquired by bilingual pivoting depends on the performance of word alignment. Previous studies based on statistical word alignment have limitations in the quality of acquired paraphrases because they do not consider word meaning. This study employs a more sophisticated neural approach for word alignment in bilingual pivoting to enhance the quality of paraphrase acquisition. Experimental results revealed that our paraphrase databases outperformed existing ones in both internal and external evaluations.

Keywords: Paraphrase Acquisition, Bilingual Pivoting, Neural Word Alignment

1. Introduction

Paraphrase databases have been utilized for various natural language processing tasks, including information retrieval (Liu et al., 2004) and machine translation (Callison-Burch et al., 2006). Even in recent years, when deep learning dominates, paraphrase databases are utilized for tasks such as representation learning (Faruqui et al., 2015; Wieting et al., 2016a,b; Sugiyama et al., 2025), pre-training (Sun et al., 2023) and fine-tuning (Zetsu et al., 2022) of masked language models, and lexical simplification (Pavlick and Callison-Burch, 2016; Nishihara and Kajiwara, 2020; Qiang et al., 2021).

For large-scale paraphrase acquisition, a framework of bilingual pivoting (Bannard and Callison-Burch, 2005) based on word alignment on bilingual parallel corpora has been widely employed. As shown in Figure 1, bilingual pivoting first estimates word alignments on a given parallel corpus. It then extracts phrase pairs in the target language ("first author" and "lead author") corresponding to common phrases in the pivot language ("筆頭著者"), treating them as paraphrases. Naturally, the quality of paraphrases acquired by bilingual pivoting depends on the performance of word alignment. Existing paraphrase databases (Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014; Pavlick et al., 2015; Mizukami et al., 2014) are based on statistical word alignment (Och and Ney, 2003) that disregards word meaning, resulting in limited paraphrase quality.

In this study, we employ a more sophisticated neural approach for word alignment in bilingual pivoting to construct higher-quality paraphrase databases.[1] However, while neural word alignment outperforms statistical word alignment, it is incompatible with traditional phrase extraction heuristics designed for statistical word alignment, generating too many phrase alignments in step 2 of Figure 1. To address this challenge, we perform both filtering of phrase pairs and reranking of paraphrase pairs, as shown in Figure 2, to acquire high-quality paraphrases. Experimental results on internal evaluation in Japanese and external evaluation in both Japanese and English reveal the usefulness of our paraphrase databases. We will release[1] both 70 million paraphrase pairs in Japanese and 40 million paraphrase pairs in English.

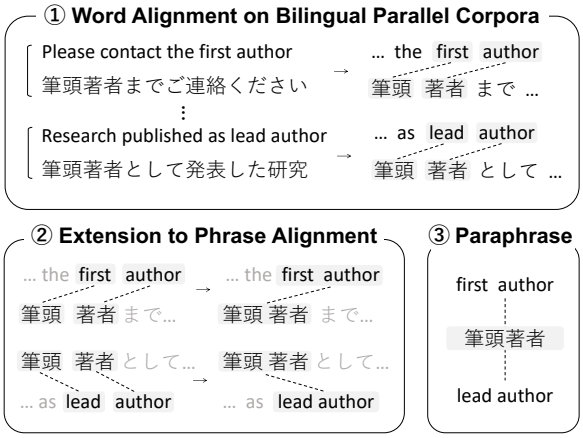

Figure 1: Paraphrase acquisition via bilingual pivoting.

