Extracting Paraphrases Grounded by an Image

Chenhui Chu[1,a)] Mayu Otani[2,b)] Yuta Nakashima[2,c)]

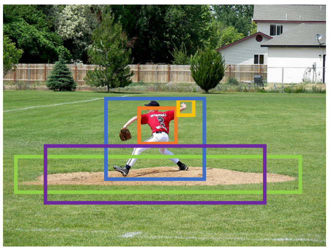
Abstract: A paraphrase is a restatement of the meaning of a text in other words. Paraphrases have been studied to enhance the performance of many natural language processing tasks. In this paper, we propose a novel task to extract visually grounded paraphrases (VGPs), which are different phrasal expressions describing the same visual concept in an image. These extracted VGPs have the potential to improve language and image multimodal tasks such as visual question answering and image captioning. How to model the similarity between VGPs is the key of VGP extraction. We apply various existing methods as well as propose a novel neural network-based method with image attention, and report the results of the first attempt toward VGP extraction.

1. Introduction

A paraphrase is a restatement of the meaning of a word, phrase, or sentence within the context of a specific language (e.g., "a red jersey" and "a red uniform shirt" in Figure 1 are paraphrases) [7]. Paraphrases have been exploited for natural language understanding, and shown to be very effective for various natural language processing (NLP) tasks, including question answering [38], summarization [50], machine translation [13], text normalization [28], textual entailment recognition [1], and semantic parsing [5].

In this paper, we propose a novel task to extract *visually grounded paraphrases (VGPs)*. We define VGPs as different phrasal expressions that describe the same visual concept in an image. 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 these describing texts are written by different people but about the same image, there are potentially large amounts of VGPs in the describing text (Figure 1). We aim to accurately extract these paraphrases using the image as a pivot to associate different phrases.

The extracted VGPs can be applied to various computer vision (CV) and NLP tasks, such as image captioning [44] and visual question answering (VQA) [47], for the better understanding of both images and languages. For example, a VQA system must understand queries of different expressions about the same visual concept (e.g., "a male" and "the pitcher" in Figure 1) in order to answer a question properly. VGPs can also be applied to the evaluation of image captioning systems in the similar way as paraphrases have been used for machine translation evaluation [41].



Caption 1: **a baseball player** in **a red jersey** throwing **a ball** at **the pitchers mound** .
Caption 2: **a baseball team pitcher** throwing **a ball** to the batter .
Caption 3: **a little league pitcher** in **a red shirt** .
Caption 4: **a male** is standing on **a base** pitching **a ball** .
Caption 5: **the pitcher** is wearing **a red uniform shirt** .

VGP set 1
a baseball player
a baseball team pitcher
a little league pitcher
a male
the pitcher

VGP set 2
a red jersey
a red shirt
a red uniform shirt

VGP set 3
the pitchers mound

VGP set 4
a base

VGP set 5
a ball

Fig. 1: An example from the Flickr30k entities dataset, in which an image is described by five captions (entities in the captions are marked in bold). Our task is to extract the entities that describe the same visual concept (represented as an image region) in the image as VGPs. Note that the image regions are not given as input but are drawn here for comprehensibility.

As a pioneering study, we work on the Flickr30k entities dataset [36]. This dataset contains 30k images with 5 captions per image annotated via crowdsourcing, which can be seen as a very small subset of the data available in the web and social media. Figure 1 shows an example image together with its five captions taken from this dataset. In the Flickr30k entities dataset, entities (i.e., noun phrases) in the captions have been manually aligned to their corresponding image regions [36]. Therefore, we can obtain a set of phrases annotated with the same image region. This set of phrases are used as the ground truth VGPs in our study. The goal of this work is to extract these VGPs.

We formulate our task as a clustering task (Section 3), where the similarity between each entity pair is crucial for the performance. We apply many different unsupervised similarity computation methods (Section 4) including phrase localization-

[1] Institute for Datability Science, Osaka University
[2] Nara Institute of Science and Technology
[a)] chu@ids.osaka-u.ac.jp
[b)] otani.mayu.ob9@is.naist.jp
[c)] n-yuta@ids.osaka-u.ac.jp

based similarity [35] (Section 4.1), translation probability-based similarity [25] (Section 4.2), and embedding-based similarity [24], [31], [36] (Section 4.3). In addition, we propose a supervised neural network (NN)-based method using both textual and visual features to explicitly model the similarity of an entity pair as VGPs (Section 5). Experiments show that our proposed NN-based method outperforms the other methods.

2. Related Work

2.1 Paraphrase Extraction

Previous studies extract paraphrases from either monolingual corpora or bilingual parallel corpora. One major approach is to use the distributional similarity [21] with regular monolingual corpora (a large collection of text in a single language) [8], [27], [30], or monolingual comparable corpora (a set of monolingual corpora that describe roughly the same topic in the same language) [4], [11]. Distributional similarity stems from the distributional hypothesis [21], stating that words/phrases that share similar meanings should appear in similar distributions. This approach sometimes suffers from noisy results, because the distributed similarity often maps antonyms to closer points. Some methods try to extract paraphrases from monolingual parallel corpora (a collection of sentence level paraphrases) [2], [29], but such monolingual parallel corpora are rarely available.

Bilingual parallel corpora (a collection of sentence-aligned bilingual text) enjoys more availability than monolingual parallel corpora as they are mandatory for training machine translation systems. Bilingual parallel corpora can be used for paraphrase extraction, with bilingual pivoting [3]. This method assumes that two source phrases are a paraphrase pair if they are translated to the same target phrase. Bilingual pivoting has been further refined by using syntax information [10] or mutual information [23]. These methods have led to the construction of a multilingual paraphrase database [17].

