Using Object Detection, NLP, and Knowledge Bases to Understand the Message of Images

Lydia Weiland[1(✉)], Ioana Hulpus[1], Simone Paolo Ponzetto[1], and Laura Dietz[2]

[1] University of Mannheim, Mannheim, Germany
lydia@informatik.uni-mannheim.de
[2] University of New Hampshire, Durham, NH, USA

Abstract. With the increasing amount of multimodal content from social media posts and news articles, there has been an intensified effort towards conceptual labeling and multimodal (topic) modeling of images and of their affiliated texts. Nonetheless, the problem of identifying and automatically naming the core abstract message (*gist*) behind images has received less attention. This problem is especially relevant for the semantic indexing and subsequent retrieval of images. In this paper, we propose a solution that makes use of external knowledge bases such as Wikipedia and DBpedia. Its aim is to leverage complex semantic associations between the image objects and the textual caption in order to uncover the intended gist. The results of our evaluation prove the ability of our proposed approach to detect gist with a best MAP score of 0.74 when assessed against human annotations. Furthermore, an automatic image tagging and caption generation API is compared to manually set image and caption signals. We show and discuss the difficulty to find the correct gist especially for abstract, non-depictable gists as well as the impact of different types of signals on gist detection quality.

1 Introduction

Recently, much work in image and language understanding has led to interdisciplinary contributions that bring together processing of visual data such as video and images with text mining techniques. Because text and vision provide complementary sources of information, their combination is expected to produce better models of understanding semantics of human interaction [3] therefore improving end-user applications [25].

The joint understanding of vision and language data has the potential to produce better indexing and search methods for multimedia content [11,31]. Research efforts along this line of work include image-to-text [10,14,22,36] and video-to-text [1,5,20] generation, as well as the complementary problem of associating images or videos to arbitrary texts [4,11]. Thus, most previous work concentrates on the recognition of visible objects and the generation of literal, descriptive caption texts. However, many images are used with the purpose of stimulating emotions [26,27], e.g., the image of polar bears on shelf ice. To understand the message behind such images, typically used for writing about complex

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 405–418, 2017.
DOI: 10.1007/978-3-319-51814-5_34

A male orangutan waits near a feeding station at Camp Leakey in Tanjung Puting National Park in Central Kalimantan province, Indonesia in this June 15, 2015.

Gist examples: *Mammals of Southeast Asia, Trees, Orangutans, Botany, Plants*

(a) Literal Pairing

Fight to save Indonesia's wildlife corridors key for endangered orangutan.

Gist examples: *Habitat, Conservation, Biodiversity, Extinction, EDGE species, Deforestation*

(b) Non-literal Pairing

Fight to save Indonesia's wildlife corridors key for endangered orangutan.

Gist examples: *Habitat Conservation, Biodiversity, Extinction, Politics, Protest*

(c) Non-literal Pairing

Fig. 1. Example image-caption pairs sharing either images or captions with their respective gist entities (a, b: http://reut.rs/2cca9s7, REUTERS/Darren Whiteside, c: http://bit.ly/2bGsvii, AP, last accessed: 08/29/2016.

topics like global warming or financial crises, semantic associations must be exploited between the depictable, concrete objects of the image and the potential abstract topics. Current knowledge bases such as Wikipedia, DBpedia, FreeBase can fill this gap and provide these semantic connections. Our previous work [35] introduces such a system for image understanding that leverages such sources of external knowledge. The approach was studied in an idealized setting where humans provided image tags and created the object vocabulary in order to make design choices.

Contribution. Building on top of our previous work, a core contribution of this paper is to study whether the performance of gist detection with external knowledge is impacted when an automatic object detector is used instead of human annotations. We make use of the Computer Vision API[1] from Microsoft Cognitive Services [8] - a web service that provides a list of detected objects and is also capable of generating a descriptive caption of the image. This way, we create a fully automatic end-to-end system for understanding abstract messages conveyed through association such as examples of Fig. 1(b) and (c).

