Understanding the Message of Images with Knowledge Base Traversals

Lydia Weiland, Ioana Hulpuş, Simone Paolo Ponzetto, and Laura Dietz
Data and Web Science Group
University of Mannheim
68131 Mannheim, Germany
{lydia, ioana, simone, dietz}@informatik.uni-mannheim.de

ABSTRACT

The message of news articles is often supported by the pointed use of iconic images. These images together with their captions encourage emotional involvement of the reader. Current algorithms for understanding the semantics of news articles focus on its text, often ignoring the image. On the other side, works that target the semantics of images, mostly focus on recognizing and enumerating the objects that appear in the image. In this work, we explore the problem from another perspective: Can we devise algorithms to understand the *message* encoded by images and their captions? To answer this question, we study how well algorithms can describe an image-caption pair in terms of Wikipedia entities, thereby casting the problem as an entity-ranking task with an image-caption pair as query. Our proposed algorithm brings together aspects of entity linking, subgraph selection, entity clustering, relatedness measures, and learning-to-rank. In our experiments, we focus on media-iconic image-caption pairs which often reflect complex subjects such as sustainable energy and endangered species. Our test collection includes a gold standard of over 300 image-caption pairs about topics at different levels of abstraction. We show that with a MAP of 0.69, the best results are obtained when aggregating content-based and graph-based features in a Wikipedia-derived knowledge base.

Keywords

Image understanding;media-iconic images;entity ranking

1. INTRODUCTION

Newspaper articles and blog posts are accompanied by figures which consist of an image and a caption. While in some cases figures are used as mere decoration, more often figures support the message of the article in stimulating emotions and transmitting intentions. This is especially the case on matters of controversial topics, such as *global warming*, where emotions are conveyed through so-called **media-icons** [29, 9]: images with high suggestive power that illustrate the topic. A picture of a polar bear on melting shelf ice is a famous example cited by advocates stopping carbon emissions [27, 26]. As such, many images are able to broadcast abstract concepts and emotions [25], beyond the physical objects they illustrate.

ICTIR '16, September 12 - 16, 2016, Newark, DE, USA

ISBN 978-1-4503-4497-5/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2970398.2970414

Previous research has focused on the identification and labeling of tangible objects that are *visible* in the image (e.g., PascalVOC [11], MS COCO [20], Im2Text [28]). While this is an important prerequisite towards image understanding, in this paper, we take a step further and study how to identify the abstract (*invisible*) concepts or themes that an image conveys. CaptionBot [38] aims at generating captions for a given image. However these also only focus on the literal aspects. For the example in Figure 1b, the generated caption states: "I think it's a sign in front of a forest."

This paper is concerned particularly with the identification and ranking of these overarching concepts, that capture the message of the image, hereafter called **gist**. Thus, we cast the problem of gist detection as an entity ranking task with the following twist:

Task (Gist detection): Given an image-caption pair as *query*, rank entities from Wikipedia according to how relevant they describe the gist expressed in the image.

To address this task, we study the combination between (1) content-based features, extracted from the analysis of Wikipedia text, and (2) graph-based features obtained by analyzing Wikipedia's underlying article-category graph. By using the knowledge base as a graph, we represent the entities therein as nodes. For consistency, in the following we use the term **node** to refer to Wikipedia entities. Consequently, a node that corresponds to the gist of an image is referred to as **gist node**.

We approach the problem of detecting the gist that represents the message conveyed by an image and its caption with the following pipeline: First, detected **objects** in the image and detected entity **mentions** in the caption are projected onto nodes of the knowledge base (called *seed nodes*). Next, the node neighborhood of the seeds in the knowledge graph is inspected as a possible set for gist candidates. Finally, several graph and text measures are combined into a node ranking.

Of course, this is only a simple representation of a much bigger problem of interpreting images. Nevertheless, we see this study as a starting point for following research on gist-based image search and classification, detection of themes in images, and recommending images from the web when writing new articles for news, blogs, and Wikipedia. But even in the simple form of casting image understanding as an *entity ranking problem*, we see immediate utility: Being able to tag and annotate images with Wikipedia concepts provides a new way of traversing large image collections, such as Wikimedia commons (more than 30 million images). It also enables the detection of fake photographic evidence on social media, or whenever pictures are taken out of context and presented with a political spin.

The underlying idea of the knowledge base expansion is based on the following hypothesis: Especially for images and captions that express an abstract meaning, such as media-icons, we assume that

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.

Mammals of Southeast Asia, Trees, Orangutans, Botany, Plants

(a) Literal Pairing

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

Habitat Conservation, Biodiversity, Extinction, EDGE species, Deforestation

(b) Non-literal Pairing

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

Habitat Conservation, Biodiversity, Extinction, Politics, Protest

(c) Non-literal Pairing

Figure 1: Example image-caption pairs sharing either images or captions with their respective gist nodes[1].

the concepts that describe the gist of the pair best, are neither visible on the image, nor explicitly mentioned in the caption. Example of such entities transmitting the message referred to as gist are *global warming*, *endangered species*, *biodiversity*, or *sustainable energy*. Despite not being visible and consequently identifiable by image recognition, the gist nodes will likely be in close proximity to the objects in the image that are visible as well as to the entities mentioned in the captions. If this hypothesis is true, the node neighborhood of the seed nodes will not only contain the true gist nodes, but entity relatedness measures will help to pinpoint them.

We show through an extensive evaluation that the graph properties of nodes in the background knowledge base indicate that local and global measures are required to select correct gist nodes. This correlates with the overall working assumption that the gist of an image-caption pair is a product of common knowledge (e.g., a polar bear is an endangered animal), while being at the same time topic-specific (e.g., the polar bear is endangered through loss of its arctic habitat, partly caused by an increase in carbon emissions).

In this paper, we study the utility of our pipeline on a gold standard of 8200 true gist annotations (using relevance levels from 0 to 5), on more than 300 image-caption pairs about the discussed topic, global warming. To study the effectiveness of different measures, we use learning-to-rank as an experimentation environment.

Outline. We detail the problem statement and definitions in Section 2 and the related work in Section 3. In Section 4, we detail well-established methods that are used by our approach which is described in Section 5. We report findings from our experimental evaluation in Section 6 before concluding the paper.

