公開

（事後評価・研究者用）

研　究　報　告　書

「画像をピボットとしたパラフレーズの抽出による自然言語と画像理解の高度化」

研究期間：平成２９年１０月～平成３１年３月
研究者番号：５０１４６
研 究 者：チョ シンキ

１．研究のねらい

Natural language and perception are two major ways for human communication and knowledge propagation, making natural language and image understanding indispensable for artificial intelligence. A paraphrase is a restatement of the meaning of a word, phrase or sentence within a single language (e.g., “A snowboarder” and “A man on a snowboard” are paraphrases). Paraphrases have been exploited for natural language understanding, and have been shown very effective for various natural language processing (NLP) tasks, including information retrieval, question answering, summarization, machine translation, text normalization, textual entailment recognition, and semantic parsing. However, as no image information is provided, paraphrases have not been exploited for image understanding.

As manual construction of paraphrases is very expensive and time consuming, two lines of studies have been conducted to automatically extract paraphrases. One is extracting paraphrases from monolingual corpora based on distributional similarity. However, it has been known that the quality of the extracted paraphrases is very low, because other related phrase pairs such as hypernyms and antonyms also share distributional similarity, and thus being extracted. The other is extracting paraphrases from parallel corpora via bilingual pivoting. Bilingual pivoting does not have the problem as distributional similarity. However, the big limitation is that large-scale parallel corpora are only available for a few languages such as European languages and languages paired with English.

In this work, we propose a novel approach that uses images as a pivot for paraphrase extraction. Nowadays, with the spread of the web and social media, it is easy to collect large amounts of images with their describing text. For example, different news sites release news with the same topic using the same image; photos with many comments are posted to social networking sites and blogs. As the describing text is written by different people but about the same image, there are potentially large amounts of paraphrases in the describing text. This work aims to **accurately extract these paraphrases together with their corresponding image regions using images as a pivot**. Together with the image information, the extracted paraphrases have the potential to significantly deepen both natural language and image understanding.

２．研究成果

（1）概要

A paraphrase is a restatement of the meaning of a text in other words. Paraphrases have been studied to enhance the performance of many natural language processing tasks. In the ACT-I research period, we studied **a novel concept to identify visually grounded paraphrases (VGPs)**, which are **different phrasal expressions describing the same visual concept in an image** [2]. Figure 1 shows an example of VGP identification. Our task is to identify the noun phrases that describe the same visual concept (represented as an image region) in the image as VGPs.

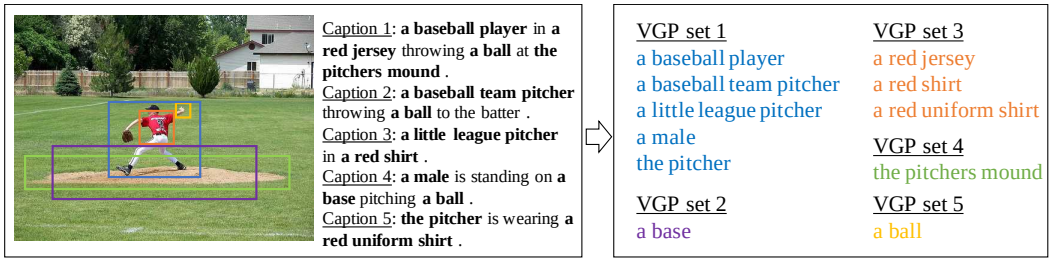

Figure 1: An example of VGP identification.

These identified VGPs have the potential to improve language and vision tasks such as visual question answering and image captioning. Because of the characteristic of VGPs, visual grounding is essential for VGP identification. We proposed two novel neural network based models with different visual grounding approaches for VGP identification, and published our work at **COLING 2018 (top conference) [2] and CVPR 2018 workshop [3]**, respectively.

