Building a Non-Trivial Paraphrase Corpus using Multiple Machine Translation Systems

Yui Suzuki **Tomoyuki Kajiwara** **Mamoru Komachi**
Graduate School of System Design, Tokyo Metropolitan University, Tokyo, Japan
{suzuki-yui, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

We propose a novel sentential paraphrase acquisition method. To build a well-balanced corpus for Paraphrase Identification, we especially focus on acquiring both non-trivial positive and negative instances. We use multiple machine translation systems to generate positive candidates and a monolingual corpus to extract negative candidates. To collect non-trivial instances, the candidates are uniformly sampled by word overlap rate. Finally, annotators judge whether the candidates are either positive or negative. Using this method, we built and released the first evaluation corpus for Japanese paraphrase identification, which comprises 655 sentence pairs.

1 Introduction

When two sentences share the same meaning but are written using different expressions, they are deemed to be a sentential paraphrase pair. Paraphrase Identification (PI) is a task that recognizes whether a pair of sentences is a paraphrase. PI is useful in many applications such as information retrieval (Wang et al., 2013) or question answering (Fader et al., 2013).

Despite this usefulness, there are only a few corpora that can be used to develop and evaluate PI systems. Moreover, such corpora are unavailable in many languages other than English. This is because manual paraphrase generation tends to cost a lot. Furthermore, unlike a bilingual parallel corpus for machine translation, a monolingual parallel corpus for PI cannot be spontaneously built.

Even though some paraphrase corpora are available, there are some limitations on them. For example, the Microsoft Research Paraphrase Corpus

Figure 1: Overview of candidate pair generation.

(MSRP) (Dolan and Brockett, 2005) is a standardized corpus in English for the PI task. However, as Rus et al. (2014) pointed out, MSRP collects candidate pairs using short edit distance, but this approach is limited to collecting positive instances with a low word overlap rate (WOR) (*non-trivial positive instances*, hereafter)[1]. In contrast, the Twitter Paraphrase Corpus (TPC) (Xu et al., 2014) comprises short noisy user-generated texts; hence, it is difficult to acquire negative instances with a high WOR (*non-trivial negative instances*, hereafter)[2].

To develop a more robust PI model, it is important to collect both *"non-trivial"* positive and negative instances for the evaluation corpus. To create a useful evaluation corpus, we propose a novel paraphrase acquisition method that has two viewpoints of balancing the corpus: positive/negative and trivial/non-trivial. To balance between positive and negative, our method has a machine translation part collecting mainly positive instances and a random extraction part collecting negative instances. In the machine translation part, we generate candidate sentence pairs using multiple machine translation systems. In the random extraction part, we extract candidate sentence pairs from a monolingual corpus. To collect both trivial and non-trivial instances, we sample candidate pairs

[1] *Non-trivial positive instances* are difficult to identify as semantically equivalent.

[2] *Non-trivial negative instances* are difficult to identify as semantically inequivalent.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 36–42
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3007

using WOR. Finally, annotators judge whether the candidate pairs are paraphrases.

In this paper, we focus on the Japanese PI task and build a monolingual parallel corpus for its evaluation as there is no Japanese sentential paraphrase corpus available. As Figure 1 shows, we use phrase-based machine translation (PBMT) and neural machine translation (NMT) to generate two different Japanese sentences from one English sentence. We expect the two systems provide widely different translations with regard to surface form such as lexical variation and word order difference because they are known to have different characteristics (Bentivogli et al., 2016); for instance, PBMT produces more literal translations, whereas NMT produces more fluent translations.

We believe that when the translation succeeds, the two Japanese sentences have the same meaning but different expressions, which is a positive instance. On the other hand, translated candidates can be negative instances when they include fluent mistranslations. This occurs since adequacy is not checked during an annotation phase. Thus, we can also acquire some negative instances in this manner.

To actively acquire negative instances, we use Wikipedia to randomly extract sentences. In general, it is rare for sentences to become paraphrase when sentence pairs are collected randomly, so it is effective to acquire negative instances in this regard.

Our contributions are summarized as follows:

- Generated paraphrases using multiple machine translation systems for the first time
- Adjusted for a balance from two viewpoints: positive/negative and trivial/non-trivial
- Released[3] the first evaluation corpus for the Japanese PI task

2 Related Work

Paraphrase acquisition has been actively studied. For instance, paraphrases have been acquired from monolingual comparable corpora such as news articles regarding the same event (Shinyama et al., 2002) and multiple definitions of the same concept (Hashimoto et al., 2011). Although these methods effectively acquire paraphrases, there are not many domains that have comparable corpora. In contrast, our method can generate paraphrase candidates from any sentences, and this allows us to choose any domain required by an application.