Microsoft Cognitive Service uses a network pre-trained on ImageNet [6]. Additionally it includes a CNN, which is capable of assigning labels to image regions [13], and trains on Microsoft COCO [23] data, thus, resulting in a vocabulary of 2,000 object categories. Together with Microsoft's language generation API, this information is used to generate a caption for an image, that did not have one before.

Our task setup defines the understanding of abstract messages as being able to describe this message with appropriate concepts from the knowledge base - called gist nodes. This way we cast the problem as an entity ranking problem

[1] https://www.microsoft.com/cognitive-services/en-us/computer-vision-api.

which is evaluated against a human-generated benchmark of relevant entities and categories.

We study the effects of automatic object detection separately for images with literal descriptive captions and captions that utilize a non-literal (i.e., abstract) meaning. While in both cases reasonable performance is obtained by our approach, experiments point towards room for improvement for object detection algorithms. We identify theoretical limits by analyzing which of the gist nodes represent depictable (e.g. tree) versus non-depictable (e.g., philosophy) concepts. These limits are complemented with experiments considering the signal from either the image or the caption as well as automatically generated captions from the Microsoft API, which ignore the original caption. We demonstrate that understanding the message of non-literal image-caption pairs is a difficult task (unachievable by ignoring the original caption) to which our approach together with MS Cognitive Services provides a large step in the right direction.

2 Problem Statement

We study the problem of identifying the gist expressed in an image-caption pair in the form of an entity ranking task. The idea is that general-purpose knowledge bases such as Wikipedia and DBpedia provide an entry for many concepts, ranging from people and places, to general concepts as well as abstract topics such as "philosophy". Some of these entries represent **depictable** objects, such as "bicycle", "solar panel", or "tree", some could be associated with visual features such as "arctic landscape" or "plants". The task is to identify (and rank) the most relevant concepts (e.g., entities or categories from Wikipedia) that describe the gist of the image-caption pair.

Problem Statement. Given an image with its respective caption as inputs, predict a ranking of concepts from the knowledge base that best represent the core message expressed in the image.

By predicting the most prominent gist of an image-caption pair, these can be indexed by a search engine and provide diverse images in response to concept queries. Our work provides a puzzle-piece in answering image queries also in response to **non-depictable** concepts such as "biodiversity" or "endangered species".

We distinguish two types of image-caption pairs: **Literal pairs**, where the caption describes what is seen on the image. In such cases the gist of the image is often a depictable concept. In contrast, in **non-literal pairs**, image and caption together allude to an abstract theme. These are often non-depictable concepts. Figure 1 displays three examples on the topic of endangered species from both classes with image, caption, and a subset of annotated gist nodes from the knowledge base. The example demonstrates how changing the picture or caption can drastically change the message expressed.

We devise a supervised framework for gist detection that is studied in the context of both styles of image-caption pairs. In particular, we center our study on the following research questions:

RQ0: What is the fraction of depictable concept?
RQ1: Does an automatic image tagging change the prediction quality?
RQ2: Does an automatic caption generation change the prediction quality?
RQ3: Would an automatic approach capture more literal or more non-literal aspects?
RQ4: What is the benefit of joint signals (in contrast to only caption or only image)?

3 Related Work

Especially with the increasing amount of multi-modal datasets, the joint modeling of cross-modal features has gained attention. Different combinations of modalities (audio, image, text, and video) are possible, we focus on those mostly related to our research. Those datasets consist of images with captions and/or textual labeled object regions, e.g., Flickr8k [30] and Flickr30k [37], SBU Captioned Photo Dataset [28], PASCAL 1k dataset [9], ImageNet [21], and Microsoft Common Objects in Context (COCO) [23].

Joint modeling of image and textual components, which utilizes KCCA [15,32] or neural networks [19,33], have shown to outperform single modality approaches. Independent from joint or single modeling, the applications are similar, e.g., multimodal topic generation [34] or retrieval tasks: Generating descriptions for images [7,14,22,29,36] and retrieving images for text [5,11]. The focus in these works lies on the generation of descriptive captions, semantic concept labeling, and depictable concepts [2,15,18,31], which results in literal pairs. In contrast, our approach benefits from external knowledge to retrieve and rank also abstract, non-depictable concepts for understanding both literal and non-literal pairs. Our previous study [35] was conducted on manually given image objects tags and captions. This work studies performance with an automatic object detection system.