Note that our definition of paraphrases may look different from the studies mentioned above, as our paraphrases are a set of noun phrases that represent the same visual concept. Our idea to extract paraphrases under this definition is to use image captioning datasets [12], [49], which usually contain several captions for each image, and currently scale to sub-million images, instead of a bilingual parallel corpus with limited availability. To the best of our knowledge, this is the first study that aims to extract paraphrases from such multimodal datasets consisting of images and their captions.

2.2 Phrase Localization

Phrase localization is a task to find an image region that corresponds to a given phrase in a caption, which is closely related to our VGP extraction task. Plummer et al. ([36]) pioneered this work, in which they annotated phrase-region alignment in the Flickr30k image-caption dataset [49] and released it as the Flickr30k entities dataset. They also proposed a method based on canonical correlation analysis (CCA) [20] that learns joint embeddings of phrases and image regions for associating them. Wang et al. ([45]) proposed joint embeddings using a two-branch NN. Fukui et al. ([15]) used a multimodal compact bilinear pooling method to combine textual and visual embeddings. Rohrbach et al. ([39]) proposed a convolutional NN (CNN)-recurrent NN (RNN)-based method for this task. They learn to detect a region for a given phrase and then reconstruct the phrase using the detected region. Wang et al. ([46]) noticed that the relationships between phrases should agree with their corresponding regions, and proposed a joint matching method, but their method only considers the "has-a" relationship that is explicitly indicated by possessive pronouns. Plummer et al. ([35]) used spatial relationships between pairs of entities connected by verbs or prepositions, which achieved the state-of-the-art performance.

In this paper, we use the current state-of-the-art phrase localization method of [35] as a baseline for VGP extraction.

2.3 Other Vision and Language Tasks

Vision and language tasks have been a hot research area recently in both the CV and NLP communities. Various efforts have been made for many multimodal tasks such as visual captioning [6], [26], [44], [48], text-image retrieval [33], visual question answering [47], and video event detection [34]. Some researchers also have employed images for improving NLP tasks, such as multimodal machine translation [42], cross-lingual document retrieval [16], and textual entailment recognition [19]. VGP extraction is a novel CV+NLP task, which to the best of our knowledge has not been studied before and can boost the performance of various multimodal and NLP tasks.

3. Paraphrase Extraction via Clustering

We formulate the paraphrase extraction from the Flickr30k entities dataset as a clustering task. Given an image and all the entities in the corresponding captions, the task is to cluster the entities[*1] to its corresponding visual concepts represented as image regions. The number of clusters (i.e., the number of paraphrase sets in a set of an image and captions) is not explicitly given in our task. Therefore, we apply the affinity propagation algorithm [14] to cluster entities, which can estimate the number of clusters as well.

Affinity propagation creates clusters by iteratively sending two types of messages between pairs of entities until convergence. The first type is the responsibility $r(i, j)$ sent from entity i to candidate representative entity j, indicating the strength that entity j should be the representative entity for entity i, which is defined as:

$$r(i, j) \leftarrow s(i, j) - \max_{\forall j' \neq j}\{a(i, j') + s(i, j')\} \qquad (1)$$

where $s(i, j)$ is the similarity between entities i and j. The second type is the availability $a(i, j)$ sent from candidate representative entity j to entity i, indicating to what degree that candidate representative entity j is the cluster center for entity i, which is defined as:

$$a(i, j) \leftarrow \min\Big\{0, r(j, j) + \sum_{\forall i' \notin \{i,j\}} \max\{0, r(i', j)\}\Big\} \qquad (2)$$

At the beginning, the values of $r(i, j)$ and $a(i, j)$ are set to zero,

[*1] In this paper, we assume that entities are given. In the case that entities are not given, we can easily extract them by chunking the noun phrases.

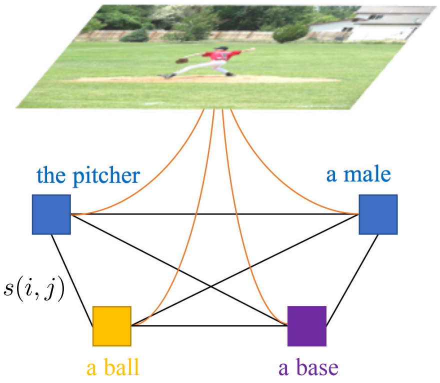


Fig. 2: An overview of our VGP extraction formulation. We extract VGP via clustering, where the entity-entity similarity $s(i, j)$ is the key. We compare both unsupervised and supervised methods using entity-image and entity-entity associations for computing this similarity.

and they are updated in every iteration until convergence. We optimize the number of clusters on a validation split by adjusting the preference (i.e., self similarity $s(i, i)$) of affinity propagation.

Figure 2 shows an overview of our formulation, where the similarity between the entities is the key. We apply various unsupervised methods for computing this similarity, and propose a supervised NN-based model.

4. Unsupervised Similarity Methods

We apply phrase localization for modeling the entity-entity similarity based on entity-image association (Section 4.1). In addition, we apply various methods for modeling the entity-entity similarity directly (Sections 4.2, and 4.3).

4.1 Phrase Localization-Based Similarity

The similarity between entities i and j is defined as:

$$s(i, j) = \sum_{r_m \in R} p(i|r_m) p(j|r_m) \tag{3}$$

where R is a set of image regions that are aligned to both entities i and j obtained with the phrase localization method of [35]; $p(i|r_m)$ is the localization probability of r_m for i, defined as:

$$p(i|r_m) = \frac{l(i, r_m)}{\sum_{r_m \in R} l(i, r_m)} \tag{4}$$

where $l(i, r_m)$ is the localization score of region r_m for entity i obtained using the method of [35].