Methods using a bilingual parallel corpus are similar to our method. In fact, our method is an extension of previous studies that acquire paraphrases using manual translations of the same documents (Barzilay and McKeown, 2001; Pang et al., 2003). However, it is expensive to manually translate sentences to create large numbers of translation pairs. Thus, we propose a method that inexpensively generates translations using machine translation and Quality Estimation.

Ganitkevitch et al. (2013) and Pavlick et al. (2015) also use a bilingual parallel corpora to build a paraphrase database using bilingual pivoting (Bannard and Callison-Burch, 2005). Their methods differ from ours in that they aim to acquire phrase level paraphrase rules and carry out word alignment instead of machine translation.

There are also many studies on building a large scale corpora utilizing crowdsourcing in related tasks such as Recognizing Textual Entailment (RTE) (Marelli et al., 2014; Bowman et al., 2015) and Lexical Simplification (De Belder and Moens, 2012; Xu et al., 2016). Moreover, there are studies collecting paraphrases from captions to videos (Chen and Dolan, 2011) and images (Chen et al., 2015). One advantage of leveraging crowdsourcing is that annotation is done inexpensively, but it requires careful task design to gather valid data from non-expert annotators. In our study, we collect sentential paraphrase pairs, but we presume that it is difficult for non-expert annotators to provide well-balanced sentential paraphrase pairs, unlike lexical simplification, which only replaces content words. For this reason, annotators classify paraphrase candidate pairs in our study similar to the method used in the TPC and previous studies on RTE.

As for Japanese, there exists a paraphrase database (Mizukami et al., 2014) and an evaluation dataset that includes some paraphrases for lexical simplification (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016). They provide either lexical or phrase-level paraphrases, but we focus on collecting sentence-level paraphrases for PI evaluation. There is also an evaluation dataset for RTE (Watanabe et al., 2013) containing 70 sentential paraphrase pairs; however, as there is a limitation in terms of size, we aim to build a larger corpus.

[3] https://github.com/tmu-nlp/paraphrase-corpus

Jaccard	# Sentence	Average Sentence Length			# Sample	# Positive	# Negative	# Unnatural	# Other
		source	PBMT	NMT					
$[0.0, 0.1)$	228	19.42	20.65	19.75	200	2	1 (0)	80	117
$[0.1, 0.2)$	2,117	21.56	24.81	22.01	200	11	14 (0)	147	28
$[0.2, 0.3)$	14,080	21.56	26.50	23.37	200	20	9 (0)	162	9
$[0.3, 0.4)$	51,316	23.48	29.69	26.29	200	24	15 (0)	161	0
$[0.4, 0.5)$	100,674	24.40	31.35	28.08	200	27	16 (0)	151	6
$[0.5, 0.6)$	134,101	23.16	29.90	27.26	200	34	16 (0)	142	8
$[0.6, 0.7)$	100,745	21.04	27.32	25.30	200	38	13 (0)	129	20
$[0.7, 0.8)$	55,610	18.83	24.57	23.04	200	53	12 (40)	131	4
$[0.8, 0.9)$	26,884	16.23	21.31	20.24	200	81	3 (80)	94	22
$[0.9, 1.0)$	8,071	13.79	18.40	17.55	200	73	3 (70)	56	68
$[1.0, 1.0]$	6,174	10.10	13.07	12.96	0	0	0 (0)	0	0
Total	500,000	19.42	24.32	22.35	2,000	363	102 (190)	1,253	282

Table 1: Statistics on our corpus. The number inside ( ) of Negative column is the number of instances extracted from Wikipedia and the other is that of machine-translated instances.

3 Candidate Generation

3.1 Paraphrase Generation using Multiple Machine Translation Systems

We use different types of machine translation systems (PBMT and NMT) to translate source sentences extracted from a monolingual corpus into a target language. This means that each source sentence has two versions in the target language, and we use the sentences as a pair.