2. PROBLEM STATEMENT AND GOLD STANDARD

In this work, we call an image instance and its associated textual caption an **image-caption pair**. Depending on the intention of the image-caption pair, we distinguish two types of pairs as follows. **Literal** pairs are those in which the caption describes or enumerates the objects depicted in the image. Figure 1a exemplifies a literal image-caption pair. **Non-literal** pairs, where the media-iconic pairs are a subclass of, are those which convey an abstract message, complemented and/or controlled by the image. As a running example, we use a topic on how deforestation affects endangered species. Figure 1 depicts three image-caption pairs, where two of them are on that topic and one (cf. Figure 1a) is a literal pair. When used to enrich text, images form a union with their captions. To understand the gist of an image-caption pair, both the image and the caption are needed. The image by itself might be given a different message by changing the caption, while the same caption will change semantics if accompanied by a different image.

For example, the pair in Figure 1a, presents an orangutan in what seems to be a national park. The gist of this figure is "Mammals of Southeast Asia". By exchanging the caption it becomes apparent that the gist is to stop deforestation to save an endangered animal (cf. Figure 1b). Considering the corresponding caption thus allows for disambiguation of the gist.

On the other hand, captions alone are often brief and when taken out from the context of the image, they fail to convey the entire message. For instance, by inspecting only the caption "Fight to save Indonesia's jungle corridors key for endangered orangutan.", it is not clear whether the focus is on the negative effects of deforestation as depicted in Figure 1b, or on people who fight against the causes of these negative effects, as depicted in Figure 1c. Only an image can disambiguate the gist. We consequently consider image-caption pairs as the targeted *queries* for which gist entities are ranked.

Furthermore, without a basic familiarity with the topic or domain of the intended message, it is very hard even for humans to grasp its gist from just looking at it. Media-icons nearly always play with prior knowledge of the user and might be differently understood in different cultural circles. (Our assessors are European.)

2.1 Research Questions

In this paper we are studying the following research questions according to our task of **gist detection** (cf. Section 1).

RQ1: How to link objects and mentions to seed nodes? We hypothesize that a simple string-match for linking image objects and entity mentions of the captions onto seed nodes without direct disambiguation, best represents the initial image-caption pair as query. The graph traversal and the re-ranking will serve as indirect disambiguation strategy.

RQ2: How close are gist nodes to seed nodes? We further hypothesize that good gist nodes are found in close proximity to the seed nodes. We study proximity in three layers: seed nodes, nodes in between seed nodes (**intermediate nodes**),

[1]http://www.reuters.com/article/us-environment-orangutan-idUSKCN0Q50HN20150801, http://www.thehindu.com/opinion/op-ed/how-oil-palm-could-kill-orangutans/article5777819.ece, last accessed: 07/18/2016.

and nodes two hops away from intermediate nodes (**border nodes**). In RQ2 we study the distribution of highly relevant gist nodes in these different layers and with respect to non-literal, literal, and both types of pairs.

RQ3: What is the benefit of the node neighborhood? We hypothesize that graph features, such as clusters and centrality measures, derived from the subgraph that includes border nodes, provide useful indicators for the true gist nodes. In RQ3 we separately study features involving border nodes from other features based on text as well as global graph properties.

2.2 Gold Standard

To the best of our knowledge there is yet no dataset covering the topic of non-literal image-caption detection, especially not including both, literal *and* non-literal pairs. To arrive at a challenging and realistic dataset, we collect images and captions for six topics related to global warming from the newspaper The Guardian, Our World magazine, and the website of the organization Union of Concerned Scientists. We confirm that all of them fall in the class of non-literal image-caption pairs.

To obtain equivalent literal image-caption pairs, we create an alternative descriptive caption for each image. The result is a balanced collection of 328 image-caption pairs (164 unique images).

Our aim is to evaluate the gist candidate selection and ranking approach without the complications from imperfect image processing (which is a research question for a different audience). Instead, we assign objects in images with bounding boxes and textual labels from a list of 43 object labels (e.g., Windmill, Solar panel, Orangutan), which are visible on the images.

To study the gist detection task, for each pair (both literal and non-literal) experts assess which of the nodes from the knowledge base represent the gist expressed in the image-caption pair. Gist nodes assessments are graded by relevance levels 0 to 5, from 0 (non-relevant), to 4 (relevant), reserving grade 5 for the six original non-literal topics such as *Biodiversity* and six corresponding literal topics such as *Orangutan*. Of the 8191 non-zero gist node annotations in total (≈25 per pair), 3100 obtain a grade of 4 or higher.

To evaluate RQ1, annotators separately assess links between entity mentions from the caption and objects of the image to nodes in the knowledge base.

Compared to other test collections in computer vision, this dataset of 328 "queries" is rather a small collection. However, this is the first test collection for literal and non-literal image-caption pairs with gold standard gist annotations and simulated object tags[2].

3. RELATED WORK

This work touches on different research communities evolving around the fields object detection from images, entity linking and retrieval, and using graph structure and content of knowledge bases.

Object detection from images. Triggered through benchmark collections, for image retrieval [37] and benchmarking tasks [33], a large body of works focuses on how to detect objects in images (e.g., PascalVOC [11], MS COCO [20], Im2Text [28]). These either train object detectors from images with bounding box annotations or use captions to guide the training or generate captions for images, based on an unsupervised model from the spatial relationship of such bounding boxes [10].

Since many images are accompanied by captions, approaches have been devised that use text in such captions to aid the detection of objects and actions depicted in the image. This idea is exploited using supervised ranking [16], using entity linking and WordNet distances [39], and using deep neural networks [35]. One application is image question answering [32]. Research to this end has thus far focused on literal image-caption pairs, where the caption enumerates the objects visible in the image. In contrast, the emphasis of this work is on non-literal image-caption pairs with media-iconic messages, which allude to an abstract gist concept that is not directly visible.

Even though datasets such as ImageNet, provide over 14 Mio. images, only 8% have bounding boxes, which are crucial for training object detectors. The lack of such training material is the only barrier for application in our domain. For this reason and to facilitate reproducibility of our research, we only simulate object detection in this work.

Entity linking. Detecting entity mentions in text and linking them to nodes in a knowledge base is a task well studied in the TAC KBP venue [24]. Most approaches include two stages: The first stage identifies candidate mentions of entities in the text with a dictionary of names. These candidates are disambiguated in the following stage using structural features from the knowledge graph, such as entity relatedness measures [5] and other graph walk features [36]. A prominent entity linking tool is the TagMe! system [12]. A simpler approach, taken by DBpedia spotlight [22], focuses on unambiguous entities and breaks ties by popularity. We evaluate both approaches in Section 6.

Entity retrieval. We cast our gist detection task as an entity retrieval task, with an image-caption pair as the query. Entity retrieval tasks have been studied widely in the IR community in INEX and TREC venues [8, 1]. The most common approach is to represent entities through textual and structural information in a combination of text-based retrieval models and graph measures [41].