4 Approach: Gist Detection

The main idea behind our approach is to use a knowledge base and the graph induced by its link structure to reason about connections between depicted objects in the image and mentioned concepts in the caption. Our basic assumption is that gist nodes may be directly referred in the image or caption or are in close proximity of directly referred concepts. To identify these gist nodes, we propose a graph mining pipeline, which mainly consists of a simple entity linking strategy, a graph traversal and expansion based on a relatedness measure, as shown in Fig. 2. Variations of the pipeline are studied in our prior work [35] but are detrimental to experiments in this work also. We clarify pipeline steps using the running example of a non-literal pair with a typical representative of endangered species in (Fig. 1(b)).

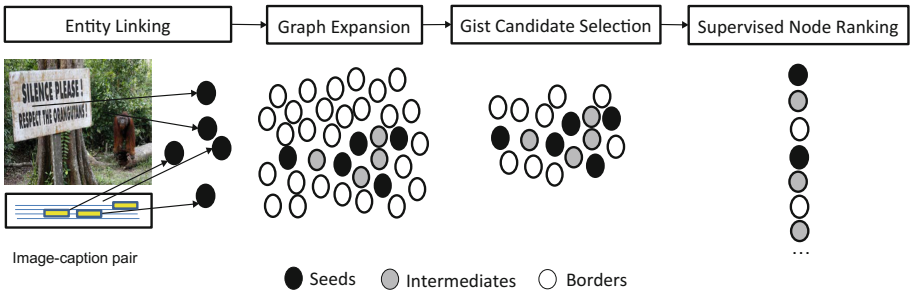

Fig. 2. Gist extraction and ranking pipeline for a given image-caption pair. For simplicity, we omit the edges betwell ween nodes in this figure.

4.1 The Knowledge Graph

Wikipedia provides a large general-purpose knowledge base about objects, concepts, and topics. Furthermore, and even more importantly for our approach, the link structure of Wikipedia can be exploited to identify topically associative nodes. DBpedia is a structured version of Wikipedia. All DBpedia concepts have their source in Wikipedia pages. In this work, our knowledge graph contains as nodes all the Wikipedia articles and categories. As for edges, we consider the following types of relations T, named by their DBpedia link property:

- **dcterms:subject.** The category membership relations that link an article to the categories it belongs to, e.g., Wildlife corridor dcterms:subject Wildlife conservation.
- **skos:broader.** Relationship between a category and its parent category in a hierarchical structure, e.g., Wildlife conservation skos:broader Conservation.
- **skos:narrower.** Relationship between a category and its subcategories, e.g., Conservation skos:narrower Water conservation.

4.2 Step 1: Entity Linking

The first step is to project the image tags from the objects detected in the image as well as the entities mentioned in the caption, e.g., wildlife corridors, onto nodes in the knowledge base, e.g., Wildlife corridor. To obtain the entities mentioned in the caption, we extract all the noun-phrases from the caption text. Each of these noun-phrases and image tags are then linked to entities in the knowledge base as follows: If the noun-phrase/image tag occurs in the knowledge base as an exact name (i.e., title of Wikipedia page, category, or redirect), this entry is selected as unambiguous. However, if it is the title of a disambiguation page, we select the disambiguation alternative with shortest connections to already projected unambiguous concepts (ignoring concepts with more than two hops). In the following, we refer to all the linked knowledge base entities as **seed nodes**.

4.3 Step 2: Graph Extraction

Our assumption is that the nodes representing the message best are not necessarily contained in the set of seed nodes, but lie in close proximity to them. Thus, we expand the seed node set to their neighborhood graph as follows: We activate all the seed nodes neighbors on a radius of n-hops, with n = 2 according to evidence from related work [16,24]. If any of the 2-hop neighbors lies on a shortest path between any two seed nodes, we call it an **intermediate node**, and further expand the graph around it on a radius of 2-hops. We name the resulting graph the **border graph** of the image-caption pair.