4.2 Translation Probability-Based Similarity

The similarity between entities i and j is defined as:

$$s(i, j) = p(i|j) p(j|i) \tag{5}$$

where $p(i|j)$ and $p(j|i)$ are the direct and inverse translation probabilities of an entity pair i and j, which are calculated using a conventional statistical machine translation (SMT) [25] method:

(1) Generate a pseudo parallel corpus using the captions in the dataset, which treats the 5 captions for each image as monolingual parallel sentences and pair each of the sentences that leads to $\binom{5}{2} = 10$ sentence pairs per image.

(2) Apply word alignment to the parallel corpus using IBM alignment models [9] in two directions with the grow-diag-final-and heuristic [25] to align the words in each caption pair.

(3) From the word-aligned parallel corpus, extract entity pairs such that the words inside an entity pair are aligned. Then $p(i|j)$ and $p(j|i)$ are calculated as follows:

$$p(i|j) = \frac{c(i, j)}{\sum_k c(i, k)}, \quad p(j|i) = \frac{c(i, j)}{\sum_k c(j, k)} \tag{6}$$

where $c(i, j)$ is the number of co-occurrence of i and j in the word-aligned corpus.

4.3 Embedding-Based Similarity

In this method, the similarity between entities i and j is defined as:

$$s(i, j) = \frac{\mathbf{v}_i^{\top} \mathbf{v}_j}{\mathbf{v}_i \mathbf{v}_j} \tag{7}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the phrase embeddings of i and j. We compare three different methods for phrase embeddings.

4.3.1 Word Embedding Average

We represent each word with a 300 dimensional word2vec [31] vector pre-trained on the Google News corpus.*2 We remove stop words in each entity, and calculate the representation of each entity using the average of all word embeddings.

4.3.2 Fisher Vector

Fisher vector is a pooling over word2vec vectors of individual words [24], which has been used in the phrase localization task for representing the entities [36]. To compute the Fisher vector for an entity, we represent the entity by the HGLMM Fisher vector encoding [24] of the word vectors, following [36].*3

4.3.3 Fisher Vector with CCA

Projecting the feature vectors of image regions and entities to a shared semantic space can provide strong associations between the image regions and entities, which has the potential to improve the performance of VGP extraction. Therefore, we learn a CCA projection on the Flickr30k entities dataset for the image region feature vectors and entity feature vectors with [36], in which the normalized CCA formulation of [18] is used. The columns of the CCA projection matrices are scaled by the eigenvalues, and the feature vectors are projected by these matrices and normalized to the dimensionality of 4,096. The image region feature vectors are extracted using Faster R-CNN [37].*4 We use Fisher vectors for entity feature vectors.

5. Supervised Similarity Model Based on Neural Network with Image Attention

We propose a NN-based supervised model. This model computes the similarities of entity pairs as VGPs by explicitly mod-

*2 https://github.com/mmihaltz/word2vec-GoogleNews-vectors

*3 The Fisher vector is constructed with 30 centers of both first and second order information, which results in a very sparse vector whose dimensionality is 300 × 30 × 2 = 18000. Therefore, we apply principal component analysis (PCA) to convert it to a lower dimensionality of 4,096.

*4 https://github.com/ShaoqingRen/faster_rcnn

ⓒ 2018 Information Processing Society of Japan

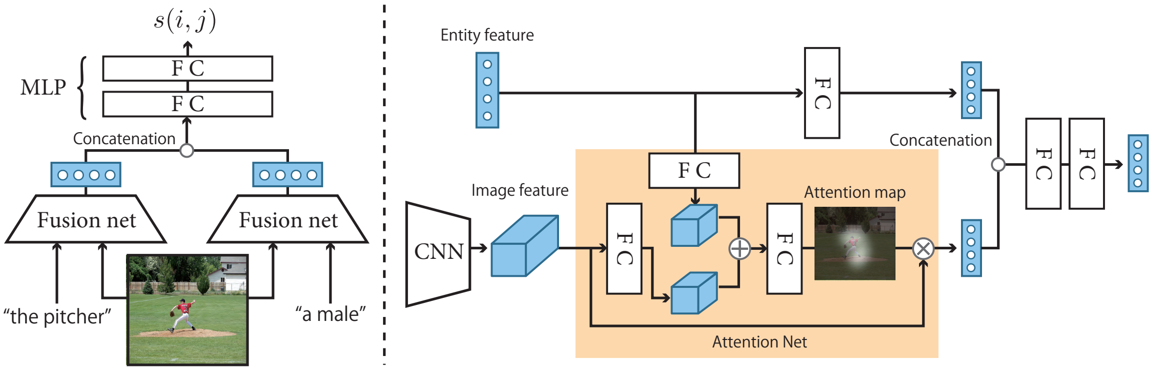


Fig. 3: Our supervised NN with image attention-based similarity model (left) and its fusion sub-network (right).

eling the associations between them and an image. Figure 3 illustrates our proposed NN model. Given an entity pair and its corresponding image, we construct two separated *fusion nets* for each entity (Figure 3 (right)). A fusion net represents an entity with a concatenation of its entity feature vector and visual context vector. The visual context vector is computed with an attention mechanism, indicating to which part of the image should be paid attention, in order to judge whether the entity pair is VGP or not. The outputs of the two fusion nets are then fed into a multilayer perceptron (MLP) to compute the similarity of the two entities.