Different definitions of entities have been explored. Recently, the definition of an entity as "anything that has an entry on Wikipedia" has become increasingly popular. Using entities from a knowledge base that are (latently) relevant for a query for ad hoc document retrieval has lead to performance improvements [6, 31]. Moreover, using text together with graphs from article links and category membership for entity ranking has been demonstrated to be effective on freetext entity queries such as "ferris and observation wheels" [7]. In contrast to this previous work, our paper focuses on a graph expansion and clustering approach.

In order to facilitate robust ranking behaviour, clustering is often combined into a back-off or smoothing framework. This has been successfully applied for document ranking by Raiber et al. [30], and our approach adopts it for the case of entity ranking.

Topic and document cluster labeling. Other research directions that are closely related to ours are concerned with labeling precomputed topic models [21, 18] and with labeling document clusters [4]. Topic model labeling is the task of finding the gist of a topic resulted from probabilistic topic modeling. Solutions to these related problems make implicit or explicit use of knowledge about words and concepts harvested from a document corpus. Such knowledge is not available for our problem rendering most of these approaches inapplicable.

Entity relatedness. The purpose of entity relatedness is to score the strength of the semantic association between pairs of concepts or entities. The research on this topic dates back several decades [40], and a multitude of approaches have been researched. Among them, we place particular emphasis on measures that use a knowledge base for computing relatedness. We distinguish two main direc-

[2]https://github.com/lweiland/GistDataset

tions: (i) works that use the textual content of the knowledge base [14, 17], particularly Wikipedia, and (ii) works that exploit the graph structure behind the knowledge base, particularly Wikipedia or Freebase hyperlinks [23], DBpedia [34, 19].

4. PRELIMINARIES

The main idea behind our approach is that a general-purpose knowledge base such as Wikipedia can aid the algorithmic understanding of the message conveyed by image and caption. We hypothesize that the way articles and categories are connected in Wikipedia can be exploited to identify nodes that capture the gist of the image-caption pair.

4.1 Knowledge Graphs

Given a knowledge base, we define a **knowledge graph** as the directed or undirected graph $\mathbf{KG}(V, E, T, \tau)$ such that the set of nodes V contains all nodes representing entities in the knowledge base, every edge $e_{ij} \in E$ corresponds to one relation in the knowledge base between two nodes v_i and v_j, the set T contains the relation types in the knowledge base, and the function $\tau : E \rightarrow T$ assigns each edge in E exactly one type in T. In this paper, we mostly consider the knowledge graph undirected, unless specified otherwise. In the following, we shortly explain preliminaries that are applied within our pipeline, but that are not part of our contribution.

Node properties. One of the most commonly used properties of nodes is their *degree* [15]. The degree of a node is the count of all edges that are adjacent to it. Another property of nodes is their tendency of being part of triangles called *local clustering coefficient* [15]. It is computed as the probability that any two random neighbors of a node are connected themselves.

Our intuition is that, these measures help to find a balance between specific and trivial nodes, and thus, the correct gist nodes. The degree and clustering coefficient of nodes are local measures that describe the nodes only in their closest vicinity.

Graph centrality measures. In the domain of network analysis, a wide range of graph centrality measures have been used with the purpose of locating the most important or influential nodes in the network. The PageRank [3] scores nodes based on their stationary probability that a random surfer will visit them. Betweeness centrality [13] defines a node as the more important the more often it lies on the shortest path between any two nodes in the graph.

Given a knowledge graph $\mathbf{KG}$, as we detail later, our approach makes use of a distance metric $\sigma^{(-1)} : V \times V \rightarrow \Re^{+}$ between two nodes. This metric captures the inverse of a similarity, relatedness, or semantic association measure between the concepts that are represented by the nodes. There are two of the main classes of measures: (i) those based on textual content associated with nodes and (ii) those based on a graph measure. In this work we are interested in using both content and graph structure.

4.2 Entity Relatedness

Content-based relatedness. Additionally we incorporate a content-based measure of relatedness. As each node in the DBpedia knowledge graph has a corresponding article in Wikipedia, we leverage a retrieval index on Wikipedia articles.

For a given entity mention, an object tag, or textual representation of the whole image-caption pair, we can use a retrieval model, which uses a query likelihood, to associate a measure of relevance with each node.

Graph-based relatedness. A great variety of semantic relatedness measures have been studied [40]. We follow Hulpus et al. [19] who introduce the exclusivity-based measure, which we use as a node metric $\sigma^{(-1)}$. The authors found that it works particularly well on knowledge graphs of categories and article membership (which we use also) for modeling concept relatedness. It was shown to outperform simpler measures that only consider the length of the shortest path, or the length of top-k shortest paths, as well as the measure proposed in [34].

The exclusivity-based measure assigns a *cost* for any edge $s \xrightarrow{r} t$ of type r between source node s and target node t. The cost function is the sum between the number of alternative edges of type r starting from s and the number of alternative edges of type r ending in t, as shown in Formula 1.

$$\text{cost}(s \xrightarrow{r} t) = |\{s \xrightarrow{r} *\}| + |\{* \xrightarrow{r} t\}| - 1, \tag{1}$$

where 1 is subtracted to count $s \xrightarrow{r} t$ only once.

The more neighbors connected through the type of a particular edge, the less informative that edge is, and consequently the less evidence it bears towards the relatedness of its adjacent concepts. By summing up the costs of all edges of a path p, one can compute the cost of that path, denoted cost(p). The higher the cost of a path, the lower its support for relatedness between the nodes at its ends. Thus, given two nodes, s and t, their relatedness is computed as the inverse of the weighted sum of the costs of the top-k shortest paths between them (ties are broken by cost function). Each path's contribution to the sum is weighted with a length based discounting factor α:

$$\sigma(s, t) = \sum_{i=1}^{k} \alpha^{\text{length}(sp_i)} \times \frac{1}{\text{cost}(sp_i)} \tag{2}$$

where sp_i denotes the i'th shortest path between s and t. $\alpha \in (0, 1]$ is the length decay parameter and k is a number of shortest paths to consider.

