SPADE: Evaluation Dataset for Monolingual Phrase Alignment

Yuki Arase[1⋆] **and Junichi Tsujii**[⋆2]
[1]Osaka University, Japan
[⋆]Artificial Intelligence Research Center (AIRC), AIST, Japan
[2]NaCTeM, School of Computer Science, University of Manchester, UK
arase@ist.osaka-u.ac.jp, j-tsujii@aist.go.jp

Abstract

We create the SPADE (Syntactic Phrase Alignment Dataset for Evaluation) for systematic research on syntactic phrase alignment in paraphrasal sentences. This is the first dataset to shed lights on syntactic and phrasal paraphrases under linguistically motivated grammar. Existing datasets available for evaluation on phrasal paraphrase detection define the unit of phrase as simply sequence of words without syntactic structures due to difficulties caused by the non-homographic nature of phrase correspondences in sentential paraphrases. Different from these, the SPADE provides annotations of gold parse trees by a linguistic expert and gold phrase alignments identified by three annotators. Consequently, $20,276$ phrases are extracted from 201 sentential paraphrases, on which $15,721$ alignments are obtained that at least one annotator regarded as paraphrases. The SPADE is available at Linguistic Data Consortium for future research on paraphrases. In addition, two metrics are proposed to evaluate to what extent the automatic phrase alignment results agree with the ones identified by humans. These metrics allow objective comparison of performances of different methods evaluated on the SPADE. Benchmarks to show performances of humans and the state-of-the-art method are presented as a reference for future SPADE users.

Keywords: phrase alignment, paraphrase detection

1. Introduction

Paraphrases have been applied to various NLP applications, and recently, they are recognized as a useful resource for natural language understanding, such as semantic parsing (Berant and Liang, 2014) and automatic question answering (Dong et al., 2017).

While most previous studies focused on sentential paraphrase detection, *e.g.*, (Dolan et al., 2004), finer grained paraphrases, *i.e.*, phrasal paraphrases, are desired by the applications. In addition, syntactic structures are important in modeling sentences, *e.g.*, their sentiments and semantic similarities (Socher et al., 2013; Tai et al., 2015). A few studies worked on phrasal paraphrase identification on sentential paraphrases (Yao et al., 2013); however, the units of correspondence in previous studies are defined as sequences of words and not syntactic phrases due to difficulties caused by the non-homographic nature of phrase correspondences. To overcome these challenges, one promising approach is phrase alignment on paraphrasal sentence pairs based on their syntactic structures derived by linguistically motived grammar. A flexible mechanism to allow non-compositional phrase correspondences is also required. We have published our initial attempt on this direction (Arase and Tsujii, 2017).

For systematic research on syntactic phrase alignment in paraphrases, an evaluation dataset as well as evaluation measures are essential. Hence, we constructed the SPADE (Syntactic Phrase Alignment Dataset for Evaluation) and released it through Linguistic Data Consortium[1] (catalog ID: LDC2018T09[2]). In the SPADE, 201 sentential paraphrases are annotated gold parse trees, on which $20,276$ phrases exist. Three annotators annotated alignments among these phrases as shown in Figure 1, resulted in $15,721$ alignments that at least an annotator regarded as paraphrases. We also propose two measures to evaluate the quality of phrase alignment on the SPADE, which have been used as official evaluation metrics in (Arase and Tsujii, 2017). These measures allow objective comparison of performances of different methods evaluated on the SPADE.

2. Related Work

Extensive research efforts have been made for sentential paraphrase detection. One of promising resources that provide paraphrases is machine translation evaluation corpora. In such a corpus, a source sentence is translated into multiple translations in a target language. These translations are called reference translations, which can be regarded as sentential paraphrases (Weese et al., 2014). Since the reference translations are constrained to convey the same information in similar structures with the source sentences, they can be regarded as authentic paraphrases containing purely a paraphrasal phenomenon.