[1] https://github.com/EhimeNLP/EhiMerPPDB

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 160–168

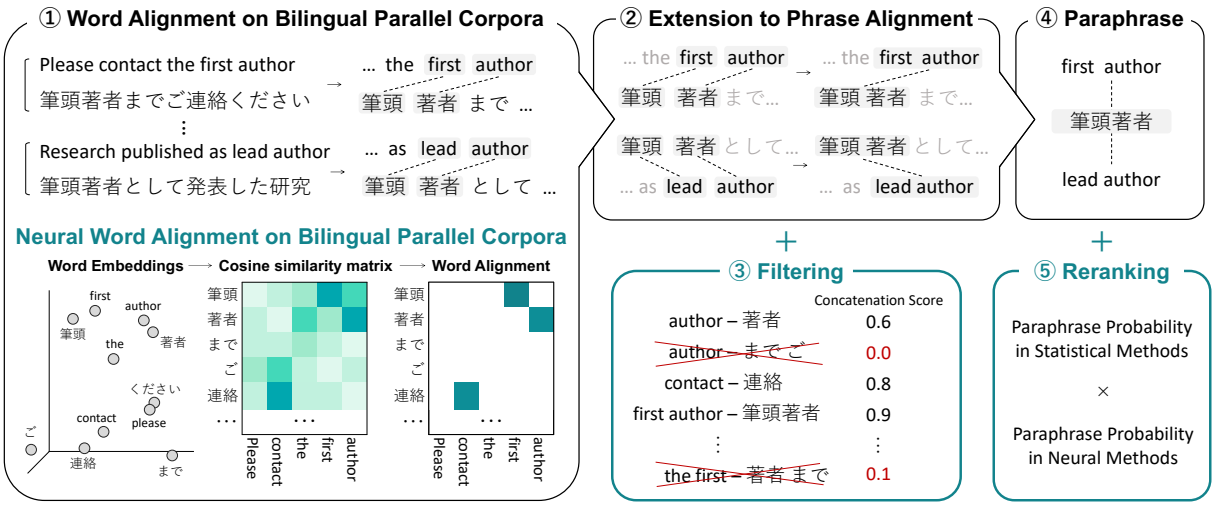

Figure 2: Overview of our proposed method. For bilingual pivoting, we employ a neural approach for word alignment. When extending this to phrase alignment, we reduce noise by filtering phrase pairs. Finally, we combine the paraphrase probability based on statistical word alignment with that of neural word alignment to rerank the paraphrases.

2. Related Work

2.1. Bilingual Pivoting

As shown in Figure 1, bilingual pivoting (Bannard and Callison-Burch, 2005) performs word alignment on bilingual parallel corpora and extends it to phrase alignment to acquire paraphrases. Paraphrases acquired in this way are phrase pairs in the target language that correspond to common phrases in the pivot language. As shown in Equation (1), its paraphrase probability $p(e_2|e_1)$ is estimated by marginalizing over f the translation probability $p(f|e_1)$ from phrase e_1 in the target language to phrase f in the pivot language and the translation probability $p(e_2|f)$ from f to phrase e_2 in the target language.

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

Bilingual pivoting has been employed in the construction of large-scale language resources such as PPDB (**p**ara**p**hrase **d**ata**b**ase) in English (Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014; Pavlick et al., 2015) and its Japanese versions. One of the previous studies in Japanese, JPPDB[2] (Mizukami et al., 2014), combined several small-scale Japanese-English parallel corpora and acquired paraphrases using bilingual pivoting based on GIZA++ (Och and Ney, 2003), a statistical word alignment. Another previous study, EhiMerPPDB,[3] utilized JParaCrawl[4] (Morishita et al., 2020, 2022), a large-scale Japanese-English parallel corpus, and similarly acquired paraphrases via bilingual pivoting based on GIZA++.

Since these previous studies rely on statistical word alignment (Och and Ney, 2003), they do not reflect information about word meaning and context. In contrast, we employ neural word alignment methods based on masked language models to enhance the quality of paraphrase acquisition.

2.2. Neural Word Alignment

Inspired by pre-trained masked language models such as BERT (Devlin et al., 2019), neural word alignment methods (Jalili Sabet et al., 2020; Azadi et al., 2023) that estimate word alignments from contextual word embeddings are being actively studied. These methods achieve high performance despite not requiring gold alignments for training.

A pioneer in neural word alignment, SimAlign (Jalili Sabet et al., 2020), employs multilingual masked language models such as mBERT[5] (Devlin et al., 2019) and XLM-R[6] (Conneau et al., 2020) to perform word alignment that maximizes word similarity based on contextual word embeddings for sentence pairs. PMIAlign (Azadi et al., 2023)

[2] http://ahcweb01.naist.jp/old/resource/jppdb/

[3] https://github.com/EhimeNLP/EhiMerPPDB

[4] https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

[6] https://huggingface.co/FacebookAI/xlm-roberta-base

improves similarity calculations in SimAlign with pointwise mutual information.

In this study, we employ these neural word alignment methods to improve paraphrase acquisition.

3. Paraphrase Acquisition

In this study, we apply bilingual pivoting (Bannard and Callison-Burch, 2005) based on neural word alignment (Jalili Sabet et al., 2020; Azadi et al., 2023) to a Japanese-English parallel corpus to enhance the quality of paraphrase databases in English and Japanese.