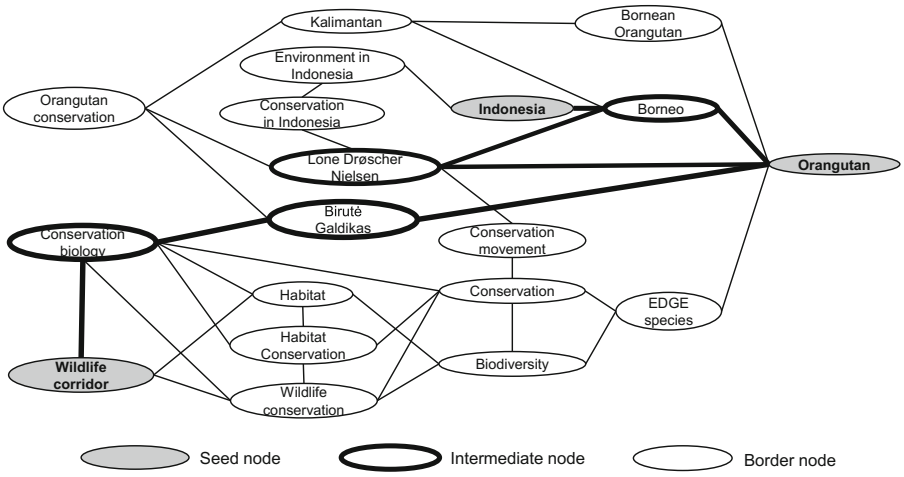

Fig. 3. Example of image-caption graph for the image-caption in Fig. 2.

Example. In Fig. 3, we show a graph excerpt of the image-caption graph extracted for the image-caption shown in Fig. 1(b). The linked seed nodes are Indonesia (as extracted from the caption), Orangutan (as extracted from the image and caption) and Wildlife corridor (as extracted from the caption). In this example, many suitable gist nodes are among the border nodes, i.e., Conservation and Orangutan Conservation.

4.4 Step 3: Gist Candidate Selection

After obtaining the border graph G, our aim is to select the nodes in this graph that are best candidates for capturing the message of the image-caption pair. We estimate for each node x its suitability as gist concept through its average semantic relatedness $\bar{\sigma}$ to all the seeds (S) and intermediate nodes (I), as shown in Formula 1.

$$\bar{\sigma}(x, G) = \frac{1}{|S \cup I|} \sum_{y \in S \cup I} \sigma(x, y) \tag{1}$$

In this paper, we use the symbol σ for representing any given relatedness measure. For calculating the relatedness we use the exclusivity-based relatedness measure of Hulpus et al. [17] with their hyperparameter settings ($\alpha = 0.25$, top-3 shortest paths). The measure for an edge is the higher the fewer alternative relations of the same edge type each of its endnode has.

Using the resulting score for each node in the border graph, we select as candidates the top-20 highest scoring nodes. These nodes are eventually ranked in a supervised manner, based on a selection of features that we present in the following step.

4.5 Step 4: Supervised Node Ranking

For each of the candidate nodes, a feature vector is created that is comprised of different measures of proximity between seed nodes and candidate nodes. The effectiveness of these features is studied within a learning-to-rank framework which ranks nodes by their utility of expressing the gist of the image-caption pair. Three boolean features {*Seed, Intermediate, Border*} indicate the expansion stage that included the node. A numerical feature refers to the semantic relatedness of the candidate nodes, to the seeds and intermediates. A retrieval-based feature is computed by considering the content of the Wikipedia articles that corresponds to the graph nodes. The feature aggregates object tags and the caption into a search query to retrieve knowledge based articles from a Wikipedia snapshot. We use a standard retrieval model called query-likelihood with Dirichlet smoothing. Using the entire graph of DBpedia described in Sect. 4.1 global node importance measures can be computed. These features are: (i) the indegree of the node which counts how many DBpedia entities point to the given entity; (ii) the clustering coefficient of the node which measures the ratio of neighbours of the node that are themselves directly connected. This feature was selected under the intuition that nodes denoting abstract concepts tend to have lower clustering coefficient, while nodes denoting specific concepts tend to have higher clustering coefficient; (iii) the node's PageRank as a very popular measure of node importance in graphs, which is also less computationally expensive on large graphs (DBpedia has approx 6 million nodes and 60 million edges) than other node centrality measures such as betweeness.