Although the amount of reference translations are relatively large thanks to efforts by the research community, they require severe human workloads for creation and expansion. To explore more abundant resources to extract paraphrases, Microsoft Research Paraphrase Corpus (Dolan et al., 2004) aligns news texts published at the same timing as paraphrases. Twitter URL Corpus takes a similar approach on news headlines and comments to them published at Twitter[3]: it uses attached URLs as a primary clew to find paraphrasal candidates (Lan et al., 2017). Since paraphrases in these datasets are not strictly constrained like in the ones extracted from machine translation evaluation corpora, they involve variety of linguistic phenomena beyond the conven-

[1]https://www.ldc.upenn.edu/

[2]https://catalog.ldc.upenn.edu/LDC2018T09 (will be effective since March 2018)

[3]https://twitter.com/

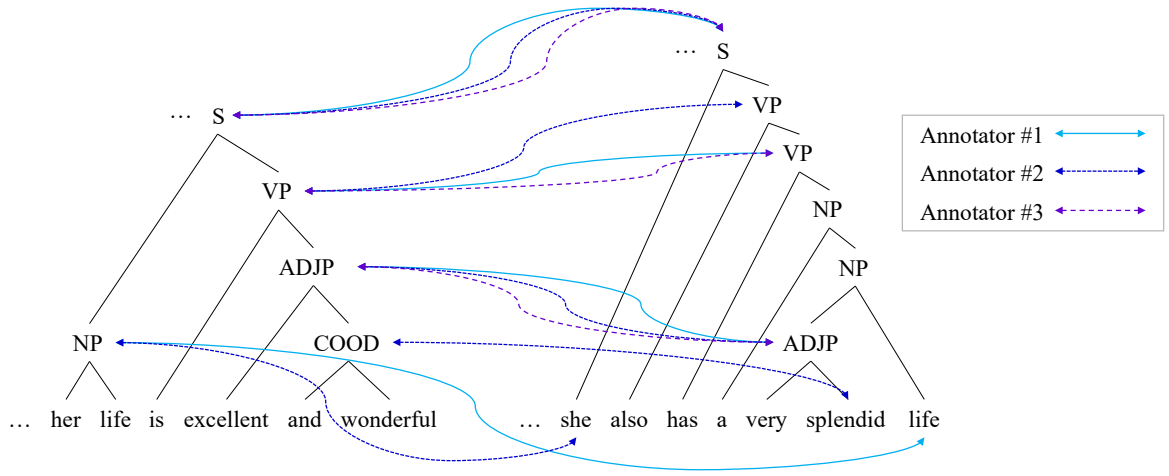

Figure 1: SPADE data example: part of gold trees and phrase alignments on a sentence pair of "Hence, I also have a reason to believe that her life is excellent and wonderful." and "So I have reason to believe that she also has a very splendid life."

tional scope of paraphrasing, such as entailment, inference, and drastic summarization.

Although costly, manually generating paraphrases is the way to produce a high-quality dataset. SICK (Marelli et al., 2014) was constructed from image and video captions through sentence alignment and careful edits to exclude undesired linguistic phenomena. In (Choe and McClosky, 2015), a linguist manually generated paraphrases to given sentences. To scale up the process trading off the quality, crowd-sourcing has been explored (Jiang et al., 2017).

As for phrasal paraphrase datasets, there are only a few; PPDB (Ganitkevitch et al., 2013) and its extension annotated levels of paraphrasability (Wieting et al., 2015). PPDB uses bilingual pivoting on parallel corpora; multiple translations of the same source phrase are regarded as paraphrases. While researchers proposed methods to identify phrasal correspondences for natural language inferences (MacCartney et al., 2008; Thadani et al., 2012; Yao et al., 2013), the unit of phrase was simply n-gram and syntax in paraphrases was out of their scope. Part of PPDB provides syntactic paraphrases under the synchronous context free grammar (SCFG); however, SCFG captures only a fraction of paraphrasing phenomenon (Weese et al., 2014). Hence, the SPADE is unique for providing fully syntactic and phrasal paraphrases.

3. Construction of SPADE

We create the SPADE for evaluation on syntactic phrase alignment in paraphrases. Two rounds of annotations were carefully conducted to annotate gold parse trees and phrase alignments.

3.1. Corpus

Paraphrasal sentence pairs to annotate were extracted from the NIST OpenMT[4] that are machine translation evaluation corpora. As discussed in Section 2., reference translations can be regarded as authentic paraphrases. Other types of paraphrase copora are left for our future study.