（2）詳細

2.1 Attention based VGP Identification

The attention based VGP identification model [2] is illustrated in Figure 2. Given a noun phrase pair and its corresponding image, we construct two separated *fusion nets* for each phrase (Figure 2 (right)). A fusion net represents a phrase with a concatenation of its feature vector and visual context feature. The visual context feature is computed with an attention mechanism, indicating to which part of the image should be paid attention, in order to judge whether the phrase pair is VGPs or not. The outputs of the two fusion nets are then fed into a multilayer perceptron to compute the similarity of the two phrases.

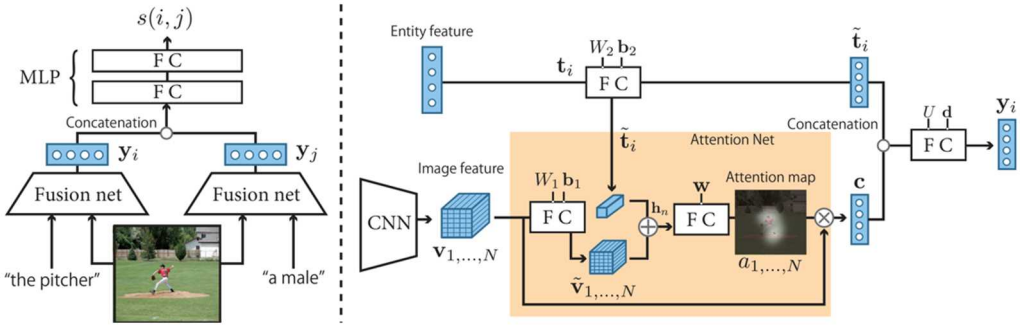

Figure 2: Attention based VGP identification model (left) and its fusion sub-network (right).

2.2 Phrase Localization based VGP Identification

The phrase localization based VGP identification model [3] is illustrated in Figure 3. This model also takes a pair of noun phrases and an associated image as input and predicts whether the input phrases are VGPs or not. The model localizes a region in the image that corresponds to each phrase. Such an image region serves as object-level visual features and allows better similarity modeling. Both language and visual features obtained for the input phrases and the image are fused and fed to a multilayer perceptron that predicts the probability of being VGPs.

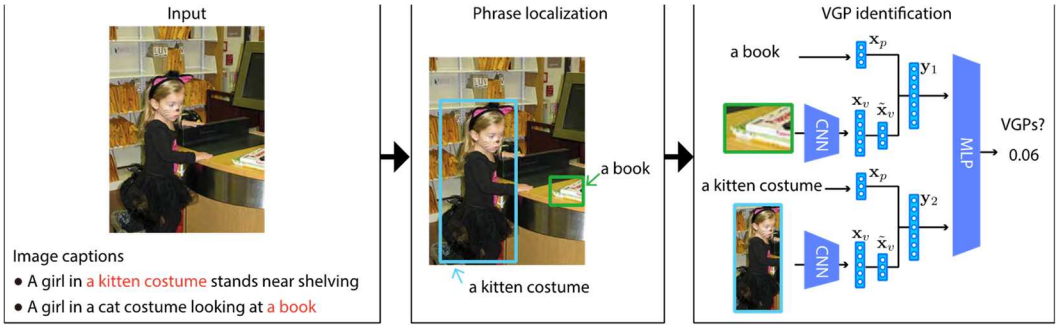

Figure 3: Phrase Localization based VGP Identification model.

2.3 Experiments

We evaluated our models on the Flickr30k entities (i.e., noun phrases) dataset (Plummer et al. 2015). This dataset contains 30k, 1k, and 1k images with 5 captions for training, validation, and testing, respectively. It also provides manually annotated image regions localizing noun phrases in captions. We treated phrases associated with the same image region as VGPs, and the other possible noun phrase pairs from different captions as non-VGPs. The following table shows the results, where phrase-only is a model that only uses phrase features. We can see that the localization based model, which explicitly models the phrase-object correspondence, shows the best performance.

	F1	Precision	Recall
Phrase-only	85.37	84.75	85.99
Attention based	84.16	82.71	85.67
Localization based	**87.47**	**86.61**	**88.35**

Reference

B. Plummer et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of ICCV, pp.2641-2649, (2015).

3. 今後の展開

We originally created the concept of VGPs and established the fundamental techniques for VGP identification in the ACT-I research period. There are several directions for future work to deepen and promote the research in both language and vision understanding with VGPs. We hope that our work can inspire other researchers work on these directions together with us.