Formally, let X be a 196×512 feature map[*5] extracted from the conv5_3 layer in the VGG-16 network [40] for an input image; $\mathbf{x}_n$ is a 512 dimensional vector at position n of X. Given an entity feature vector $\mathbf{v}_i$ and $\mathbf{x}_n$, we first transform them with fully connected (FC) layers whose unit sizes are 512:

$$\tilde{\mathbf{x}}_n = \mathrm{norm}_{\mathrm{L2}}(W_v \mathbf{x}_n + \mathbf{b}_v) \quad (8)$$

$$\tilde{\mathbf{v}}_i = \mathrm{norm}_{\mathrm{L2}}(W_p \mathbf{v}_i + \mathbf{b}_p) \quad (9)$$

where $\mathrm{norm}_{\mathrm{L2}}(\cdot)$ indicates L2 normalization to an input vector. We then compute an attention value a_n for $\mathbf{x}_n$ as:

$$\mathbf{h}_n = \mathrm{relu}(\tilde{\mathbf{x}}_n + \tilde{\mathbf{v}}_i) \quad (10)$$

$$e_n = \mathbf{w}^\top \mathbf{h}_n \quad (11)$$

$$a_n = \frac{\exp(e_n)}{\sum_{n=1}^{N} \exp(e_n)} \quad (12)$$

where $N = 196$. After obtaining a_n, we fuse a visual and an entity feature vector to $\mathbf{y}_i$ as:

$$\mathbf{c} = \sum_{n=1}^{N} a_n \mathbf{x}_n \quad (13)$$

$$\mathbf{y}_i = U[\mathrm{norm}_{\mathrm{L2}}(\mathbf{c}), \tilde{\mathbf{v}}_i] + \mathbf{d} \quad (14)$$

where $[\cdot, \cdot]$ indicates the concatenation of two vectors, $\mathbf{c}$ is a visual context vector. We compute fusion feature vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ with the corresponding image. Finally, we feed them to a two-layer MLP network with ReLU non-linearity, whose unit sizes are 128 and 1, respectively, to produce the similarity of the entity pair.

6. Experiments

6.1 Settings

We conducted experiments on the Flickr30k entities dataset [36]. This dataset contains 31,837 images, which is described with 5 captions annotated via crowdsourcing. We followed the 29,873 training, 1,000 validation, and 1,000 test image splits used in the phrase localization task [36]. Our task is to automatically cluster the entities in the captions that describe the same visual concept (i.e., region in the dataset) in the image as VGPs. Entities that share the same ID and group type (e.g., "a red jersey," "a red shirt" and "a red uniform shirt" in Figure 1 share the same entity ID and group type "/EN#19026/clothing") are treated as the ground truth VGP clusters in our evaluation.[*6] As stop words should not be considered for computing the entity similarities, we preprocessed the entities in the dataset by removing stop words for all the methods.

We evaluated both clustering and pairwise performance. The entity clustering performance for each image was measured with adjusted Rand index (ARI) [22]. We used the implementation in the Scikit-learn machine learning toolkit [43][*7] for computing ARI. We report the mean of ARI scores for all the images in the test split. To evaluate the performance for clustering, we optimized the number of clusters by adjusting the preference for affinity propagation on the validation split to maximize the ARI using the Bayesian optimization algorithm [32] implemented in GPyOpt.[*8] The pairwise performance was evaluated with precision, recall, and F-score, defined as:

[*5] An image is split into $14 \times 14 = 196$ sub-images, and represented as a 196×512 feature map.

[*6] There is an entity type named "notvisual" in the dataset (e.g., "the batter" in Figure 1), which means this entity has no corresponding visual regions in the image. In our evaluation, we excluded this "notvisual" type, because all entities that are not visual are annotated with the same entity ID and thus ground truth VGPs for these "notvisual" entities are unavailable in the dataset. There are entity pairs in the dataset that are the same after removing the stop words (e.g., "a man" and "the man"), we treated them as one entity for evaluation. In addition, entities that do not have corresponding regions in the image were excluded from evaluation.

[*7] http://scikit-learn.org/stable/modules/clustering.html
#adjusted-rand-index

[*8] https://github.com/SheffieldML/GPyOpt

$$\text{precision} = \frac{\#\text{predicted\_positive}}{\#\text{predicted}} \tag{15}$$

$$\text{recall} = \frac{\#\text{predicted\_positive}}{\#\text{true\_positive}} \tag{16}$$

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{17}$$

where an entity pair with a similarity higher than a threshold is treated as *predicted*, which is compared against the ground truth to judge whether it is *predicted_positive* or not. We report the performance using the similarity threshold tuned on the validation split that maximizes the F-score.

We used the affinity propagation implementation[*9] in Scikit-learn for clustering. We compared the performance of the different similarity methods described in Sections 4 and 5, where the detailed settings for the methods were as follows:

- Phrase localization (PL): we used the pl-clc toolkit,[*10] which is an implementation of the localization method of [35]. For R in equation 3, we used the top 30 localization candidates for each entity. The localization scores for each entity and region pair obtained with [35] were used to compute the similarity.
- Translation probability (TP): to get the entity translation probabilities, we first applied the GIZA++ toolkit[*11] that is an implementation of the IBM alignment models [9] on the pseudo parallel corpus, and then a phrase table was extracted and the phrasal translation probabilities were calculated using state-of-the-art SMT toolkit Moses [25].
- Word embedding average (WEA): see the detailed setting in Section 4.3.1.
- Fisher vector (FV): entity feature vectors were computed using the Fisher vector toolkit released by the authors,[*12] following the settings described in Section 4.3.2.
- Fisher vector w/ CCA (FV+CCA): image region feature vectors and entity feature vectors were projected into a 4,096 dimensional space CCA trained on the training split of the Flickr30k entity dataset (Section 4.3.3).
- Supervised NN (SNN): to show the effectiveness of the fusion net (Section 5), we compared a supervised NN-based setting that only feeding the entity feature vectors to the MLP (Figure 3 (left)) for paraphrase similarity prediction. This setting only uses entity feature vectors as input for the NN. It was trained on the training split of the Flickr30k entity dataset. We used all the ground truth VGP pairs in the training split as positive instances. During training, we constructed mini-batches with 15% of positive instances and 85% of randomly sampled negative instances. We used Adam for optimization with a mini-batch size of 300 and weight decay of 0.0001. The learning rate was initialized to 0.01, which was halved at every epoch. We terminated training after 5 epochs, where we observed the loss converged on the validation split. For the entity feature vectors, we compared three different settings described above namely: WEA, FV, and FV+CCA.
- SNN+image: this setting is for our proposed supervised NN-based method described in Section 5. We again compared the three different entity feature vectors. We used VGG-16 [40] for the image features. The model was trained with the same configuration as the SNN setting.
- Ensemble: the ensemble of the SNN and SNN+image models that takes the average similarity given by both models. The motivation of this setting is to complement these two models to each other.

6.2 Results

Table 1 shows the results of all the different methods. We report the performance based on the entity types to better understand the performance difference of each method, i.e., "all" evaluates on all entities, whereas "single" and "multi" only evaluate on entities with one single token and multiple tokens, respectively, after removing stop words. For the unsupervised methods, we can see that PL does not show good performance. This is due to the low performance of phrase localization.[*13] TP shows a fairly high F-score, but a very low ARI score. The reason for this is that the translation probabilities are computed based on word alignment, leading to a similarity score of 0 to the unrelated entity pairs, which is not suitable for affinity propagation. WEA shows relatively good performance that is better than FV. This is because 45.84% of the entities in our task are single word type after removing the stop words, and converting the low dimensional word embedding to high dimensional and sparse Fisher vectors is harmful for these single word entity pairs. However, for the performance of entities containing multiple words, the Fisher vector is better than word embedding average in the perspective of F-score. FV+CCA significantly outperforms FV. This is because it uses visual information in the training split that transforms the entity vectors and visual vectors into the semantic space that is helpful for detecting VGPs.

Regarding the supervised methods, NN-based methods using any entity feature vectors outperforms the methods that uses them in an unsupervised way. The reason for this is that it directly uses the paraphrase supervision in the training split, while the unsupervised methods do not. Using entity representation with better ARI and F-score for the SNN method can achieve better results. The performance improvement by SNN on FV is not as large as WEA and FV+CCA, and we suspect the reason for this is the sparseness of the Fisher vectors. Our proposed method (SNN+image) that uses both textual and visual features shows better performance compared to SNN that uses textual features only, indicating that the usage of visual features is helpful for our VGP extraction task. However, the performance improvements are not very large. We discuss the reason for this in detail in Section 6.3.1. The ensemble of SNN and SNN+image further improves the performance, which means that these two models complement each other.

[*9] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

[*10] https://github.com/BryanPlummer/pl-clc

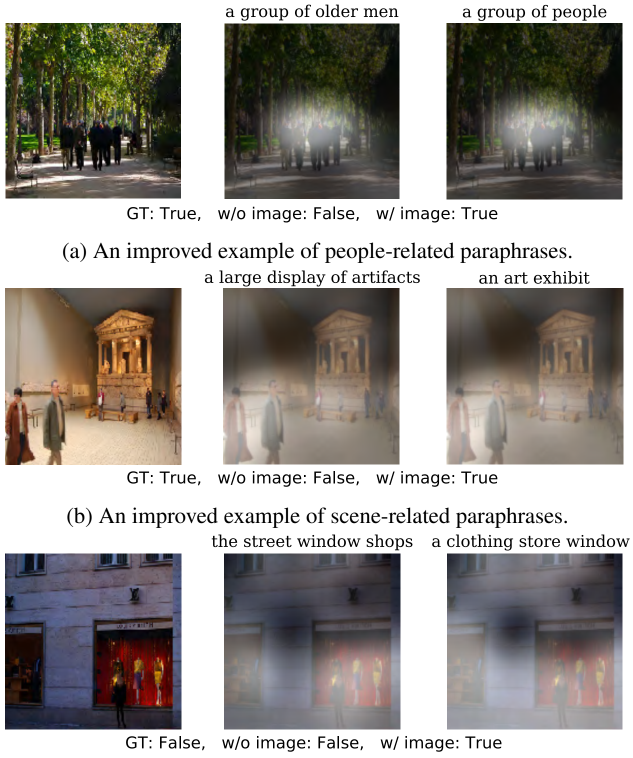
[*11] http://code.google.com/p/giza-pp

[*12] https://owncloud.cs.tau.ac.il/index.php/s/vb7ys8Xe8J8s8vo

[*13] Although [35] is the current state-of-the-art for phrase localization, the accuracy is only 55.85%.