5. APPROACH: GIST DETECTION

As previously stated, the main idea behind our approach is that given a knowledge graph that covers the subject of the image-caption pair, the gist nodes lie in the proximity of the concepts mentioned in the caption or illustrated in the image. Furthermore, we define features of candidate gist nodes based on their graph relations. These come from different pipeline steps where we introduce them and are used in a final supervised reranking (Step 5). We present these steps as a pipeline that is illustrated in Figure 2. To further transmit the intuition of the approach, we make use of a running example (Figure 1) that shows each step according to the pipeline in Figure 2. The running example will be the media-iconic pair of Figure 1b. As explained in the gold standard (cf. Section 2), objects are given, e.g., "orangutan" and "sign". From the caption we extract mentions, e.g.,"wildlife corridor", "Indonesia", and "orangutan". Those objects and mentions serve as input data for Step 1, where we continue with our running example.

5.1 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 Wikipedia Categories. As for

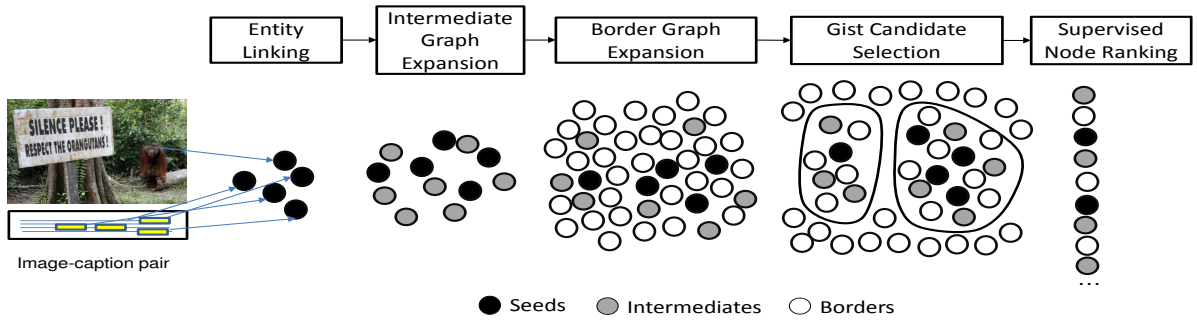

Figure 2: Gist extraction and ranking pipeline for a given image-caption pair. For simplicity, in this figure we omit the edges between nodes.

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 EDGE species.

5.2 Step 1: Image and Caption Node Linking

The first step is to project objects depicted in the image and entities mentioned in the caption onto nodes in the knowledge base.

To identify and create entity links for concepts mentioned in the caption to Wikipedia, we identify nouns and noun phrases in the caption as candidates for entity mentions. These candidate mentions are linked to the knowledge base nodes, which we call **caption nodes** in the following.

Where entity linking for text is a well established field, methods for linking objects in images to knowledge base nodes is still an open research question. As a consequence, out-of-the-box object detectors only provide limited object vocabulary trained on a dataset with bounding box information, i.e., bicycles, people, fruits, and houses. For the purpose of this work, we simulate an object recognizer and provide manual tags for a variety of theoretically recognizable object classes such as "grass", "orangutan", "vegetation", or "sign". Linking these object class labels to knowledge base nodes yields a set of **image nodes**. The mapped caption and image nodes are called **seed nodes** in the following.

We use a simple entity linking strategy that is both applicable to caption nodes and image nodes which are linked to nodes in the knowledge graph that represent articles as follows: First, we attempt to link noun phrases and image labels with exact matches to the article title. Whenever we would link to a title of a disambiguation page, we include all redirected articles that are within 2-hop distance to already-linked nodes. (Using unambiguous links for disambiguation is a standard entity linking strategy.)

In the experimental evaluation in Section 6, we demonstrate that this approach is as successful as using the TagMe [12] entity linking system for the domain at hand.

Example. As shown in the pipeline (Figure 2) objects in the image (orangutan, sign, trunk, tree, ground, and vegetation) and the entity mentions of the caption (fight, Indonesia, jungle, corridor, key, and orangutan) are linked to seed nodes, e.g., *Wildlife corridor* and *Orangutan* (Figure 3, depicted in grey).

5.3 Step 2: Intermediate Graph Expansion

Especially for media-iconic pairs, the gist refers to an abstract non-depictable concept, such as *Endangered Species*. Therefore, Step 1 may not be sufficient to identify such a gist by simple entity linking. However, one of our hypotheses is that gist nodes will be found in the knowledge base on paths between seed nodes.

In order to extract the nodes that enable the connection between the seed nodes, we extract all the paths with length shorter than 4, i.e., 2-hop, that connect all pairs of seeds. We call the nodes on these paths, except the seed nodes, **intermediate nodes**. The graph resulted from combining all the nodes on these paths (including the seeds) as well as the edges of the paths, is what we call the **intermediate graph**.

Example. Figure 3 shows the paths shorter than 4 edges that connect the three concepts: *Orangutan*, *Indonesia* and *Wildlife corridor*. Note that there is no path shorter than 4 directly connecting *Indonesia* to *Wildlife corridor*.

5.4 Step 3: Border graph & Metric

In order to further assess the graph properties of the seeds and intermediates, we expand the graph that contains all their 2-hop neighbors and the edges between them. We name this graph the **border graph**. The nodes that are added to the graph as a consequence of the expansion are called **border nodes**, as they lie between the seeds, intermediates, and the rest of the knowledge graph. Figure 4 shows a part of the border graph obtained from expanding the intermediate graph shown in Figure 3.

For all the nodes in the border graph, we compute a distance metric σ between all nodes in the border graph. Actually, it suffices to compute $\sigma(x, y)$ for $x \in S \cup I$ and $y \in S \cup I \cup B$.

between each seed and intermediate node as well as all border nodes and seed/intermediate nodes. This metric is used in two ways when extracting the top gist candidates:

- Step 4a) clustering the seed and intermediate nodes;
- Step 4b) selecting border nodes close to clusters.

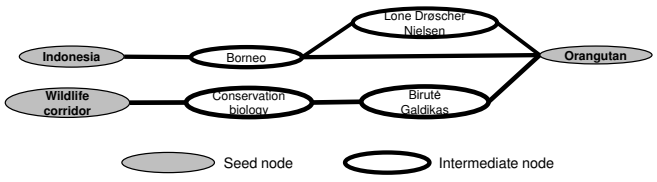

Figure 3: Example of intermediate graph for the image-caption pair in Figure 1 b.