5 Experimental Evaluation

We begin our evaluation by studying whether the relevant gist nodes (scored 4 or 5 by human annotators) are in general depictable or not (**RQ 0**) by analyzing a subset of the gold standard. Research questions **RQ 1–4** (cf. Sect. 2) evaluate the end-to-end system using Microsoft Cognitive Service in combination with our approach presented in Sect. 4.

Dataset and Gold Standard. To conduct our evaluation we create a dataset[2] containing 328 pairs of images and captions with a balanced amount of literal

[2] dataset and gold standard: https://github.com/gistDetection/GistDataset.

and non-literal pairs (164 literal, 164 media-iconic pairs). The non-literal pairs are collected from news portals such as www.theguardian.com. The literal pairs use the same images as the non-literal pairs, but have a descriptive caption that is created by annotators. One of our goals is to be able to evaluate the proposed gist detection approach independently from automated object detection systems, thus annotators also manually assign labels to objects in the images.

The gold standard annotation of gist nodes is conducted by annotators selecting nodes from the knowledge-base representing the gist and assigning ranks to each of the nodes, on a Likert scale ranging from 0 (irrelevant) to 5 (core gist). Nodes with level 4 and 5 are referred to as *relevant gists*. For each pair there is only one gist which is annotated with level 5, following the assumption of having one *core gist* which represents the message best. On a subset, annotators also assessed whether a gist is depictable.

Experimental Setup. The feature set (cf. Sect. 4.5) is used in a supervised learning-to-rank framework (RankLib[3]). As ranking algorithm, we use Coordinate Ascent with a linear kernel. We perform 5-fold cross validation, with training optimized towards the target metric Mean Average Precision (MAP), which is a recall-oriented measure averaging the precision at those ranks where the recall changes. Besides MAP, the evaluations contain normalized discounted cumulative gain (NDCG@10), which is a graded retrieval measure, and Precision (P@10) of the top ten nodes from the ranked lists.

For every image, besides the original and manually set captions and object tags, we also evaluate automatically generated captions and image tags (in the following abbreviated with **MS tags** and **MS captions**, respectively) by using Computer Vision API from Microsoft Cognitive Services.

5.1 RQ 0: Relevant Gists: Depictable or Not?

We study whether gist concepts as selected by annotators tend to be depictable or non-depictable. For a subset of the gold standard pairs, annotators decided for all relevant gist concepts whether this concept is *depictable, not-depictable, or undecided*. On average the fraction of depictable core gists is 88% for literal pairs versus only 39% for the non-literal pairs. On the larger set of all relevant gists, 83% are depictable for literal pairs versus 40% for the non-literal pairs. The decision what an automatic image tagging system might be able to detect is difficult for humans, reflected in an inter-annotator agreement (Fleiss' kappa [12]) of $\kappa = 0.42$ for core gists and $\kappa = 0.73$ for relevant gists.

Discussion RQ 0. These results are in line with our initial assumption that literal pairs tend to have depictable concepts as gist, and the non-literal pairs have a predominant amount of non-depictable concepts as gist (cf. Sect. 2). This finding underlines the fact that the core message of images does not necessarily correspond to objects that are depicted in the image. This reinforces the need for approaches that are able to reason with semantic associations between depictable and abstract concepts.

[3] http://lemurproject.org/ranklib.php.

5.2 RQ 1: Manual vs. Automatic Image Tagging

To answer our first research question, we study the performance impact of using an automatic object detector for image tagging, **MS tags**, as opposed to manual tags (**tags**). In both cases we are using the original literal and non-literal captions (**captions**). We refer to the combination of MS tags with original captions as **the realistic, end-to-end approach** considering that images with captions are typically published without image object tags. We compare the performance difference on different stages: First, we note that the manual tags arise from 43 different entities with 640 instances over the complete dataset. The MS tags are from 171 different entities with 957 instances. There are 131 overlapping instances between manual and automatic tags, which amounts to less than one shared tag per image and 20% overlap over the complete dataset.