Reference translations of 10 to 30 words were randomly extracted for annotation. To diversify the data, only one reference pair per source sentence was chosen.

3.2. Gold-Standard Parse Tree

We essentially need phrase structure grammars to recognize phrases. In addition, we believe that rich syntactic information is useful for deriving rules or applying machine learning techniques in phrase alignment process. Hence, we decided to use Head-driven phrase structure grammar (HPSG) (Pollard and Sag, 1994) to assign gold parse trees to sentences.

We asked a linguistic expert with rich experience on annotating HPSG trees to annotate gold-trees to paraphrasal sentence pairs. Consequently, 201 paraphrased pairs with gold-trees (containing 20, 276 phrases) were obtained.

3.3. Gold-Standard Phrase Alignment

Next, three professional English translators identified paraphrased pairs including phrases with no correspondences given sets of phrases extracted from the gold-trees. These annotators independently annotated the same set.

The annotators were given an annotation guideline with detailed instructions and examples. They were also provided an annotation tool developed by us for simpler annotation process and easier management of progress. Figure 2 shows a screenshot of the annotation tool, in which three panes show (1) phrases extracted from a sentence, (2) phrases extracted from another sentence, and (3) annotation results, from left to right. Annotators select a phrase in each pane and assign an alignment label to declare if the alignment is either *sure* or *possible* according to their confidence in judgment. They can also quickly refer to tree structures by clicking the "Show parse trees" button, then trees are visualized as Figure 3 shows[5]. In the panes of (1) and (2), already-aligned phrases turn their surface colors into gray for easier recognition. The annotators can modify their

[4]LDC catalog numbers: LDC2010T14, LDC2010T17, LDC2010T21, LDC2010T23, LDC2013T03

[5]Script for visualization is borrowed from `http://www.nactem.ac.uk/enju/demo.html`.