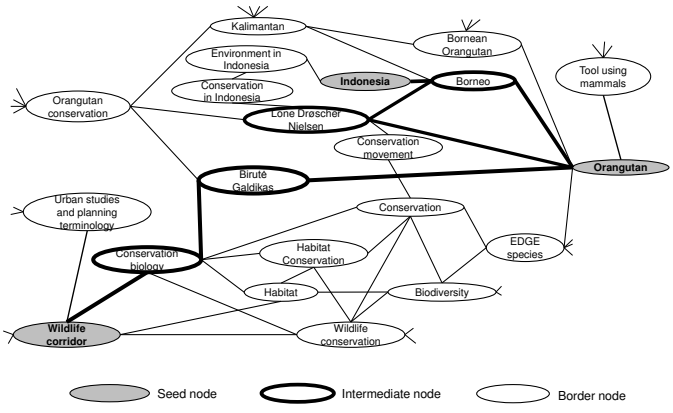

Figure 4: Border graph example for image-caption pair in Figure 1 b. For simplicity, the image does not make the distinction between article nodes and category nodes, and it also omits edge directions and edge costs.

As metric choices for the metric σ, we focus on the semantic relatedness measure defined in Equation 2, Section 4. This measure ingests information from the border nodes into the metric for two nodes from the joint seed/intermediate set (which is one way in which we exploit the border nodes). For this and many other graph-based measures it is sufficient to only consider the border graph. As the border graph is expected to be much smaller than the complete knowledge graph, this provides an upper bound on the computational complexity of this approach.

Example. According to Figure 4, with this step we include many border nodes such as, *Habitat*, *Biodiversity*, and *EDGE species*. Some of these constitute good candidates for gist nodes. But more importantly they affect the pairwise metric distance between the seed nodes. From the sparse information in the intermediate graph (Figure 3), it seems that *Indonesia* and *Wildlife Corridor* are equally far apart from *Orangutan*. However, metric uses information from the border graph to identify that *Indonesia* and *Orangutan* are much closer than *Wildlife Corridor* and *Orangutan*. An illustration of the distances is given in Figure 5.

5.5 Step 4a: Cluster Seed and Intermediates

After the previous step, we obtain a graph that contains all the concepts from the image and its caption, as well as multiple other concepts from the knowledge graph that lie in close proximity. As previously stated, our assumption is that the gist nodes are part of this graph, and that their graph properties will make them identifiable. However, a challenge is that often, an image-caption pair covers multiple sub-topics. Directly using the border graph between in the presence of multiple topics, will most often result in semantic drift and low-quality results.

To overcome this issue, we propose to identify weakly related sub-topics of an image-caption pair by clustering the joint set of seed and intermediate nodes based on their pairwise metric σ. We therefore anchor the set of gist candidate nodes to concepts that only arise in the context of the particular image-caption pair. To this end, we use the metric σ between all pairs of the joint seed and intermediate set, and apply Louvain clustering [2], a non-parametric modularity optimization network clustering algorithm. This clustering results in groups of seeds and intermediates that broadly correspond to different sub-topics of the image-caption pair.

Example. For the example in Figure 4, this step identifies several clusters, of which two are illustrated in Figure 5: Cluster C_1 is about wildlife conservation, containing the seed node *Wildlife Corridor*, and cluster C_2 represents topics about Indonesia, including both *Orangutan* and *Indonesia*.

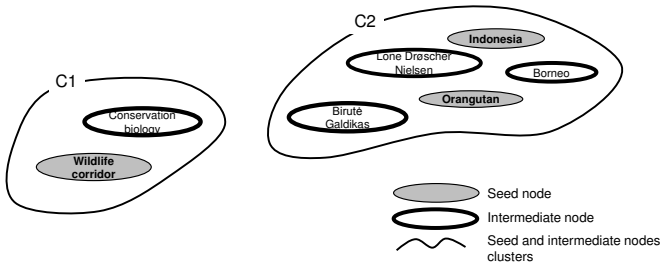

Figure 5: Example of seeds and intermediates, clustered based on their pairwise metric.

5.6 Step 4b: Selecting Top Border Nodes

In this step, we identify a subset of suitable border nodes that would make good gist candidates. We hypothesize that these border nodes are close to any of the clusters according to the metric σ. We therefore compute for every border node x, its average metric distance $\bar{\sigma}$ for every cluster C as shown in Formula 3.

$$\bar{\sigma}(x, G^C) = \frac{1}{|S^C \cup I^C|} \sum_{y \in S^C \cup I^C} \sigma(x, y) \tag{3}$$

In order to keep the number of gist candidates reasonable, for each cluster we select the top-k closest borders as well as all seeds and all intermediates. These nodes constitute the candidate node set which is ranked in the following step.

Example. The association of top border nodes with the two example clusters is illustrated in Figure 6. For instance, the wildlife cluster C_1 includes *Habitat* and *Biodiversity*, where both *Orangutan Conservation* and the the geographic region *Kalimantan* are associated with cluster C_2. The border node *Conservation* is associated with both clusters. These border nodes are included in the candidate set, in contrast to borders with a high distance such as *Urban studies and planning terminology* which are left out.

5.7 Step 5: Supervised Node Ranking

For each of the candidate nodes, a feature vector is created and ranked for relevance with supervised learning-to-rank. The feature vector consists of features listed in Table 1 collected from the various steps of the pipeline:

Seed and intermediate features (Step 1–2). Seed and intermediate nodes are distinguished by two binary features. For all the nodes in the intermediate graph, we compute and retain their betweenness centrality and their PageRank score as features.

Border features (Step 3–4). After the expansion into the border graph, we introduce a feature indicating the border nodes.

We leverage information from the clustering step by associating each node with its proximity $\bar{\sigma}(n, c)$ (inverse metric) to the nearest cluster c. This feature is also used as an unsupervised baseline in the experimental evaluation.

As border nodes can be associated with more than one cluster (cf. *Conservation* in Figure 6) we additionally add features capturing the sum (and average) proximity to all clusters $\sum_c \bar{\sigma}(n, c)$.

We assume that the more seed nodes are member of a cluster, the more relevant this cluster is for expressing the gist of the image-caption pair. This assumption is expressed in two features: a binary

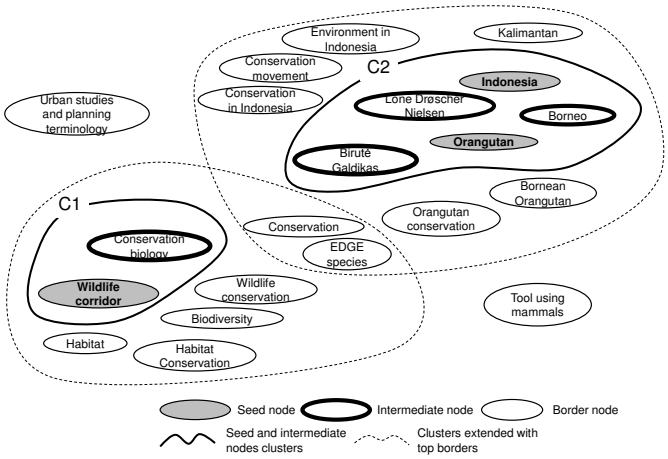

Figure 6: Example of clusters of seeds and intermediates, extended with their most highly related border nodes (top borders). The border nodes that have only weak semantic associations to the clusters are filtered out (e.g., *Tool using mammals* and *Urban studies and planning terminology*).