Second, we compare the performance of the manual and MS tags, both combined with the original captions (cf. Table 1). As expected, a higher performance is achieved with manual tags (Tags&Captions, MAP: 0.74), but the realistic approach achieves a reasonable quality as well (MAP: 0.43).

Discussion RQ 1. The overlap between MS and manual image tags is rather low (20%) and the detected concepts are not always correct (e.g., a polar bear is detected as herd of sheep). However, the MS tags in combination with the original captions achieve a reasonable ranking, which indicates the ability of automatic detectors to find relevant concepts and our method of being capable to handle certain levels of noise.

5.3 RQ 2: Manual vs. Automatic Caption Generation

Spinning the use of automatic detectors further, based on the detected objects, descriptive captions are generated and used as an alternative input. Doing so, we would ignore the original caption and ask, whether this is sufficient. We compare to the same stages as in RQ1. The manual captions use around 300 and 700 different entities (seed nodes) for the literal (l) and non-literal (nl) pairs, respectively. The MS caption results in 130 different entity nodes. 10% (l) and 3% (nl) of the instances overlap between the nodes from the manual and the MS captions across all image-caption pairs. However, the assessment of the non-literal pairs is restricted by the fact that automatic detectors are trained on models with a descriptive purpose.

In the following, we compare the manual captions to the MS captions within our approach. We combine each caption with the manual image tags and provide it as input for the pipeline described in Sect. 4. We study the combinations with respect to the complete dataset (cf. Table 1) and compare again to the pure manual input signals (MAP: 0.74). A better ranking than for the realistic approach can be achieved across all pairs (Tags & MS caption, MAP: 0.48).

In contrast to the strong results of the manual input signals, the pure automatic signals perform worse across all pairs (MAP: 0.14, cf. Table 1).

Discussion RQ 2. The overlap between MS and manual caption is low (3–10%), the MS captions are short, and the focus of the captions does not always match

Table 1. Ranking results (grade 4 or 5) according to different input signals and feature sets. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Both				Non-Literal				Literal			
	MAP	Δ%	NDCG@10	P@10	MAP	Δ%	NDCG@10	P@10	MAP	Δ%	NDCG@10	P@10
Tag& Caption	0.74	0.00	0.71	0.71	0.64	0.00	0.59	0.58	0.83	0.00	0.84	0.84
Tag&MS Caption	0.48	−34.75*	0.63	0.56	0.36	−44.44*	0.45	0.38	0.61	−27.33*	0.80	0.73
MS tags& Caption	0.43	−41.67*	0.58	0.53	0.40	−37.55*	0.49	0.44	0.46	−44.95*	0.68	0.61
MS tags&MS caption	0.14	−80.49*	0.28	0.23	0.09	−86.06*	0.17	0.14	0.20	−76.11*	0.39	0.32
Tags only	0.48	−36.79*	0.65	0.57	0.28	−47.25*	0.40	0.33	0.68	−28.29*	0.89	0.82
MS tags only	0.13	−84.02*	0.24	0.20	0.06	−89.50*	0.13	0.11	0.20	−79.83*	0.35	0.29
Caption only	0.38	−49.68*	0.54	0.49	0.31	−51.79*	0.40	0.35	0.45	−47.59*	0.67	0.63
MS caption only	0.07	−91.49*	0.15	0.12	0.05	−93.25*	0.10	0.08	0.09	−89.18*	0.19	0.16

the focus of the manual caption (e.g., example Fig. 1 receives the caption "There is a sign", without considering the orangutan, although it was detected as monkey by the automatic image tagging). However, the ranking of the combined automatic and manual approach with respect to the complete dataset performs reasonably well. This shows promising opportunities for using our approach together with semi-automatic image tagging and/or caption creation in a real-world pipeline. In the following, we study these results in more detail with respect to the distinction between literal and non-literal pairs.