To avoid collecting ungrammatical sentences as much as possible, we use Quality Estimation and eliminate inappropriate sentences for paraphrase candidate pairs. At WMT2016 (Bojar et al., 2016) in the Shared Task on Quality Estimation, the winning system YSDA (Kozlova et al., 2016) shows that it is effective for Quality Estimation to employ language model probabilities of source and target sentences, and BLEU scores between the source sentence and back-translation. Therefore, we calculate the language model probabilities of source sentences and translate them in the order of their probabilities. To further obtain better translations, we select sentence pairs in the descending order of machine translation output quality, which is defined as follows:

$$QE_i = \mathrm{SBLEU}(e_i, \mathrm{BT}_{\mathrm{PBMT}}(e_i)) \times \mathrm{SBLEU}(e_i, \mathrm{BT}_{\mathrm{NMT}}(e_i)) \quad (1)$$

Here, e_i denotes the i-th source sentence, $\mathrm{BT}_{\mathrm{PBMT}}$ denotes the back-translation using PBMT, $\mathrm{BT}_{\mathrm{NMT}}$ denotes the back-translation using NMT, and SBLEU denotes the sentence-level BLEU score (Nakov et al., 2012). When this score is high, it indicates that the difference in sentence meaning before and after translation is small for each machine translation system.

3.2 Non-Paraphrase Extraction from a Monolingual Corpus

This extraction part of our method is for acquiring non-trivial negative instances. Although the machine translation part of our method is expected to collect non-trivial negative instances too, there will be a certain gap between positive and negative instances. To fill the gap, we randomly collect sentence pairs from a monolingual corpus written in the target language.

To check whether the negative instances acquired by machine translation and those extracted directly from a monolingual corpus are discernible, we asked three people to annotate randomly extracted 100 instances whether a pair is machine-translated or not. The average F-score on the annotation was 0.34. This means the negative instances are not distinguishable, so this does not affect the balance of the corpus.

3.3 Balanced Sampling using Word Overlap Rate

To collect both trivial and non-trivial instances, we carefully sample candidate pairs. We classify the pairs into eleven ranges depending on the WOR and sample pairs uniformly for each range, except for the exact match pairs. The WOR is calculated as follows:

$$\mathrm{Jaccard}(\mathrm{T}_{\mathrm{PBMT}}(e_i), \mathrm{T}_{\mathrm{NMT}}(e_i)) = \left| \frac{\mathrm{T}_{\mathrm{PBMT}}(e_i) \cap \mathrm{T}_{\mathrm{NMT}}(e_i)}{\mathrm{T}_{\mathrm{PBMT}}(e_i) \cup \mathrm{T}_{\mathrm{NMT}}(e_i)} \right| \quad (2)$$

Label	Example		
Positive	Input:	*My father was a very strong man.*	
	PBMT:	私の父は非常に強い男でした。	*My father was a very strong man.*
	NMT:	父はとても強い男だった。	*My father was a very strong man.*
Negative	Input:	*It is available as a generic medication.*	
	PBMT:	これは、一般的な薬として利用可能です。	*It is available as a generic medicine.*
	NMT:	ジェネリック医薬品として入手できます。	*It is available as a generic medication.*
Unnatural	Input:	*I want to wake up in the morning*	
	PBMT:	私は午前中に目を覚ますしたいです*	*I wake up want to in the morning**
	NMT:	私は朝起きたい	*I want to wake up in the morning*
Other	Input:	*Academy of Country Music Awards :*	
	PBMT:	アカデミーオブカントリーミュージックアワード：	*Academy of Country Music Awards :*
	NMT:	アカデミー・オブ・カントリー・ミュージック賞：	*Academy of Country Music Awards :*

Table 2: Annotation labels and examples.

Here, T_{PBMT} and T_{NMT} denote the sentence in the target language translated by PBMT and NMT respectively.

4 Corpus Creation

4.1 Acquiring Candidate Pairs in Japanese

We built the first evaluation corpus for Japanese PI using our method. We used Google Translate PBMT[4] and NMT[5] (Wu et al., 2016) to translate English sentences extracted from English Wikipedia [6] into Japanese sentences[7]. We calculated the language model probabilities using KenLM (Heafield, 2011), and built a 5-gram language model from the English Gigaword Fifth Edition (LDC2011T07). Then we translated the top 500,000 sentences and sampled 200 pairs in the descending order of machine translation output quality for each range, except for the exact match pairs (Table 1).