	ARI	Precision	Recall	F-score
Method	all / single / multi	all / single / multi	all / single / multi	all / single / multi
PL	43.30 / 45.92 / 46.35	59.32 / 51.53 / 62.86	63.12 / 47.99 / 74.14	61.16 / 49.70 / 68.04
TP	37.61 / 50.50 / 36.79	66.23 / 63.20 / **82.17**	64.20 / 66.10 / 56.31	65.20 / 64.62 / 66.83
WEA	49.55 / 48.48 / 49.31	62.95 / 46.15 / 62.77	69.67 / 67.04 / 79.23	66.14 / 54.66 / 70.05
FV	45.42 / 43.55 / 41.80	66.60 / 37.23 / 67.89	58.59 / 31.32 / 77.05	62.34 / 34.02 / 72.18
FV+CCA	54.97 / 51.84 / 50.76	64.79 / 55.79 / 68.24	82.20 / 75.83 / 84.98	72.46 / 64.28 / 75.69
SNN (WEA)	60.44 / 55.06 / 53.26	77.86 / 83.66 / 74.50	84.58 / 75.16 / **88.96**	81.08 / 79.18 / 81.09
SNN+image (WEA)	60.55 / 55.42 / **55.82**	79.47 / 81.01 / 77.26	84.56 / 79.35 / 87.06	81.94 / 80.17 / 81.86
Ensemble (WEA)	61.04 / 55.02 / 54.83	80.65 / 78.68 / 77.38	84.79 / 83.14 / 88.85	82.67 / 80.85 / 82.72
SNN (FV)	48.13 / 46.04 / 47.22	64.21 / 45.92 / 66.40	65.93 / 50.89 / 76.51	65.06 / 48.28 / 71.10
SNN+image (FV)	48.00 / 47.83 / 48.31	63.49 / 52.62 / 66.86	68.20 / 55.62 / 78.01	65.76 / 54.08 / 72.01
Ensemble (FV)	50.14 / 49.86 / 48.25	65.48 / 54.87 / 70.51	71.43 / 56.24 / 76.54	68.33 / 55.55 / 73.40
SNN (FV+CCA)	60.68 / 56.58 / 54.04	**83.11 / 85.19** / 77.44	82.13 / 79.30 / 87.69	82.62 / 82.14 / 82.25
SNN+image (FV+CCA)	61.56 / 54.86 / 54.14	82.51 / 84.52 / 80.28	84.19 / 81.85 / 86.82	83.34 / 83.16 / 83.43
Ensemble (FV+CCA)	**62.42 / 56.83** / 54.86	82.71 / 84.10 / 80.91	**85.67 / 83.50** / 87.06	**84.16 / 83.80 / 83.87**

Table 1: VGP extraction results ("all" evaluates on all entities, "single" and "multi" only evaluate on entities consist of one single token and multiple tokens after removing stop words, respectively; the methods above and below the double line are unsupervised and supervised, respectively).


a group of older men
a group of people
GT: True, w/o image: False, w/ image: True
a large display of artifacts
an art exhibit
GT: True, w/o image: False, w/ image: True
the street window shops
a clothing store window
GT: False, w/o image: False, w/ image: True

(a) An improved example of people-related paraphrases.

(b) An improved example of scene-related paraphrases.

(c) A worsened example of scene-related paraphrases.

Fig. 4: Examples comparing SNN (FV+CCA) with SNN+image (FV+CCA) (the leftmost images are the original ones, the images in the middle and on the right show attention of the entity pairs on the images, the degree of whiteness indicates the strength of attention, the identification results are shown under the images).

6.3 Discussion

6.3.1 Neural Network w/ and w/o Images

We compared the SNN and SNN+image results, and found that image attention is helpful for identifying people-related paraphrases, which are difficult to be determined based on the textual information only. In addition, the attention for these people-related paraphrases are well learned. We believe the reason for this is that many entities in the training split are people-related and thus they are well modeled. Figure 4a shows such an example, where the SNN (FV+CCA) model fails to identify these two entities "a group of order men" and "a group of people" as VGPs due to the diverse textual descriptions of the the same visual concept. Our proposed NN+image (FV+CCA) model correctly identifies these VGPs by paying attention to the image region of people in the image. In some cases, the visual information is also helpful for the identification of other types of paraphrases, although the attention is not accurate. Figure 4b shows an example, where the SNN (FV+CCA) model could not identify two entities "a large display of artifacts" and "an art exhibit" as VGPs.

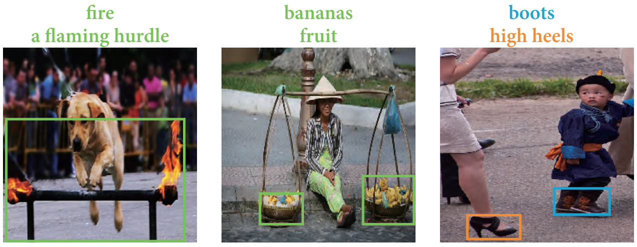


Fig. 5: Failed examples.

In some cases, visual information could bring negative effects for paraphrase identification. Figure 4c shows an example that "the street window shops" and "a clothing store window" are mistakenly judged as a paraphrase after using the image information while using textual information judges correctly. Although, the attention for these entities refer to the same visual concept in the image, the entities actually refer to different concepts (i.e., "shop" and "window").

6.3.2 Failed Examples

Even the best method, namely Ensemble (FV+CCA), only achieves a ARI of 62.42 and a F-score of 84.16. We found that most false negative examples are sparse entity pairs that describe a image region in an image in a very diverse way, for example "fire" and "a flaming hurdle" (Figure 5 (left)), "bananas" and "fruit" (Figure 5 (middle)). These pairs are difficult not only for using textual features, but also for using image attentions. Most false positive examples are produced by the noisy phrase embedding method. For example, "boots" and "high heels" referring to the shoes on a boy and a lady, respectively, are identified as a paraphrase pair because of their closeness in the embedding space

(Figure 5 (right)). Some of the false positive examples are caused by the noise introduced by the wrong attention in an image. For example, "a green snowman" and "his new toy" are attended to the similar image regions.