Table 1: Features for supervised re-ranking. Border features marked with X; baselines marked with †.

Description	Step	feat. set
is seed node?	1	
is intermediate node?	2	
is border node?	3	X
Page Rank on intermediate graph	2	
Betweenness Centrality on intermediate graph	2	
max node-cluster relatedness	4	X †
avg node-cluster relatedness	4	X
sum node-cluster relatedness	4	X
is member of cluster with most seed nodes	4	X
is member of cluster with most seed/intermediate nodes	4	X
fraction of seeds in cluster	4	X
fraction of seeds and intermediates in cluster	4	X
query likelihood on KB text	cont.	†
indegree of node	glob.	
clustering coefficient	glob.	

feature indicating nodes which are members of the cluster that contains the highest number of seed nodes; and a fraction of all seed nodes that this node is sharing a cluster with (summing fractions for nodes with multiple cluster memberships).

Exploiting potential benefit of the joint set of seed and intermediate nodes, we further indicate membership of the cluster with highest number of nodes that are seed or intermediate nodes, as well as the fraction of all seed or intermediate nodes in shared clusters.

Content features. We include a content-based similarity measure for image-caption pairs. For this we concatenate all (distinct) entity mentions from the caption and all object annotations from the image as a keyword query. We use the query to retrieve textual content associated with article and category nodes using the query likelihood model with Dirichlet smoothing. We use this retrieval model to only rank nodes in the candidate set relative to each other.

We use this ranking as a baseline for the experimental evaluation and include the reciprocal rank as a node feature.

Global features. Finally we also include global node features that are independent of the image-caption pair. These include the indegree of a node in the graph as well as the clustering coefficient.

These generated feature vectors serves as an input for a listwise learning-to-rank model, trained with respect to the target metric Mean-average precision (MAP). The model is evaluated with five-fold cross validation based on MAP, NDCG, and Precision@k.

Table 2: Number of image and caption nodes after entity linking.

	Non-Literal		Literal		Overall
	Image	Caption	Image	Caption	
Unique nodes	43	684	43	306	806
Total occurrences	640	1412	640	843	3535

Table 3: Correctness of different entity linking methods for image and caption nodes. ("No answer" counted as mistake.)

Linking Method	P	R
String2Article	0.9	0.97
String2Category	1.0	0.27
TagMe	0.7	0.83
Wikipedia index	0.81	0.98

6. EXPERIMENTAL EVALUATION

The utility of our pipeline is evaluated on gold standard annotations on 328 image-caption pairs from the wide topical domain of global warming. The evaluation is presented according to the research questions (**RQ**) of Section 2.1. In the gold standard gist nodes are assessed with a grade, ranging from 0-5, where 0 is non-relevant and 5 represents maximally relevant. Unless noted otherwise, we binarize the assessments and call a node relevant if and only if it is rated with a grade of 4 or 5. Among all relevant nodes in this study: 54,6 % of all gist nodes are entities and 45,4% are categories.

Experimental setup. We use a combined knowledge base aligning Wikipedia (WEX dump from 2012), Freebase (from 2012), and DBpedia (from 2014). This knowledge base is used to entity linking, deriving edges for the graph, and the content-based retrieval methods. The knowledge base is indexed with Galago.[3] With respect to the relatedness measure we use for metric $\sigma^{(-1)}$, we follow Hulpus et al. [19], Formula 2. We use their settings for hyperparameters $\alpha = 0.25$ and take the $k = 3$ shortest paths.

6.1 RQ 1: Seed node linking

We first evaluate the entity linking performance of the simple string-match method used in Step 1 to produce a set of image nodes and caption nodes. These together form the set of seed nodes. We use a separate gold standard to evaluate the correctness of the established links (i.e., not the gold standard in Section 2, which is used in RQ2 and RQ3). We verify all links for correctness, especially those with multiple meanings. In our application domain of global warming, most entities and objects denote general concepts as common nouns – people and organizations are an exception.

Image and caption nodes. We find that a total of 806 different nodes are reached from image or caption across all pairs. These result in 3535 links from image objects or caption mentions across the dataset. In total, only 5 noun phrases in captions are not linkable to the knowledge base (e.g., underwater view).

Table 2 presents the total number of different image nodes and caption nodes, where we distinguish between literal and non-literal image-caption pairs. We see that all images make use of an object vocabulary that can be disambiguated to 43 different nodes (different senses are counted multiple times) with a total frequency of

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

Table 4: Quality of gist candidate selection method. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Avg cands.	P	R	F1	ΔF1%
Seeds	8.6	0.18	0.16	0.17	0.0
Intermediates	11.4	0.19	0.22	0.21	+19.0%*
Top Borders	31	0.09	0.30	0.14	-21.4%*

Table 5: Statistics about proportion of highly ranked gists (grade > 3 and grade > 4), with respect to the found gist nodes in Table 4).

	Grade 4 or 5			Grade 5	
	All	Non-Lit.	Literal	Non-Lit.	Literal
Seeds	53.79%	53.46%	53.96%	57.89%	70.75%
Intermediates	21.05%	21.70%	20.73%	07.89%	17.92%
Borders	25.16%	24.84%	25.32%	34.21%	11.32%

640 linkable objects across all images. As each image gives rise to a literal and non-literal pair, there is no difference between these columns. We observe a much wider range of nodes when linking entity mentions in the caption. In particular we notice a smaller vocabulary for literal image-caption pairs (306 unique nodes) compared to non-literal pairs (684 unique nodes), where each node is mentioned about three times on average. However, we find that the caption nodes set from literal versus non-literal pairs have nearly no overlap.

Entity linking. The set of seed nodes is formed by the union of image and caption nodes. In Step 1 we link objects/mentions to article nodes. However, the same procedure could have been applied to category names as well. We first compare these two methods in comparison to entity links produced by the TagMe system. Furthermore, we use the retrieval index of texts associated with nodes and output the top ranked node (Wikipedia index).