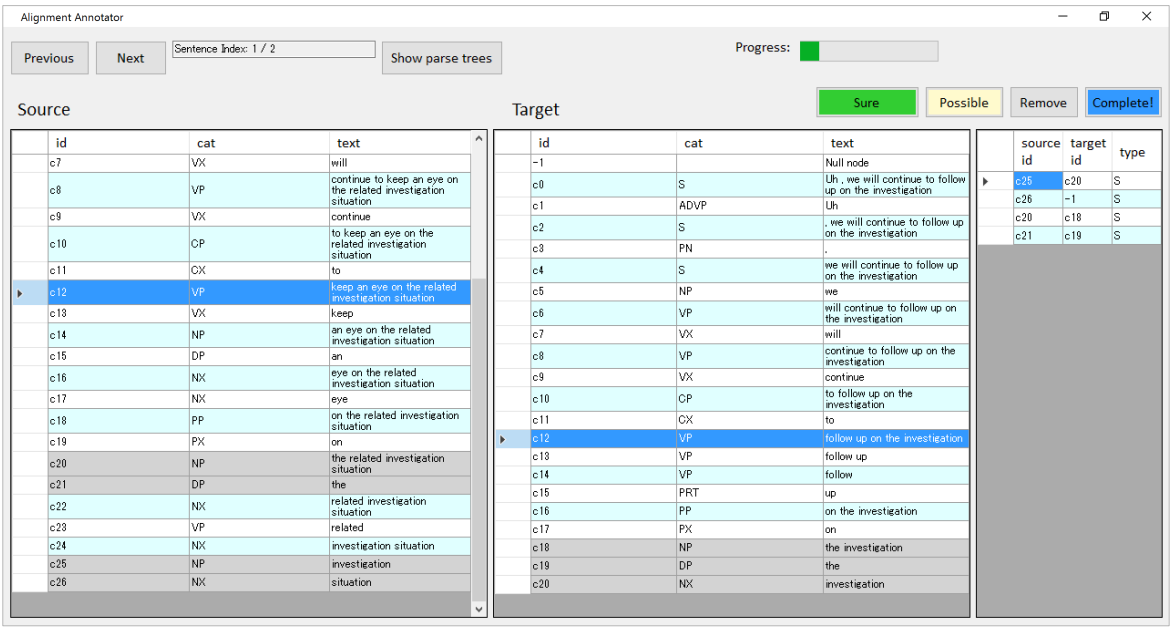

Figure 2: Annotation tool

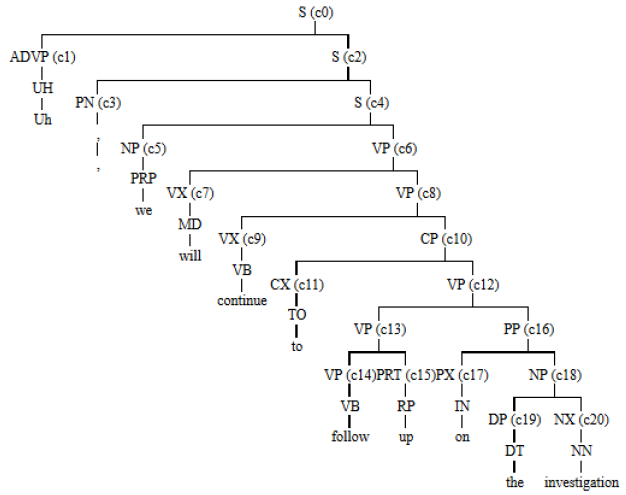

Figure 3: Tree structures of annotating sentences can be quickly referred on the annotation tool.

alignments by directly editing an alignment at the annotation result pane, or erasing it using the "Remove" button.

3.4. Statistics in SPADE

As results of annotation, 15,721 phrase alignments were obtained that at least one annotator regarded as paraphrases. These alignments contain ones in which a phrase does not have any correspondence in another sentence. Such a phrase is regarded as being aligned to a *null* phrase. Figure 1 visualizes gold parse trees and phrase alignments, where alignment types of *sure* and *possible* are not distinguished.

Table 1 shows detailed statistics of the SPADE. The annotated sentence pairs were split into 50 pairs for development and another 151 for testing. In the development and test sets, 3,932 and 11,789 alignments are the ones regarded as paraphrases by at least an annotator, respectively. Among them, 2,518 and 7,134 alignments are agreed by all annotators, respectively. Hence, the overall agreement rate is 61.4%. Although the numbers of sentences in the development and test sets sound limited, we deem that those of phrase alignments are sufficient for evaluation.

	Development	Test
# of sentence pairs	50	151
# of tokens	2,494	7,276
# of types	736	1,573
# of phrases (w/o tokens)	5,201	15,075
# of alignments (∪)	3,932	11,789
# of alignments (∩)	2,518	7,134

Table 1: Statistics of development and test sets in SPADE

4. Evaluation Measure

Phrase alignment quality should be evaluated by measuring the extent that automatic alignment results of a certain method agree with those of humans. We propose two measures, named ALIR (alignment recall) and ALIP (alignment precision) based on conventional recall and precision. These are used in (Arase and Tsujii, 2017) as official evaluation metrics. Together with the SPADE, researchers objectively compare performances of their methods measured by ALIR and ALIP with those of others .

Specifically, ALIR evaluates how gold-alignments can be replicated by automatic alignments and ALIP measures how automatic alignments overlap with alignments that at least an annotator aligned as:

$$\text{ALIR} = \frac{|\{\mathbb{h}|\mathbb{h} \in \mathbb{H}_a \wedge \mathbb{h} \in \mathbb{G} \cap \mathbb{G}'\}|}{|\mathbb{G} \cap \mathbb{G}'|},$$

$$\text{ALIP} = \frac{|\{\mathbb{h}|\mathbb{h} \in \mathbb{H}a \wedge \mathbb{h} \in \mathbb{G} \cup \mathbb{G}'\}|}{|\mathbb{H}a|},$$

Method	ALIR	ALIP
Human	90.65	88.21
(Arase and Tsujii, 2017)	83.64	78.91

Table 2: Benchmarks of ALIR and ALIP on the SPADE

where $\mathbb{H}a$ is a set of automatic alignments, while $\mathbb{G}$ and $\mathbb{G}'$ are the ones that two of annotators produce, respectively. The function of $|\cdot|$ counts the elements in a set.

Since we have three annotators, there are three combinations for $\mathbb{G}$ and $\mathbb{G}'$. The final ALIR and ALIP values are calculated by taking the averages.

