Linking Entities Across Images and Text

Weegar, Rebecka; Åström, Karl; Nugues, Pierre

Published in:
Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL 2015)

2015

Linking Entities Across Images and Text

Rebecka Weegar
DSV
Stockholm University

Kalle Åström
Dept. of Mathematics
Lund University

Pierre Nugues
Dept. of Computer Science
Lund University

rebeckaw@dsv.su.se, kalle@maths.lth.se, Pierre.Nugues@cs.lth.se

Abstract

This paper describes a set of methods to link entities across images and text. As a corpus, we used a data set of images, where each image is commented by a short caption and where the regions in the images are manually segmented and labeled with a category. We extracted the entity mentions from the captions and we computed a semantic similarity between the mentions and the region labels. We also measured the statistical associations between these mentions and the labels and we combined them with the semantic similarity to produce mappings in the form of pairs consisting of a region label and a caption entity. In a second step, we used the syntactic relationships between the mentions and the spatial relationships between the regions to rerank the lists of candidate mappings. To evaluate our methods, we annotated a test set of 200 images, where we manually linked the image regions to their corresponding mentions in the captions. Eventually, we could match objects in pictures to their correct mentions for nearly 89 percent of the segments, when such a matching exists.

1 Introduction

Linking an object in an image to a mention of that object in an accompanying text is a challenging task, which we can imagine useful in a number of settings. It could, for instance, improve image retrieval by complementing the geometric relationships extracted from the images with textual descriptions from the text. A successful mapping would also make it possible to translate knowledge and information across image and text.

In this paper, we describe methods to link mentions of entities in captions to labeled image segments and we investigate how the syntactic structure of a caption can be used to better understand the contents of an image. We do not address the closely related task of object recognition in the images. This latter task can be seen as a complement to entity linking across text and images. See Russakovsky et al. (2015) for a description of progress and results to date in object detection and classification in images.

2 An Example

Figure 1 shows an example of an image from the Segmented and Annotated IAPR TC-12 data set (Escalantea et al., 2010). It has four regions labeled *cloud*, *grass*, *hill*, and *river*, and the caption:

> a flat landscape with a dry meadow in the foreground, a lagoon behind it and many clouds in the sky

containing mentions of five entities that we identify with the words *meadow*, *landscape*, *lagoon*, *cloud*, and *sky*. A correct association of the mentions in the caption to the image regions would

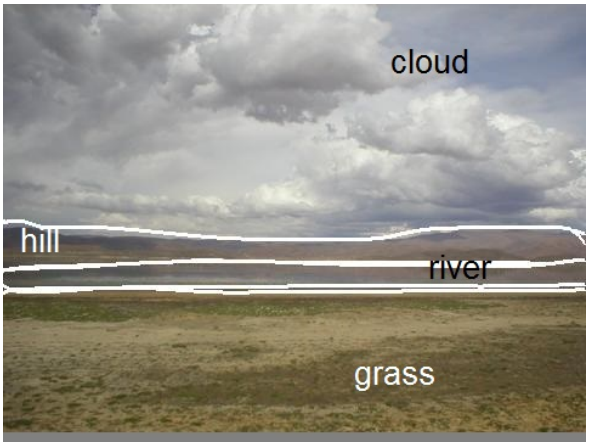

Figure 1: Image from the Segmented and Annotated IAPR TC-12 data set with the caption: *a flat landscape with a dry meadow in the foreground, a lagoon behind it and many clouds in the sky*

Proceedings of the 19th Conference on Computational Language Learning, pages 185–193,
Beijing, China, July 30-31, 2015.

map *clouds* to the region labeled *cloud*, *meadow* to *grass*, and *lagoon* to *river*.

This image, together with its caption, illustrates a couple of issues: The objects or regions labelled or visible in an image are not always mentioned in the caption, and for most of the images in the data set, more entities are mentioned in the captions than there are regions in the images. In addition, for a same entity, the words used to mention it are usually different from the words used as labels (the categories), as in the case of *grass* and *meadow*.

3 Previous Work

Related work includes the automatic generation of image captions that describes relevant objects in an image and their relationships. Kulkarni et al. (2011) assign each detected image object a visual attribute and a spatial relationship to the other objects in the image. The spatial relationships are translated into selected prepositions in the resulting captions. Elliott and Keller (2013) used manually segmented and labeled images and introduced *visual dependency representations* (VDRs) that describe spatial relationships between the image objects. The captions are generated using templates. Both Kulkarni et al. (2011) and Elliott and Keller (2013) used the BLEU-score and human evaluators to assess grammatically the generated captions and on how well they describe the image.