Preprocessing for Parallel Corpus As with the EhiMerPPDB,[3] we also used JParaCrawl (Morishita et al., 2020, 2022), the largest Japanese-English parallel corpus. For tokenization, we used Moses Tokenizer (Koehn et al., 2007) for English and MeCab with IPADIC (Kudo et al., 2004) for Japanese. We then filtered out[7] sentence pairs containing blank lines, unnecessary spaces, or longer sentences exceeding 100 words.

Word Alignment We employed SimAlign (Jalili Sabet et al., 2020) and PMIAlign (Azadi et al., 2023) for neural word alignment. These are based on multilingual masked language models, and we employ either mBERT[5] (Devlin et al., 2019) or XLM-R[6] (Conneau et al., 2020). To select a masked language model suitable for neural word alignment, we conducted a preliminary experiment comparing word alignment performance on the KFTT[8] Japanese-English parallel corpus. Evaluation results in Table 1 revealed the effectiveness of mBERT in neural word alignment. Therefore, we employ mBERT-based neural word alignment for paraphrase acquisition.

Extension to Phrase Alignment To extend the word-level alignments obtained on JParaCrawl to phrase-level alignments, we employed the phrase extraction heuristic, which corresponds to Step 5 of the Moses toolkit (Koehn et al., 2007). This process extracts candidate phrases by identifying contiguous sequences of words—including those not explicitly aligned—that are consistent with the word alignments produced in the previous step. Following a previous study (Mizukami et al., 2014), the maximum length of phrases was set to 7 tokens.

Filtering Phrases To ensure the quality of the paraphrase database, we filtered out phrase pairs containing noise. Specifically, we removed phrases that consisted solely of symbols (e.g., punctuation marks or mathematical signs) or did not contain any lexical characters from the target language, as such pairs are unlikely to be meaningful paraphrases. Furthermore, since phrases with low co-occurrence probabilities are often noisy, we filtered them using the phrase extraction tool[9] included in word2vec (Mikolov et al., 2013). This tool scores how likely words are to appear consecutively. To systematically eliminate unreliable candidates, we performed this scoring and phrase formation process iteratively four times for both languages, removing phrases with scores below 200, 100, 50, and 25 in each respective step.

	#Align	Recall	Precision	F1
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

Table 1: Preliminary experiment: performance of word alignment on the KFTT evaluation dataset.

Reranking Paraphrases Having obtained the phrase pairs and their translation probabilities, we calculated the paraphrase probability using Equation (1). The phrase pairs and their paraphrase probabilities will be added to our paraphrase database. Here, previous study on lexical paraphrase acquisition (Kajiwara et al., 2017) has reported the effectiveness of combining statistical-based and embedding-based methods. Also in this study, the paraphrase probability based on statistical word alignment and the paraphrase probability based on neural word alignment are complementary, and we expect to acquire high-quality paraphrases by combining them. To obtain high-quality paraphrases, we multiplied the paraphrase probabilities of both methods. Since this multiplication takes the intersection of the paraphrase pairs acquired by each method, it reduces the total number of candidates while prioritizing those identified as reliable by both statistical and neural alignments. Here, statistical word alignment using GIZA++ (Och and Ney, 2003) was bidirectionalized with a grow-diag-final heuristic after calculating word alignments for each language direction based on the IBM model 2. Table 2 shows the quantities of paraphrase pairs we acquired.

[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

[8] https://www.phontron.com/kftt/

[9] https://github.com/tmikolov/word2vec/blob/master/word2phrase.c

		Number of paraphrases
English	PPDB 2.0 (XXXL)	56M
	EhiMerPPDB	251M
	SimAlign	28,176M
	PMIAlign	28,274M
	SimAlign x GIZA++	37M
	PMIAlign x GIZA++	38M
Japanese	JPPDB (10best)	15M
	EhiMerPPDB	387M
	SimAlign	37,231M
	PMIAlign	40,210M
	SimAlign x GIZA++	69M
	PMIAlign x GIZA++	71M

Table 2: Number of paraphrase pairs.