4.2 Annotation

We used four types of labels; *Positive*, *Negative*, *Unnatural*, and *Other* (Table 2). When both sentences of a candidate pair were fluent and semantically equivalent, we labeled it as *Positive*. In contrast, when the sentences were fluent but semantically inequivalent, the pair was labeled as *Negative*. *Positive* and *Negative* pairs were included in our corpus. The label *Unnatural* was assigned to pairs when at least one of the sentences was ungrammatical or not fluent. In addition, the label *Other* was assigned to sentences and phrases that comprise named entities or that have minor differences such as the presence of punctuation, even though they are paraphrases. *Unnatural* or *Other* pairs were discarded from our corpus.

One of the authors annotated 2,000 machine-translated pairs, then another author annotated the pairs labeled either *Positive* or *Negative* by the first annotator. The inter-annotator agreement (Cohen's Kappa) was κ=0.60. Taking into consideration the fact that PI deals with a deep understanding of sentences and that there are some ambiguous instances without context (e.g., *good child* and *good kid*), the score is considered to be sufficiently high. There were 89 disagreements, and the final label was decided by discussion. As a result, we acquired 363 positive and 102 negative machine-translated pairs.

Although the machine translation part of our method successfully collected non-trivial positive instances, it acquired only a few non-trivial negative instances as we expected. To fill the gap between positive and negative in higher WOR, we randomly collected sentence pairs from Japanese Wikipedia[8] and added 190 non-trivial negative instances. At the end of both parts of our method, we acquired 655 sentence pairs in total, comprising 363 positive and 292 negative instances.

Figures 2 and 3 indicate the distribution of the instances in each corpus. Compared to MSRP and TPC, our corpus covers all ranges of WOR both for positive and negative instances.

[4]GOOGLETRANSLATE function on Google Sheets.

[5]https://translate.google.co.jp/

[6]https://dumps.wikimedia.org/enwiki/20160501/

[7]We trained Moses and translated the sentences from Wikipedia; however, it did not work well. This is the reason why we chose Google machine translation systems, which work sufficiently well on Wikipedia.

[8]https://dumps.wikimedia.org/jawiki/20161001/

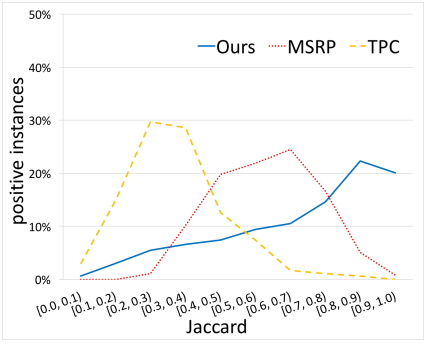

Figure 2: Distributions of positive sentence pairs in each WOR.

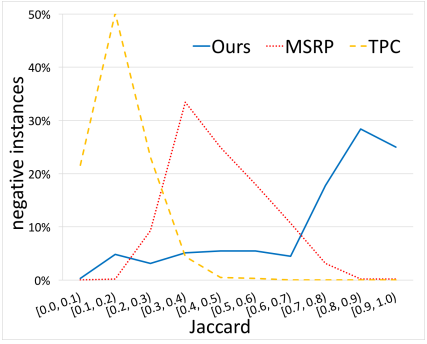

Figure 3: Distributions of negative sentence pairs in each WOR.

Category	%
Content word replacement	63.1
Phrasal/Sentential replacement	25.0
Function word replacement	23.2
Function word insertion/deletion	14.3
Content word insertion/deletion	9.5
Word order	6.5
Lexical entailment	4.2

Table 3: The result of corpus analysis.

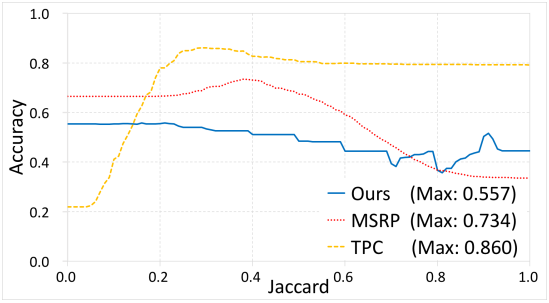

Figure 4: Accuracy of PI using WOR.

5 Discussion

5.1 Corpus Analysis

Table 3 shows the result of corpus analysis on machine-translated instances. We randomly sampled ten pairs from each range of WOR for both positive and negative pairs, i.e., 168 pairs in total, and investigated what type of pairs are included.