5.4 RQ 3: Literal vs. Non-literal Aspect Coverage by Automatic Detector

Next, we study the input combinations with respect to the non-literal and literal pairs and compare again with the pure manual input (cf. Table 1). Analyzing MS tags&MS captions as input shows a moderate ranking for the literal pairs (MAP: 0.20). However, the performance for the non-literal pairs is bisected (MAP: 0.09). This result is expected because without any context, it is currently impossible for an automatic caption generator to recommend non-literal captions. The realistic approach has a performance decrease of less than 40% (MAP: 0.40 (nl), 0.46 (l)). Substituting the manual captions by the automatic captions results in an even better performance for the literal pairs, but a lower performance than with the realistic approach for the non-literal pairs (Tags&MS caption MAP: 0.36 (nl), 0.61 (l)).

Discussion RQ 3. The evaluation results across all input signal combinations confirm our intuition that gists of non-literal pairs are more difficult to detect. These non-literal pairs, however, are the ones found in news, blogs, and twitter,

which are our main interest. Automatic approaches can address descriptive pairs by detecting important objects in the image and describe those in the caption. However, the automatic approaches lack mentioning things that are salient to detect the gist of non-literal pairs. With respect to RQ 3 we have shown that pure automatic detectors achieve fair results for pairs where a descriptive output is wanted. A good performance of automatic object detectors is also achieved within the realistic approach. However, these results indicate that differentiating captions - which is currently done by setting the captions manually - is necessary to detect the gist.

5.5 RQ 4: Comparison of Single Signals vs. Signal Combination

Since experiments from the field of multi-modal modeling have demonstrated improvements by combining textual and visual signals, we study whether this effect also holds for our case—especially with respect to non-literal pairs (cf. Table 1). For a detailed analysis, we also study MS tags Only and MS captions Only.

Given the image tags as input signal only, the literal pairs - apart from ndcg - are nearly as good as using the combined signal as input (MAP: 0.68). In contrast, the non-literal pairs are worse than combining signals (Δ%: −47.25%). The MS tags have an informative content for the literal, but achieve only a fifth of the performance for the non-literal pairs compared to the manual input (MAP: 0.20 vs. 0.68 (l), 0.06 vs 0.28 (nl)). The same study is conducted on the captions as single input signal (Caption Only and MS caption Only). Interestingly, the caption only performs better than the image tags only for the non-literal pairs. Especially for non-literal pairs the results degrade significantly when the caption is replaced by a MS caption.

Discussion RQ 4. The results of Table 1 show that image-only signals cannot completely convey abstract and/or associative topics and thus, cannot fully address the requirements of non-literal pairs. However, these results prove also another hypothesis, that concrete objects which can be detected in the image, are important pointers towards the relevant gists. We remark that for both types of pairs the performance benefit from the combination of signals. Apart from the manual tags for the literal pairs, we conclude that the gist cannot be detected with only the caption or only the image signal.

6 Conclusion

Our aim is to understand the gist of image-caption pairs. For that we address the problem as a concept ranking, while leveraging features and further gist candidates from an external knowledge base. We compare manually to automatically gathered information created by automatic detectors. The evaluation is conducted on the complete test collection of 328 image-caption pairs, with respect

to the different input signals, signal combination, and single signal analysis. Furthermore, we study both, non-literal and literal pairs. Our finding is that combining signals from image and caption improves the performance for all types of pairs. An evaluation of inter-annotator agreement has shown that literal pairs in most of the cases have a depictable gist and non-literal pairs have a non-depictable gist. This analysis result is in line with the finding that non-literal more benefit from the (manual) caption signal, whereas literal more benefit from image signals. Within the realistic scenario, we test the performance of object detectors in the wild, which shows level for improvement of 10%.

Acknowledgements. This work is funded by the RiSC programme of the Ministry of Science, Research and the Arts Baden-Wuerttemberg, and used computational resources offered from the bwUni-Cluster within the framework program bwHPC. Furthermore, this work was in part funded through the Elitepostdoc program of the BW-Stiftung and the University of New Hampshire.