4. Experiments

We experimentally evaluate the quality and usefulness of paraphrase databases both internally, using recall and precision, and externally, on sentence pair modeling tasks. For comparison, we use PPDB 2.0 (Pavlick et al., 2015) in English, JPPDB (Mizukami et al., 2014) in Japanese, and EhiMerPPDB[3] in both languages. EhiMerPPDB is a paraphrase database constructed from JParaCrawl (Morishita et al., 2020, 2022) like ours, but based on statistical word alignment unlike ours.

4.1. Internal Evaluation

In this section, we evaluate the quality of paraphrase databases using recall (automatic evaluation) and precision (human evaluation). However, since PPDB 2.0 is currently private and inaccessible, we report only the experimental results in Japanese.

4.1.1. Setting

Dataset We constructed an evaluation dataset by extracting phrasal paraphrase pairs from a manually built Japanese paraphrase corpus created by experts. To avoid bias toward specific domains and ensure high-quality annotations, we employed three corpora created by different groups of experts: JADES[10] (Hayakawa et al., 2022) for news domains created by a researcher in Japanese text simplification, MATCHA[11] (Miyata et al., 2024) for tourism domains created by Japanese language teachers, and JASMINE[12] (Horiguchi et al., 2024) for medical domains created by annotators with expertise in Japanese medical NLP.[13]

We first randomly selected 200 sentence pairs from each corpus. These pairs were then distributed among three university students, all native Japanese speakers, who manually extracted phrasal paraphrases. The annotators were instructed to identify the minimum units that correspond as paraphrases among the segments tokenized by MeCab (Kudo et al., 2004). As a preprocessing step, similar to Section 3, we removed noise such as phrases containing symbols or those not containing Japanese. Finally, we collected a total of 1,776 Japanese phrasal paraphrases: 895 from JADES, 452 from MATCHA, and 429 from JASMINE.

Recall We automatically evaluated the coverage of paraphrase databases. For a fair comparison with JPPDB, our paraphrase database also evaluated the top 10 entries with the highest paraphrase probability for each query.

Precision We manually evaluated the quality of paraphrase databases. Three university students who are native Japanese speakers evaluated the top 10 paraphrases for 100 randomly selected phrases using the following four-point scale.

1. Not semantically equivalent.
2. Semantically equivalent but not substitutable.
3. Substitutable depending on context.
4. Always substitutable.

We defined paraphrases rated as 1 or 2 as negative examples and those rated as 3 or 4 as positive examples, then calculated the precision based on the majority vote of the evaluators. Here, the inter-annotator agreement was evaluated using the weighted kappa coefficient, showing substantial agreement ranging from 0.56 to 0.72.

4.1.2. Result

Table 3 shows the experimental results. Across all evaluation metrics, the proposed method combining statistical word alignment and neural word alignment (PMIAlign x GIZA++) consistently achieved the highest performance. Neural word alignments collect an extremely large number of paraphrases, and may contain a lot of noise when used on its own. The combination of statistical word alignment and neural word alignment is promising for removing such noise, resulting in the acquisition of a high-quality paraphrase database.

[10] https://github.com/naist-nlp/jades
[11] https://github.com/EhimeNLP/matcha
[12] https://github.com/EhimeNLP/JASMINE

[13] We excluded corpora created via crowdsourcing platforms such as SNOW (Katsuta and Yamamoto, 2018) and corpora generated by machine translation systems such as TMUP (Suzuki et al., 2017).

	Alignment	#Paraphrase	Recall	Precision	F1 Score
JPPDB[2]	GIZA++	15M	0.172	0.379	0.236
EhiMerPPDB[3]	GIZA++	387M	**0.209**	0.439	0.283
Ours	SimAlign	37,231M	0.186	0.344	0.241
Ours	PMIAlign	40,210M	0.191	0.380	0.255
Ours	SimAlign x GIZA++	69M	0.207	0.450	0.284
Ours	PMIAlign x GIZA++	71M	**0.209**	**0.466**	**0.288**

Table 3: Internal evaluation of Japanese paraphrase databases.