7. Conclusion

In this paper, we proposed a novel task to extract VGPs describing the same visual concept in an image. We not only applied various existing techniques for this task, but also proposed a NN-based method that uses both the textual and visual information to model the similarity between the VGPs. Experiments on the Flickr30k entities dataset showed that we achieved a good performance.

For future work, we plan to study a multi task method for both VGP extraction and phrase localization to further improve the performance. Extracting other types of paraphrases (e.g., prepositional and verb paraphrases) is another possible extension, which requires a much deeper understanding of the relation between phrases and image regions. We also plan to apply the VGPs for CV and NLP multimodal tasks, such as VQA.

Acknowledgments

This work was supported by ACT-I, JST.

References

[1] Androutsopoulos, I. and Malakasiotis, P.: A Survey of Paraphrasing and Textual Entailment Methods, *Journal of Artificial Intelligence Research*, Vol. 38, No. 1, pp. 135–187 (online), available from ⟨http://dl.acm.org/citation.cfm?id=1892211.1892215⟩ (2010).

[2] Arase, Y. and Tsujii, J.: Monolingual Phrase Alignment on Parse Forests, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 1–11 (online), available from ⟨https://www.aclweb.org/anthology/D17-1001⟩ (2017).

[3] Bannard, C. and Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 597–604 (online), DOI: 10.3115/1219840.1219914 (2005).

[4] Barzilay, R. and Lee, L.: Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment, *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Association for Computational Linguistics, pp. 16–23 (2003).

[5] Berant, J. and Liang, P.: Semantic Parsing via Paraphrasing, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, Association for Computational Linguistics, pp. 1415–1425 (online), available from ⟨http://www.aclweb.org/anthology/P14-1133⟩ (2014).

[6] Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. and Plank, B.: Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, *Journal of Artificial Intelligence Research*, Vol. 55, No. 1, pp. 409–442 (online), available from ⟨http://dl.acm.org/citation.cfm?id=3013558.3013571⟩ (2016).

[7] Bhagat, R. and Hovy, E.: What Is a Paraphrase?, *Computational Linguistics*, Vol. 39, No. 3, pp. 463–472 (2013).

[8] Bhagat, R. and Ravichandran, D.: Large Scale Acquisition of Paraphrases for Learning Surface Patterns, *Proceedings of the 46rd Annual Meeting of the Association for Computational Linguistics: the Human Language Technology Conference*, Columbus, Ohio, Association for Computational Linguistics, pp. 674–682 (online), available from ⟨http://www.aclweb.org/anthology/P/P08/P08-1077⟩ (2008).

[9] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263–312 (1993).

[10] Callison-Burch, C.: Syntactic Constraints on Paraphrases Extracted from Parallel Corpora, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, Association for Computational Linguistics, pp. 196–205 (online), available from ⟨http://www.aclweb.org/anthology/D08-1021⟩ (2008).

[11] Chen, D. and Dolan, W.: Collecting Highly Parallel Data for Paraphrase Evaluation, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Association for Computational Linguistics, pp. 190–200 (online), available from ⟨http://www.aclweb.org/anthology/P11-1020⟩ (2011).

[12] Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P. and Zitnick, C. L.: Microsoft COCO Captions: Data Collection and Evaluation Server, *CoRR*, Vol. abs/1504.00325 (online), available from ⟨http://arxiv.org/abs/1504.00325⟩ (2015).

[13] Chu, C. and Kurohashi, S.: Paraphrasing Out-of-Vocabulary Words with Word Embeddings and Semantic Lexicons for Low Resource Statistical Machine Translation, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, PortoroÅ, Slovenia, European Language Resources Association (ELRA), pp. 644–648 (2016).

[14] Frey, B. J. and Dueck, D.: Clustering by Passing Messages Between Data Points, *Science*, Vol. 315, No. 5814, pp. 972–976 (2007).

[15] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 457–468 (online), available from ⟨https://aclweb.org/anthology/D16-1044⟩ (2016).

[16] Funaki, R. and Nakayama, H.: Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 585–590 (online), available from ⟨http://aclweb.org/anthology/D15-1070⟩ (2015).

[17] Ganitkevitch, J. and Callison-Burch, C.: The Multilingual Paraphrase Database, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, European Language Resources Association (ELRA), pp. 4276–4283 (2014).

[18] Gong, Y., Ke, Q., Isard, M. and Lazebnik, S.: A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics, *International Journal of Computer Vision*, Vol. 106, No. 2, pp. 210–233 (online), DOI: 10.1007/s11263-013-0658-4 (2014).

[19] Han, D., Martínez-Gómez, P. and Mineshima, K.: Visual Denotations for Recognizing Textual Entailment, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 2843–2849 (online), available from ⟨https://www.aclweb.org/anthology/D17-1304⟩ (2017).

[20] Hardoon, D. R., Szedmak, S. R. and Shawe-taylor, J. R.: Canonical Correlation Analysis: An Overview with Application to Learning Methods, *Neural Computation.*, Vol. 16, No. 12, pp. 2639–2664 (online), DOI: 10.1162/0899766042321814 (2004).

[21] Harris, Z. S.: Distributional structure, *Word*, Vol. 10, No. 23, pp. 146–162 (1954).

[22] Hubert, L. and Arabie, P.: Comparing partitions, *Journal of Classification*, Vol. 2, No. 1, pp. 193–218 (online), DOI: 10.1007/BF01908075 (1985).

[23] Kajiwara, T., Komachi, M. and Mochihashi, D.: MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting, *Proceedings of the 8th International Joint Conference on Natural Language Processing*, Taipei, Taiwan, Asian Federation of Natural Language Processing, pp. 80–89 (2017).