Firstly, VGP identification for visual relations. We have studied VGP identification for entities

(noun phrases) only. For the next, we want to work on VGPs of visual relations, which requires semantic understanding of the relationships between objects in an image. A visual relation is a triplet of (entity1, entity2, relation), where the relation is the interaction between the two entities. We have created a dataset of VGPs for visual relations upon the Flickr30k entities dataset [1]. We plan to propose novel models for identifying visual relation based VGPs.

Secondly, VGP typology creation. Our research until now treats VGP identification as a binary classification task, which ignores various phenomena behind VGPs. E.g., "field hockey" and "lacrosse" are linguistic paraphrases; "competitors" and "a group of bicyclists" describe the same visual concept from different aspects; however, these two pairs of VGPs have been treated equally, which is obvious undesirable. In the future, we aim to create typology of VGPs to elucidate the phenomena behind VGPs and open up novel ways of utilizing VPGs for various language and vision tasks.

Thirdly, VGP identification in the noisy real world. We have only worked on the image captioning dataset, where sentential level VGPs are already annotated. In the real world, such as e-commerce or social networking sites and blogs, sentential level VGPs are unavailable. How to identify VGPs in these noisy scenarios is obviously a more challenging task, for which we want to study in the future.

4. 自己評価

・研究目的の達成状況

We have successfully achieved our research goal, which was extracting paraphrases pivoted by image regions accurately. We further proposed the novel concept of VGPs.

・研究の進め方（研究実施体制及び研究費執行状況）

Regarding to the research implementation system, because of the multimodality of the project, we closely collaborated with researchers in the computer vision (CV) field. We also employed RA students for creating visually grounded paraphrase datasets.

Regarding to the research funding execution status, we reasonably used it for purchasing GPU servers, dataset annotation, RA employment, and travel expenses.

・研究成果の科学技術及び学術・産業・社会・文化への波及効果（今後の見込みも重視してください。）

VGP identification is a novel task, which is crucial for both language and vision understanding. We believe that our pioneering work on VGPs in the ACT-I research period has the potential to make VGPs a new language and vision research field in both NLP and CV fields. We are eager to promote the VGP research field with other researchers.

・研究課題の独創性・挑戦性

We originally proposed the concept of VGPs, which is an achievement inspired by our study in both NLP and CV. Due to the pioneer and multimodality characteristics, this work is very

challenging. We overcame the problems via close collaboration between NLP and CV researchers.

5. 主な研究成果リスト

(1) 論文 (原著論文) 発表:0件

(2) 特許出願

研究期間累積件数:0 件

(3) その他の成果 (主要な学会発表、受賞、著作物、プレスリリース等)

[1] 竹林佑斗, C. Chu, 中島悠太. 画像内物体間の視覚的関係の真偽判定データセット. 言語処理学会第 25 回年次大会, pp.1383-1386, (2019).

[2] C. Chu, M. Otani, Y. Nakashima. iParaphrasing: Extracting Visually Grounded Paraphrases via an Image. In Proc. of COLING, pp.3479-3492, (2018).

[3] M. Otani, C. Chu, Y. Nakashima. Visually Grounded Paraphrase Extraction via Phrase Grounding. Workshop on Language and Vision at CVPR, (2018).

[4] M. Otani, C. Chu, Y. Nakashima. Phrase Localization-based Visually Grounded Paraphrase Identification. MIRU 2018.

[5] C. Chu, M. Otani, Y. Nakashima. Visually Grounded Paraphrase Extraction. In Proc. of NLP, pp.979-982, (2018).

[6] C. Chu, M. Otani, Y. Nakashima. Extracting Paraphrases Grounded by an Image. CVIM 211, (2018).

[7] C. Chu, M. Otani, Y. Nakashima. Towards Image-pivoted Paraphrase Extraction. YANS, (2017).