Table 3 presents precision and recall achieved by these four methods on the set of all 806 unique image/caption nodes. We find that all methods perform reasonably well, where the category-based linking strategy cannot associate a vast majority of 581 objects / mentions. In particular, we find that our heuristics in Step 1 outperforms TagMe and is better in precision than retrieving from the Wikipedia index.

Discussion Step 1. The TagMe system is a strong state-of-the-art entity linking system. How can it be outperformed by such simple heuristics? TagMe is particularly strong whenever interpretation and association is required, for instance to disambiguate ambiguous people names, organizations, and abbreviations. In contrast, the concepts we are linking in this domain are mostly common nouns, for which Wikipedia editors have done the work for us already. In the remaining cases that need disambiguation, our heuristic is likely to encounter a disambiguation page. At this point, we are using a well known disambiguation heuristic by using graph connections to unambiguous contextual mentions/objects.

We conclude that our simple entity linking method on articles works much better than on categories and as well as TagMe, thus its use is justified in our pipeline.

6.2 RQ2: Distribution of Relevant Gist Nodes

The research question is, whether good gist nodes are found in close proximity neighborhood to the directly depicted and mentioned seed nodes. We distinguish proximity in the three expansion layers of seed nodes, intermediate nodes, and top border nodes and evaluate the benefits of each graph expansion step.

Benefits of expansions. Taking nodes with gist grades of 4 and 5 as relevant, we study how precision and recall changes with the different expansion/filtering steps along the pipeline (Step 1, 2, 4b). The results are presented in Table 4, where we give precision, recall, and F1 together with the number of average candidates per image-caption pair. In order to judge the significance of improvement for F1 we evaluate with the relative increase in precision, on a per-image-caption-pair basis, and report the average (denoted Δ). Significance is verified with a paired-t-test with level 0.05.

We find that especially the expansion into the intermediate graph increases both recall and precision. While the increase in F1 is relatively small, it is statistically significant across the image-caption pairs, where it yields an average increase of 19%.

The expansion into the border graph of Step 3 and its contraction to the closest border nodes in Step 4b yields the new set of top border nodes. While it increases recall quite drastically, the loss in precision leads to a significant loss in F1 (over the seed set).

Visible versus invisible gists. We subdivide the set of true gists (grade 4 or 5) into visible gists and invisible gists. Visible gists can be depicted, such as *Orangutan* or *Plants*. In contrast, invisible gists are non-tangible concepts such as *Biodiversity*. Confirming our intuition, we find a much higher fraction of non-visible gists in the candidate set of non-literal image-caption pairs (34%) than in literal image-caption pairs (10%). This trend is even more pronounced when we look at the seed sets associated with image-caption pairs, where 99% of relevant seeds for literal image-caption pairs are visible.

Distribution of high quality gists. We change perspective and ask in which expansion set the majority of high-quality gists are to be found. Initially, we hypothesized that especially for non-literal caption pairs, fewer good gists will be found in the seed set, which motivated the graph expansion approach. Accordingly we separately report findings on literal and non-literal subsets.

We study two relevance thresholds in Table 5, for relevant gists (grade 4 or 5) as well as a stricter threshold including grade 5 only. Gists graded with 5 are limited to the one major gist of the image-caption pair, which in almost all non-literal cases refers to an abstract topic such as *Biodiversity*.

Focusing on the distribution relevant gists (grade 4 or 5) in Table 5, left, we notice that more than half of the gists are already contained in the seed set and about 20% are found in the intermediate set. The much larger border set still contains a significant portion of relevant gists. Focusing on the differences between literal and non-literal pairs, we verify that there are no significant differences between the distributions. Where gists with grade 4 or 5 are highly relevant, they still include the most important visible concepts for non-literal image-caption pairs.

However, if we focus only on the distribution of gists with grade 5 in Table 5, right, we notice that 71% of high-quality gists in literal pairs are found in the seed set which is in stark contrast to only 58% for non-literal pairs. Also, for non-literal image-caption pairs we found the most useful gists in the set of border nodes with high cluster proximity.

Discussion of Steps 2–4. We confirm that many relevant (grade 4) and high-quality (grade 5) gists are found in the seed set and the node neighborhood. The large fraction of nodes available in the border set (compared to the intermediate set) suggests that limiting the intermediate graph expansion in Step 2, to be between seed nodes is too restrictive. We see our initial assumptions confirmed in that literal image-caption pairs, which is where most of the related work is focusing on, contain more visible gists, and those are directly visible/mentioned. For non-literal pairs, the high-quality

Table 6: Entity ranking results (grade 4 or 5) of supervised learning-to-rank. 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
All Features	0.69	0.0	0.73	0.7	0.56	0.0	0.6	0.56	0.82	0.0	0.87	0.84
All But Borders	0.66	-4.35%*	0.7	0.67	0.54	-3.5%	0.57	0.55	0.78	-4.88%*	0.83	0.8
Only Borders	0.63	-8.70%*	0.64	0.64	0.52	-7.14%*	0.54	0.52	0.73	-10.98%*	0.74	0.76
Wikipedia ranking	0.43	-37.68%*	0.48	0.37								
max node-cluster relatedness	0.29	-57.97%*	0.57	0.40								

gists are not only invisible, but also more often only implicitly given. Nevertheless, the graph-based relatedness measures are able to identify a reasonable candidate set.

6.3 RQ3: Value of Border Features

The last research question assesses the quality of our supervised node ranking. We further inspect the question whether features generated by global and local graph centrality measures, especially those derived from border graph expansions enhances the overall gist node ranking.

For this study we use a supervised learning-to-rank approach to evaluate the benefit of the feature sets. Here we use the RankLib[4] package using all features, in comparison to the subset of border features only (cf. Step 5), and all but the border features. We train RankLib using the ground truths of gist nodes (grade 4 or 5 as relevant) optimizing for the metric mean-average precision. We use coordinate ascent with a linear kernel and perform 5-fold cross validation using each image-caption pair as one "query". This way we predict 328 node rankings for all image-caption pairs, while keeping training and testing data separate.

To assess the different aspects of content and semantic relatedness, we compare the results of Step 5 with the three feature sets and two baselines (marked with † in Table 1): One retrieves Wikipedia text using the query likelihood model on all entity mentions and object annotations concatenated, the other is based on an unsupervised ranking according to the maximal node-cluster relatedness measure, as described in Step 4. We use a learning-to-rank a model to study the value of border features (marked with b in Table 1). The results in Table 6 are tested for significance (p-value ≤ 0.05).