References

1. Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S.J., Fidler, S., Zhang, Z.: Video in sentences out. In: UAI, pp. 102–112 (2012)
2. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Plank, B.: Automatic description generation from images: a survey of models, datasets, and evaluation measures. arXiv preprint arXiv:1601.03896 (2016)
3. Bruni, E., Uijlings, J., Baroni, M., Sebe, N.: Distributional semantics with eyes: using image analysis to improve computational representations of word meaning. In: MM, pp. 1219–1228 (2012)
4. Das, P., Srihari, R.K., Corso, J.J.: Translating related words to videos and back through latent topics. In: WSDM, pp. 485–494 (2013)
5. Das, P., Xu, C., Doell, R.F., Corso, J.J.: A thousand frames in just a few words: lingual description of videos through latent topics and sparse object stitching. In: CVPR, pp. 2634–2641 (2013)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Elliott, D., Keller, F.: Image description using visual dependency representations. In: EMNLP, pp. 1292–1302 (2013)
8. Fang, H., Gupta, S., Iandola, F.N., Srivastava, R., Deng, L., Dollár, P., Zweig, G.: From captions to visual concepts and back. In: CVPR, pp. 1473–1482 (2015)
9. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 15–29. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_2
10. Feng, Y., Lapata, M.: How many words is a picture worth? Automatic caption generation for news images. In: ACL, pp. 1239–1249 (2010)
11. Feng, Y., Lapata, M.: Topic models for image annotation and text illustration. In: NAACL-HLT, pp. 831–839 (2010)
12. Fleiss, J., et al.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
13. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR (2013)

14. Gupta, A., Verma, Y., Jawahar, C.V.: Choosing linguistics over vision to describe images. In: AAAI, pp. 606–612 (2012)
15. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. IJCAI **47**, 853–899 (2013)
16. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised graph-based topic labelling using DBpedia. In: Proceedings of the WSDM 2013, pp. 465–474 (2013)
17. Hulpuş, I., Prangnawarat, N., Hayes, C.: Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9366, pp. 442–457. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25007-6_26
18. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & WordNet. In: MM, pp. 706–715 (2005)
19. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR, pp. 3128-3137. IEEE Computer Society (2015)
20. Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., Guadarrama, S.: Generating natural-language video descriptions using text-mined knowledge. In: AAAI (2013)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
22. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: understanding and generating image descriptions. In: CVPR, pp. 1601–1608 (2011)
23. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_48
24. Navigli, R., Ponzetto, S.P.: Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
25. Nikolaos Aletras, M.S.: Computing similarity between cultural heritage items using multimodal features. In: LaTeCH at EACL, pp. 85–92 (2012)
26. O'Neill, S., Nicholson-Cole, S.: Fear won't do it: promoting positive engagement with climate change through imagery and icons. Sci. Commun. **30**(3), 355–379 (2009)
27. O'Neill, S., Smith, N.: Climate change and visual imagery. Wiley Interdisc. Rev.: Clim. Change **5**(1), 73–87 (2014)
28. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: describing images using 1 million captioned photographs. In: NIPS (2011)
29. Ortiz, L.G.M., Wolff, C., Lapata, M.: Learning to interpret and describe abstract scenes. In: NAACL HLT 2015, pp. 1505–1515 (2015)
30. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's mechanical turk. In: CSLDAMT at NAACL HLT (2010)
31. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: MM, pp. 251–260 (2010)
32. Socher, R., Fei-Fei, L.: Connecting modalities: semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
33. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. ACL **2**, 207–218 (2014)

34. Wang, C., Yang, H., Che, X., Meinel, C.: Concept-based multimodal learning for topic generation. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8935, pp. 385–395. Springer, Heidelberg (2015). doi:10.1007/978-3-319-14445-0_33
35. Weiland, L., Hulpus, I., Ponzetto, S.P., Dietz, L.: Understanding the message of images with knowledge base traversals. In: Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, 12–16 September 2016, pp. 199–208 (2016)
36. Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: EMNLP, pp. 444–454 (2011)
37. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. In: ACL, pp. 67–78 (2014)