[24] Klein, B., Lev, G., Sadeh, G. and Wolf, L.: Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation, *CoRR*, Vol. abs/1411.7399 (online), available from ⟨http://arxiv.org/abs/1411.7399⟩ (2014).

[25] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W. ., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, Association for Computational Linguistics, pp. 177–180 (online), available from ⟨http://www.aclweb.org/anthology/P/P07/P07-2045⟩ (2007).

[26] Laokulrat, N., Phan, S., Nishida, N., Shu, R., Ehara, Y., Okazaki, N., Miyao, Y. and Nakayama, H.: Generating Video Description using Sequence-to-sequence Model with Temporal Attention, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, The COLING 2016 Organizing Committee, pp. 44–52 (online), available from ⟨http://aclweb.org/anthology/C16-1005⟩ (2016).

[27] Lin, D. and Pantel, P.: DIRT – Discovery of Inference Rules from Text, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, New York, NY, USA, ACM, pp. 323–328 (online), DOI: 10.1145/502512.502559 (2001).

[28] Ling, W., Dyer, C., Black, A. W. and Trancoso, I.: Paraphrasing 4 Microblog Normalization, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, Association for Computational Linguistics, pp. 73–84 (online), available from ⟨http://www.aclweb.org/anthology/D13-1008⟩ (2013).

[29] MacCartney, B., Galley, M. and Manning, C. D.: A Phrase-Based Alignment Model for Natural Language Inference, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, Association for Computational Linguistics, pp. 802–811 (online), available from ⟨http://www.aclweb.org/anthology/D08-1084⟩ (2008).

[30] Marton, Y., Callison-Burch, C. and Resnik, P.: Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, Association for Computational Linguistics, pp. 381–390 (online), available from ⟨http://www.aclweb.org/anthology/D/D09/D09-1040⟩ (2009).

[31] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781 (online), available from ⟨http://arxiv.org/abs/1301.3781⟩ (2013).

[32] Mockus, J.: *Bayesian approach to global optimization: theory and applications*, Mathematics and its applications: Soviet series, Kluwer Academic (1989).

[33] Otani, M., Nakashima, Y., Rahtu, E., Janne, H. and Yokoya, N.: Learning joint representations of videos and sentences with web image search, *European Conference on Computer Vision (ECCV)*, pp. 651–667 (2016).

[34] Phan, S., Miyao, Y., Le, D.-D. and Satoh, S.: Video Event Detection by Exploiting Word Dependencies from Image Captions, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, The COLING 2016 Organizing Committee, pp. 3318–3327 (online), available from ⟨http://aclweb.org/anthology/C16-1313⟩ (2016).

[35] Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J. and Lazebnik, S.: Phrase Localization and Visual Relationship Detection With Comprehensive Image-Language Cues, *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1928–1937 (2017).

[36] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J. and Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649 (2015).

[37] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in Neural Information Processing Systems 28* (Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. and Garnett, R., eds.), Curran Associates, Inc., pp. 91–99 (2015).

[38] Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V. and Liu, Y.: Statistical Machine Translation for Query Expansion in Answer Retrieval, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, Association for Computational Linguistics, pp. 464–471 (online), available from ⟨http://www.aclweb.org/anthology/P07-1059⟩ (2007).

[39] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T. and Schiele, B.: Grounding of Textual Phrases in Images by Reconstruction, *European Conference on Computer Vision (ECCV)*, pp. 817–834 (2016).

[40] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recoginition, *International Conference on Learning Representations (ICLR)*, pp. 1–14 (2015).

[41] Snover, M. G., Madnani, N., Dorr, B. and Schwartz, R.: TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate, *Machine Translation*, Vol. 23, No. 2-3, pp. 117–127 (online), DOI: 10.1007/s10590-009-9062-9 (2009).

[42] Specia, L., Frank, S., Sima'an, K. and Elliott, D.: A Shared Task on Multimodal Machine Translation and Crosslingual Image Description, *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, Association for Computational Linguistics, pp. 543–553 (online), available from ⟨http://www.aclweb.org/anthology/W16-2346⟩ (2016).

[43] Thirion, B., Duschenay, E., Michel, V., Varoquaux, G., Grisel, O., VanderPlas, J., alexandre granfort, fabian pedregosa, Mueller, A. and Louppe, G.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).

[44] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D.: Show and Tell: A Neural Image Caption Generator, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164 (2015).

[45] Wang, L., Li, Y. and Lazebnik, S.: Learning Deep Structure-Preserving Image-Text Embeddings, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5005–5013 (2016).

[46] Wang, M., Azab, M., Kojima, N., Mihalcea, R. and Deng, J.: Structured Matching for Phrase Localization, *European Conference on Computer Vision (ECCV)*, pp. 696–711 (2016).

[47] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. and Hengel, A. v. d.: Visual question answering: A survey of methods and datasets, *Computer Vision and Image Understanding*, pp. 1–20 (online), available from ⟨http://dx.doi.org/10.1016/j.cviu.2017.05.001⟩ (2017).

[48] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, Lille, France, PMLR, pp. 2048–2057 (online), available from ⟨http://proceedings.mlr.press/v37/xuc15.html⟩ (2015).

[49] Young, P., Lai, A., Hodosh, M. and Julia, H.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association of Computational Linguistics*, Vol. 2, No. 1, pp. 67–78 (2014).

[50] Zhou, L., Lin, C.-Y., Munteanu, D. S. and Hovy, E.: ParaEval: Using Paraphrases to Evaluate Summaries Automatically, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Main Conference*, New York City, USA, Association for Computational Linguistics, pp. 447–454 (online), available from ⟨http://www.aclweb.org/anthology/N/N06/N06-1057⟩ (2006).