We study the research question with respect to both, non-literal, and literal pairs in Table 6, where we report ranking quality in terms of mean-average precision (MAP), NDCG@10, and precision (P@10) of the top ten ranks.

Overall entity ranking results. On the whole, our approach achieves relative high ranking performance of 0.69 MAP across all image-caption pairs. As expected, ranking non-literal image-caption pairs is much harder (MAP:0.56) than for literal pairs (MAP:0.82). Yet, even in the non-literal case, more than half of the nodes in the top 10 are relevant.

In contrast, both baselines are much worse. The baseline which ranks nodes by the query likelihood model on all entity mentions and objects achieves a MAP of 0.43 (being 38% worse). The baseline which just includes the max node-cluster relatedness obtains an even worse performance of 0.29 MAP, even though both achieve the same P@10 performance).

Value of border features. One research question was to study whether the border features, which include the metric (relatedness), clustering, and membership of largest cluster provide value. We therefore compare the performance changes of the learning-to-rank approach, when we vary the feature set from full, to border features (and also an ablation study of all but border features). In both cases, the ranking quality drops significantly by 4–11%, where the literal pairs seem to benefit slightly more from border features.

Discussion of Step 4–5. The fact that our full re-ranking pipeline improves so drastically over both a retrieval baseline and a cluster-relatedness baseline demonstrates the benefit of our approach. Here, both baselines are incorporated with further information such as centrality, degree, and expansion zone. Regarding research question RQ3, we can verify that measures that are derived from border nodes (Step 4) are contributing a significant performance benefit.

7. CONCLUSION

Given an image-caption pair, our aim is to automatically understand the message it conveys. To this end, we focus on non-literal image-caption pairs with media-iconic elements as found in news articles.

Using a test collection of 328 image-caption pairs as "queries", we cast the problem of message understanding as an entity ranking task. First, objects in the image and entities in the caption are linked to nodes in a knowledge base. Then graph expansions and clustering are used to select a candidate set of knowledge base entities that represent the message. Candidates are ranked based on graph-based and content-based features, which outperform baselines using either as well as feature subsets.

We find that easy heuristics are sufficient to link objects to the knowledge base (RQ1). Furthermore, while for literal pairs, good gist nodes are often directly linked and often visible on the image, we find that for non-literal pairs about 40% of very important entities are only found in the expanded graph (RQ2). Simple graph expansion will introduce too much noise, but by using a supervised learning-to-rank approach which is capable of exploiting even far away nodes (RQ3) achieving a MAP 0.56 for image-caption pairs with a message.

Acknowledgements

This work is funded by the Research Seed Capital (RiSC) programme of the Ministry of Science, Research and the Arts Baden-Württemberg, 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.

8. REFERENCES

[1] K. Balog, P. Serdyukov, and A. P. d. Vries. Overview of the TREC 2010 entity track. Technical report, DTIC Document, 2010.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008:1–12, Oct. 2008.

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

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, pages 107–117, 1998.

[4] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *SIGIR*, pages 139–146, 2009.

[5] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Learning relatedness measures for entity linking. In *CIKM*, pages 139–148, 2013.

[6] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *SIGIR*, pages 365–374, 2014.

[7] G. Demartini, C. S. Firan, T. Iofciu, and W. Nejdl. *WISE 2008*, chapter Semantically Enhanced Entity Ranking, pages 176–188. Springer, Berlin, Heidelberg, 2008.

[8] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the inex 2009 entity ranking track. In *Focused Retrieval and Evaluation*, pages 254–264. Springer, 2009.

[9] B. Drechsel. The berlin wall from a visual perspective: comments on the construction of a political media icon. *Visual Communication*, 9(1):3–24, 2010.

[10] D. Elliott and A. de Vries. Describing images using inferred visual dependency representations. In *ACL*, pages 42–52, 2015.

[11] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[12] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*, pages 1625–1628. ACM, 2010.

[13] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, page 215, 1978.

[14] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611, 2007.

[15] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers. Topics in semantic representation. *Psychological Review*, 114:2007, 2007.

[16] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[17] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *CIKM*, pages 545–554, 2012.

[18] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using DBpedia. In *WSDM*, pages 465–474, 2013.

[19] I. Hulpuş, N. Prangnawarat, and C. Hayes. Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *ISWC*, pages 442–457, 2015.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[21] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *SIGKDD*, pages 490–499, 2007.

[22] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *I-Semantics*, pages 1–8, 2011.

[23] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, pages 509–518, 2008.

[24] NIST. *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*, 2013.

[25] S. O'Neill and S. Nicholson-Cole. Fear won't do it: promoting positive engagement with climate change through imagery and icons. *Science Communication*, 30(3):355–379, 2009.

[26] S. O'Neill and N. Smith. Climate change and visual imagery. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):73–87, 2014.

[27] S. J. O'Neill and M. Hulme. An iconic approach for respresenting climate change. *Global Environmental Change*, 19(4), 2009.

[28] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[29] D. D. Perlmutter and G. L. Wagner. The anatomy of a photojournalistic icon: Marginalization of dissent in the selection and framing of 'a death in Genoa'. *Visual Communication*, 3(1), Feb. 2004.

[30] F. Raiber and O. Kurland. Ranking document clusters using markov random fields. In *SIGIR*, pages 333–342, 2013.

[31] H. Raviv, O. Kurland, and D. Carmel. Document retrieval using entity-based language models. In *SIGIR*, 2016.

[32] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2935–2943, 2015.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[34] M. Schuhmacher and S. P. Ponzetto. Knowledge-based graph document modeling. In *WSDM*, pages 543–552, 2014.

[35] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *ACL*, 2:207–218, 2014.

[36] P. P. Talukdar, J. Reisinger, M. Paşca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In *EMNLP*, pages 582–590, 2008.

[37] B. Thomee and A. Popescu. Overview of the ImageCLEF 2012 Flickr photo annotation and retrieval task. In *CLEF*, 2012.

[38] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz. Rich image captioning in the wild. *arXiv preprint arXiv:1603.09016*, 2016.

[39] R. Weegar, L. Hammarlund, A. Tegen, M. Oskarsson, K. Åström, and P. Nugues. Visual entity linking: A preliminary study. In *AAAI-14 Workshop on Computing for Augmented Human Intelligence*, 2014.

[40] Z. Zhang, A. L. Gentile, and F. Ciravegna. Recent advances in methods of lexical semantic relatedness - a survey. *Natural Language Engineering*, 19:411–479, 10 2013.

[41] N. Zhiltsov, A. Kotov, and F. Nikolaev. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *SIGIR*, pages 253–262, 2015.