4.1. Benchmark

As a benchmark of evaluation using the SPADE, Table 2 shows performances of humans and (Arase and Tsujii, 2017). Although we have *sure* and *possible* alignments, we did not distinguish them in the evaluation due to variance in annotators' decision to assign either label[6].

The performance of the human annotators was assessed by considering one annotator as the test and the other two as the gold-standard, and then taking the averages, which is the same setting as the automatic method. We regard this as the pseudo inter-annotator agreement, since the conventional inter-annotator agreement is not directly applicable due to variations in combinations of aligned phrases as depicted in Figure 1.

ALIR and ALIP of (Arase and Tsujii, 2017) reach 92% and 89% of those of humans, respectively, though there still significant gaps to the human performance. Research efforts from variety of groups are desired for further progress in syntactic phrase alignment research.

5. Conclusion

We created the SPADE that provides annotations of gold HPSG trees and phrase alignments on sentential paraphrases extracted from machine translation evaluation corpora. This dataset was released through LDC. Two evaluation measures, ALIR and ALIP, were also proposed, which allow to compare the extent that automatic phrase alignment results agree with the ones produced by humans.

The creation of SPADE is just initiated, and there remains room for future development. The most important task is employing other sentential paraphrase corpora for annotation. We are analyzing the linguistic phenomena observed in Twitter URL Corpus and Microsoft Research Paraphrase Corpus, where drastic summarization or lengthening by adding context, as well as entailments and inference due to people's world knowledge are happening. These are beyond the conventional scope of paraphrasing; however, considering such paraphrases spontaneously arise and widely available, we should definitely need technologies to handle such paraphrases in the wild. We will annotate paraphrases in these corpora to scale up the SPADE in terms of its size as well as the variety of paraphrasal phenomena.

[6]We also observed results using only *sure* alignments, and confirmed that they show the same trends.

Acknowledgments

We thank Dr. Yuka Tateishi for her contribution to HPSG tree annotation. This project is funded by Microsoft Research Asia and the Kayamori Foundation of Informational Science Advancement.

Bibliographical References

Arase, Y. and Tsujii, J. (2017). Monolingual phrase alignment on parse forests. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–11, Copenhagen, Denmark, September.

Berant, J. and Liang, P. (2014). Semantic parsing via paraphrasing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1415–1425, Baltimore, Maryland, June.

Choe, D. K. and McClosky, D. (2015). Parsing paraphrases with joint inference. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1223–1233, Beijing, China, July.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceesings of the International Conference on Computational Linguistics (COLING)*, pages 350–356, Geneva, Switzerland, August.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017). Learning to paraphrase for question answering. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Copenhagen, Denmark, September.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, Atlanta, Georgia, June.

Jiang, Y., Kummerfeld, J. K., and Lasecki, W. S. (2017). Understanding task design trade-offs in crowdsourced paraphrase collection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 103–109, Vancouver, Canada, July.

Lan, W., Qiu, S., He, H., and Xu, W. (2017). A continuously growing dataset of sentential paraphrases. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1235–1245, Copenhagen, Denmark, September.

MacCartney, B., Galley, M., and Manning, C. D. (2008). A phrase-based alignment model for natural language inference. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 802–811, Honolulu, Hawaii, October.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceesings of the Language Re-*

sources and Evaluation Conference (LREC), pages 216–223, Reykjavik, Iceland.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. CSLI Publications.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA, October.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *ACL-IJCLNLP*, pages 1556–1566, Beijing, China, July.

Thadani, K., Martin, S., and White, M. (2012). A joint phrasal and dependency model for paraphrase alignment. In *Proceesings of the International Conference on Computational Linguistics (COLING)*, pages 1229–1238, Mumbai, India, December.

Weese, J., Ganitkevitch, J., and Callison-Burch, C. (2014). PARADIGM: Paraphrase diagnostics through grammar matching. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 192–201, Gothenburg, Sweden, April.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association of Computational Linguistics (TACL)*, 3(1):345–358.

Yao, X., Van Durme, B., Callison-Burch, C., and Clark, P. (2013). Semi-Markov phrase-based monolingual alignment. In *Proceesings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 590–600, Seattle, Washington, USA, October.