	English			Japanese		
Rank	GIZA++	PMIAlign	GIZA++×PMIAlign	GIZA++	PMIAlign	GIZA++×PMIAlign
1	spooky	eerie	spooky	不気味	不気味	不気味
2	eerie	spooky	eerie	恐ろしい	不吉な	恐ろしい
3	uncanny	an eerie	uncanny	キモ	恐ろしい	な
4	weird	uncanny	weird	な	不思議	奇妙な
5	eerily	weird	an eerie	奇妙な	ような	不吉な
6	scary	eerily	eerily	creepy 気味悪い	奇妙な	キモ
7	macabre	ominous	scary	気味悪い	奇妙	気味悪い
8	disgusting	sinister	macabre	で不気味	不思議な	奇妙
9	kimonos	scary	strange	奇妙	気味悪い	不思議な
10	strange	bad	disgusting	おかしい	な	おかしい

Table 4: Top 10 paraphrases for the expression "creepy" / "不気味な"

	Train	Dev	Test
Shopping Queries	1,254,438	138,625	425,762
STS-B	5,749	1,500	1,379
SICK	4,439	495	4,906
SNLI	549,367	9,842	9,824
PAWS	49,401	8,000	8,000

Table 5: Number of sentence pairs in English.

	Train	Dev	Test
Shopping Queries	294,874	32,272	118,907
JSTS	11,205	1,246	1,457
JSICK	4,500	500	4,927
JNLI	18,065	2,008	2,434
PAWS-X	49,401	2,000	2,000

Table 6: Number of sentence pairs in Japanese.

Among the neural word alignment methods, PMIAlign consistently outperformed SimAlign. As shown in Table 1, PMIAlign demonstrates higher performance in word alignment. These consistent results suggest that employing high-performance word alignment methods can improve the quality of paraphrase acquisition based on bilingual pivoting.

4.2. Qualitative Evaluation

Table 4 shows examples of paraphrases obtained for the word "creepy" (不気味な) using each method. It can be observed that for English, noisy paraphrases such as "kimonos", semantically unrelated to "creepy", were successfully removed from the top results by incorporating the neural method. Furthermore, for Japanese, noisy expressions such as "creepy 気味悪い" (creepy creepy) in the statistical method and "ような" (like/as if) in the neural method were effectively eliminated from the top rankings by combining GIZA++ and PMIAlign.

4.3. External Evaluation

In this section, we evaluate the usefulness of paraphrase databases by applying them to sentence pair modeling tasks. Following previous work (Sugiyama et al., 2025), we apply paraphrase-based contrastive learning, where paraphrases serve as positive examples and other sentences within the mini-batch as negative examples, to fine-tune pre-trained sentence encoders for sentence-pair modeling tasks.

4.3.1. Setting

Following previous work (Sugiyama et al., 2025), we evaluate the effectiveness of paraphrase-based contrastive learning across four tasks: product retrieval, sentence similarity estimation, recognizing textual entailment, and paraphrase identification (Tables 5 and 6). For pre-trained sentence encoders, as in previous work (Sugiyama et al., 2025),

English	Retrieval	Similarity		Entailment		Paraphrase	
	Shopping Queries	STS-B	SICK	SNLI	SICK	PAWS	Avg.
w/o Paraphrasing	0.654	0.824	0.815	**0.904**	0.858	0.913	0.828
PPDB 2.0	**0.655**	0.841	0.842	**0.904**	**0.866**	0.918	0.838
EhiMerPPDB[3]	**0.655**	**0.860**	0.829	**0.904**	0.858	**0.928**	0.839
Ours (PMIAlign)	**0.655**	0.859	**0.844**	0.900	0.860	0.925	**0.841**

Japanese	Retrieval	Similarity		Entailment		Paraphrase	
	Shopping Queries	JSTS	JSICK	JNLI	JSICK	PAWS-X	Avg.
w/o Paraphrasing	0.576	0.859	0.890	0.785	0.839	0.793	0.790
JPPDB[2]	0.586	0.857	0.896	0.824	0.854	0.795	0.802
EhiMerPPDB[3]	0.587	**0.861**	0.896	**0.828**	**0.856**	0.791	0.803
Ours (PMIAlign)	**0.588**	0.860	**0.901**	0.820	**0.856**	**0.803**	**0.805**

Table 7: External evaluation of paraphrase databases in sentence pair modeling tasks. The scores for "w/o Paraphrasing" and "PPDB 2.0" are taken from (Sugiyama et al., 2025).

we also used English BERT[14] (Devlin et al., 2019) and Japanese RoBERTa[15] (Liu et al., 2019).

Retrieval Product retrieval is a four-class classification task of the relationships between product titles and their search queries, and we employed both English and Japanese versions of the Shopping Queries dataset[16] (Reddy et al., 2022).