We found that most of the data comprises content word replacement (63.1%). Further investigation of this category shows that 30.2% are related to a change in the origin of words and transliterations. In Example # 1 in Table 4, PBMT outputs a transliteration of a *member*, and NMT outputs a Japanese translation. Next, the second most common type of pair is phrasal/sentential replacement (25.0%). When a pair has a bigger chunk of sentence or the sentence as a whole is replaced, it is assigned to this category. This implies that our method, which focuses on sampling by WOR, works to collect non-trivial instances like Examples # 2 and # 3. On the contrary, Example # 4 is an example of instances where machine translations demonstrate each characteristic like that mentioned in Section 1 (PBMT is more literal and NMT is more fluent), so negative instances are produced as we expected. The outputs are semantically close, but the surface is very different. In this example, the PBMT output entails the NMT output.

5.2 Paraphrase Identification

We conducted a simple PI experiment — an unsupervised binary classification. Here, we classified each sentence pair as either paraphrase or non-paraphrase using WOR thresholds and evaluated its accuracy. Figure 4 shows the results from each corpus. Achieving around accuracy of 80% does not mean that the corpus is well built in any language. In that respect, this result proves that our corpus includes more instances that are difficult to be solved with only superficial clues, which helps develop a more robust PI model.

6 Conclusion

We proposed a paraphrase acquisition method to create a well-balanced corpus for PI. Our method generates positive instances using machine translations, extracts negative instances from a monolingual corpus, and uses WOR to collect both triv-

#	Type of Replacement	Jaccard	Label	Trivial/Non-Trivial	Example	
1	Lexical	0.60	P	Trivial	Input:	*He was a **member** of the Republican Party.*
					PBMT:	彼は共和党のメンバーでした。
					NMT:	彼は共和党の一員だった。
2	Lexical	0.90	N	Non-Trivial	Input:	*There is also a strong Roman Catholic **presence**.*
					PBMT:	強力なローマカトリックの存在感もあります。
					NMT:	強力なローマカトリックの存在もあります。
3	Phrasal	0.07	P	Non-Trivial	Input:	*It is **rarely** used.*
					PBMT:	めったに使われることはありません。
					NMT:	まれに使用されます。
4	Phrasal	0.15	N	Trivial	Input:	*Why do **you work so hard**?*
					PBMT:	なぜあなたは一生懸命働くのですか？
					NMT:	どうしてそんなに頑張ってるの？

Table 4: Examples from our corpus. Bold words/phrases were replaced.

ial and non-trivial instances. With this method, we built the first evaluation corpus for Japanese PI. According to our PI experiment, our method made the corpus difficult to be solved.

Our method can be used in other languages, as long as machine translation systems and monolingual corpora exist. In addition, more candidates could be added by including additional machine translation systems. A future study will be undertaken to explore these possibilities.

Acknowledgments

We thank Naoaki Okazaki, Yusuke Miyao, and anonymous reviewers for their constructive feedback. The first author was partially supported by Grant-in-Aid for the Japan Society for Promotion of Science Research Fellows.

References

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.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*. pages 50–57.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 257–267.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. pages 131–198.

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.

David Chen and William Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pages 190–200.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 504–514.

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

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*. pages 9–16.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 1608–1618.

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.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting Paraphrases from Definition Sentences on the Web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pages 1087–1097.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. pages 187–197.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*. pages 35–40.

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*. pages 1–7.

Anna Kozlova, Mariya Shmatova, and Anton Frolov. 2016. YSDA Participation in the WMT'16 Quality Estimation Shared Task. In *Proceedings of the First Conference on Machine Translation*. pages 793–799.

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.

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 (COCOSDA 2014)*. pages 1–4.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Technical Papers*. pages 1979–1994.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. pages 102–109.

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.

Vasile Rus, Rajendra Banjade, and Mihai Lintean. 2014. On Paraphrase Identification Corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. pages 2422–2429.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of the 2002 Human Language Technologiy Conference*. pages 1–6.

Chenguang Wang, Nan Duan, Ming Zhou, and Ming Zhang. 2013. Paraphrasing Adaptation for Web Search Ranking. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 41–46.

Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. 2013. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*. pages 385–404.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In *arXiv preprint arXiv:1609.08144*. pages 1–23.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting Lexically Divergent Paraphrases from Twitter. *Transactions of the Association for Computational Linguistics* 2(1):435–448.