Visual Description Paraphrase Corpus Creation
with Various Elementary Operations [*]

Johanes Effendi[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2] ([1]NAIST, [2]RIKEN AIP)

1 Introduction

This paper reports the construction of a multi-paraphrase corpus of visual description through proposed paraphrase elementary operations (i.e., reordering, substitution, deletion, insertion) using crowdsourcing platform. Unlike existing paraphrase corpus, the use of elementary operations enables the possibility to track what operation has been done to the original sentence. Furthermore, the combinations of this elementary operations can be utilized for various applications such as abstractive summarization using deletion, and data augmentation using substitution or reordering.

2 Multi-paraphrase Corpus Creation

2.1 Possible Variants of Paraphrases

Most existing paraphrase corpora cover only one-to-one parallel sentences, overlooking the fact that possible variants of paraphrases can be generated from a single source sentence. On the other hand, Bhagat and Hovy categorized the variations of how humans paraphrase [1] and argued that there should be a broader definition of paraphrasing that is accepted by linguist. Based on this idea, they analyzed the paraphrase characteristics in various studies and corpora and established 25 quasi-paraphrase classes, such as change of tenses, metaphor substitution, function word variations, etc. Each quasi-paraphrase class has its own way of implementing the semantic equivalence standards of a paraphrase.

As reported in their paper, many quasi-paraphrases have very small frequency when it is surveyed on the real corpora. We argue that these 25 classes can actually be grouped into fewer classes. We grouped them and created a constraint for each of our sub-corpus into the following four operations: deletion, insertion, reordering, and substitution.

2.2 Corpus Creation

Following [3], we used image as the basis of paraphrasing, and paraphrase the visual description using these four operations. We collected 10 k images
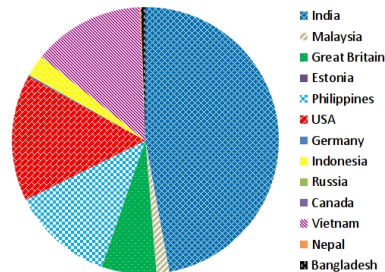

Fig. 1　Country distribution of crowdworkers from Visual Genome [2], which consists of such diverse objects as animals, people, and vehicles, and asked the crowdworkers to describe these images and create paraphrases of the captions into four different paraphrases based on each operation. The corpus consists of 5-pairs of paraphrases with 50 k sentences. Table 1 show an example of the resulting image description and the paraphrased expressions by the crowdworkers.

We used Crowdflower (now Figure Eight) as a crowdsourcing platform. For one session, each crowdworker had to describe and paraphrase at least two images and also be available for additional sessions of image annotation. We limited this task to English-speaking countries or those countries where English is the second language. We also monitored the crowdworker results. If the resulting sentences were not valid paraphrases, we discarded them from the data.

We successfully gathered more than 200 workers from 13 countries for this corpus creation. Most workers came from India, the United States, and Philippines. The remaining countries and distribution are shown in Fig. 1. Each worker created an average of 52.56 images took an average of 7 minutes 54 seconds to describe an image and write its four paraphrases.

3 Corpus Analysis

3.1 Operation Characteristics

To calculate the deletion operation effect on the source sentences, we compared the ratio of the num-

Table 1　Visual description and its paraphrase examples

Image	Operation		Sentence
	Image Caption		A cat is scratching the seats in the park with its nail.
	Paraphrase	**Deletion**	A cat is scratching the seats in the park
		Insertion	A cat is scratching the seats in the park with its sharp nail.
		Substitution	A cat is scratching the bench in the park with its nail.
		Reordering	With its nail, a cat is scratching the seats in the park

ber of words in the target sentence to the number in the source sentence and found that the number of words in the target sentences decreased by 19.53%, where an average of 2.23 words was deleted per sentence (Table 2). We did the same calculation in the insertion sub-corpus and found that the number of words in the target sentences increased by 25.15%, where an average of 2.93 words were inserted per sentence.

Table 2　Operation characteristics

Parameter	# word
Avg. deleted words per sentence	2.2286
Avg. inserted words per sentence	2.9311
Avg. distance of reordered word	4.4252
Avg. word substitutions per sentence	1.6770

To measure the reordering elementary operation for the sub-corpus, we calculated the shift distance of a word in the source and target sentences and found that those in the latter shifted on average by as many as 4.42 words. The distance calculated in the reordering happened when the source sentence was paraphrased into its passive form, or to exchange the order of such sentence information as time, place, and tool. As seen in Table 2, we found an average of 1.68 word substitutions per sentence in the substitution sub-corpus. This means that at most 1 or 2 words were substituted in a sentence.

3.2　Evaluation Based On Word Type

We counted the types of words that were most deleted and inserted. For reordering and substitution, we counted a pair of source and target word types. This evaluation identifies which word type was usually preferred by the crowdworkers.

Table 3　Top 5 Most Operated POS Tags

Deletion	Insertion	Reordering	Substitution
NN	NN	DT DT	NN NN
JJ	IN	NN NN	NN NNS
IN	JJ	IN IN	NN JJ
DT	DT	JJ JJ	NNS NNS
VBG	NNS	NNS NNS	NNS NN

Table 3 shows that nouns (NN), adjectives (JJ), and conjunctions (IN) were usually deleted or inserted. This correlates with how most sentences are deleted or inserted: by adding or removing time, place, or tool information. For reordering the sub-corpus, we found that no word type is changed by reordering and substitutions which implies no major alterations to the words or the semantics.

4　Conclusions

This paper presented a construction of a multi-paraphrase corpus consisting of four elementary operations, enable multi-paraphrase sentences to be generated from a source sentence. Our corpus was developed using a crowdsourcing platform that collected 5-pairs of paraphrases of 10 k images resulting in 50 k sentences. In the future, we will utilize our paraphrase corpus for various applications, including summarization or machine translation.

5　Acknowledgements

References

[1] R. Bhagat and E. Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.

[2] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[3] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016*, pp. 2923–2934, Osaka, Japan, December 2016.