Similarity Sentence similarity estimation is a regression task that estimates the semantic similarity between two sentences, and we employed datasets of STS-B[17] (Cer et al., 2017) and SICK[18] (Marelli et al., 2014) for English and JSTS[19] (Kurihara et al., 2022) and JSICK[20] (Yanaka and Mineshima, 2022) for Japanese.

Entailment Recognizing textual entailment is a three-class classification task of semantic relationships between two sentences, and we employed datasets of SNLI[21] (Bowman et al., 2015) and SICK for English and JNLI[19] (Kurihara et al., 2022) and JSICK for Japanese.

Paraphrase Paraphrase identification is a two-class classification task of synonymity between two sentences, and we employed datasets of PAWS[22] (Zhang et al., 2019) for English and PAWS-X[22] (Yang et al., 2019) for Japanese.

For evaluation metrics, we employed the same metrics as previous work (Sugiyama et al., 2025): the micro-F1 score for the retrieval task, Spearman's rank correlation coefficient for the similarity task, and the macro-F1 score for both the entailment and paraphrase tasks. All other experimental details also followed (Sugiyama et al., 2025).

4.3.2. Result

Table 7 shows the experimental results. In both English and Japanese settings, paraphrase-based contrastive learning with our paraphrase database achieved the highest average performance. These experimental results suggest that we can acquire paraphrases that are effectively utilized to train better sentence encoders through contrastive learning.

5. Conclusion

In this study, we applied bilingual pivoting to a large-scale Japanese-English parallel corpus and constructed a Japanese paraphrase database containing 70 million pairs and an English paraphrase database containing 40 million pairs.[1] We successfully improved the quality of paraphrase acquisition by replacing the word alignment method for bilingual pivoting from a statistical approach to a neural approach. Although neural word alignment and traditional phrase extraction heuristics are incompatible, we achieved high-quality paraphrase acquisition by filtering and reranking noisy phrases.

[14] https://huggingface.co/google-bert/bert-base-uncased
[15] https://huggingface.co/rinna/japanese-roberta-base
[16] https://github.com/amazon-science/esci-data
[17] http://ixa2.si.ehu.eus/stswiki/stswiki/index.php/Special:Random.html
[18] https://zenodo.org/records/2787612
[19] https://github.com/yahoojapan/JGLUE
[20] https://github.com/verypluming/JSICK
[21] https://nlp.stanford.edu/projects/snli/
[22] https://github.com/google-research-datasets/paws

6. Bibliographical References

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2023. PMI-Align: Word Alignment With Point-Wise Mutual Information Without Requiring Parallel Training Data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12366–12377.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4276–4283.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. 2022. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*, pages 179–187.

Koki Horiguchi, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2024. Evaluation Dataset for Japanese Medical Text Simplification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 219–225.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Tomoyuki Kajiwara, Mamoru Komachi, and Daichi Mochihashi. 2017. MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 80–89.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. 2004. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 266–272.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 216–223.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations*.

Rina Miyata, Hyuga Koretaka, Hiroki Yamauchi, Daiki Yanamoto, Tomoyuki Kajiwara, Takashi Ninomiya, and Yasuhiro Nishiwaki. 2024. MATCHA: Parallel Corpus for Japanese Text Simplification Based on Professionally Simplified Articles. *Journal of Natural Language Processing*, 31(2):590–609.

Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Building a Free, General-domain Paraphrase Database for Japanese. In *Proceedings of the 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques*, pages 1–4.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609.

Daiki Nishihara and Tomoyuki Kajiwara. 2020. Word Complexity Estimation for Japanese Lexical Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3114–3120.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 143–148.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. LSBert: Lexical Simplification Based on BERT. *Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. *arXiv:2206.06588*.

Seiji Sugiyama, Risa Kondo, Tomoyuki Kajiwara, and Takashi Ninomiya. 2025. Paraphrase-based Contrastive Learning for Sentence Pair Modeling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 400–407.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In *Findings of the*

Association for Computational Linguistics: ACL 2023, pages 9345–9355.

Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Building a Non-Trivial Paraphrase Corpus Using Multiple Machine Translation Systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages 36–42.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016a. Charagram: Embedding Words and Sentences via Character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016b. Towards Universal Paraphrastic Sentence Embeddings. In *Proceedings of the 4th International Conference on Learning Representations*.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional Evaluation on Japanese Textual Entailment and Similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3687–3692.

Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*, pages 147–153.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308.