Although much work has been done to link complete images to a whole text, there are only a few papers on the association of elements inside a text and an image. Naim et al. (2014) analyzed parallel sets of videos and written texts, where the videos show laboratory experiments. Written instructions are used to describe how to conduct these experiments. The paper describes models for matching objects detected in the video with mentions of those objects in the instructions. The authors mainly focus on objects that get touched by a hand in the video. For manually annotated videos, Naim et al. (2014) could match objects to nouns nearly 50% of the time.

Karpathy et al. (2014) proposed a system for retrieving related images and sentences. They used neural networks and they show that the results are improved if image objects and sentence fragments are included in the model. Sentence fragments are extracted from dependency graphs, where each edge in the graphs corresponds to a fragment.

4 Entity Pairs

4.1 Data Set

We used the *Segmented and Annotated IAPR TC-12 Benchmark* data set (Escalantea et al., 2010) that consists of about 20,000 photographs with a wide variety of themes. Each image has a short caption that describes its content, most often consisting of one to three sentences separated by semicolons. The images are manually segmented into regions with, on average, about 5 segments in each image.

Each region is labelled with one out of 275 predefined image labels. The labels are arranged in a hierarchy, where all the nodes are available as labels and where `object` is the top node. The labels `humans`, `animals`, `man-made`, `landscape/nature`, `food`, and `other` form the next level.

4.2 Entities and Mentions

An image caption describes a set of entities, the caption entities CE, where each entity CE_i is referred to by a set of mentions M. To detect them, we applied the Stanford CoreNLP pipeline (Toutanova et al., 2003) that consists of a part-of-speech tagger, lemmatizer, named entity recognizer (Finkel et al., 2005), dependency parser, and coreference solver. We considered each noun in a caption as an entity candidate. If an entity CE_i had only one mention M_j, we identified it by the head noun of its mention. We represented the entities mentioned more than once by the head noun of their most representative mention. We applied the entity extraction to all the captions in the data set, and we found 3,742 different nouns or noun compounds to represent the entities.

In addition to the caption entities, each image has a set of labeled segments (or regions) corresponding to the image entities, IE. The Cartesian product of these two sets results in pairs P generating all the possible mappings of caption entities to image labels. We considered a pair (IE_i, CE_j) a correct mapping, if the image label IE_i and the caption entity CE_j referred to the same entity. We represented a pair by the region label and the identifier of the caption entity, *i.e.* the head noun of the entity mention. In Fig. 1, the correct pairs are $(\text{grass}, \text{meadow})$, $(\text{river}, \text{lagoon})$, and $(\text{cloud}, \text{clouds})$.

4.3 Building a Test Set

As the Segmented and Annotated IAPR TC-12 data set does not provide information on links between the image regions and the mentions, we annotated a set of 200 randomly selected images from the data set to evaluate the automatic linking accuracy. We assigned the image regions to entities in the captions and we excluded these images from the training set. The annotation does not always produce a 1:1 mapping of caption entities to regions. In many cases, objects are grouped or divided into parts differently in the captions and in the segmentation. We created a set of guidelines to handle these mappings in a consistent way. Table 1 shows the sizes of the different image sets and the fraction of image regions that have a corresponding entity mention in the caption.

Set	Files	Regions	Mappings	%
Data set	19,176	–	–	–
Train. set	18,976	–	–	–
Test set	200	928	730	78.7

Table 1: The sizes of the different image sets.

5 Ranking Entity Pairs

To identify the links between the regions of an image and the entity identifiers in its caption, we first generated all the possible pairs. We then ranked these pairs using a semantic distance derived from WordNet (Miller, 1995), statistical association metrics, and finally, a combination of both techniques.

5.1 Semantic Distance

The image labels are generic English words that are semantically similar to those used in the captions. In Fig. 1, *cloud* and *clouds* are used both as label and in the caption, but the region labeled *grass* is described as a *meadow* and the region labeled *river*, as a *lagoon*. We used the WordNet Similarity for Java library, (WS4J), (Shima, 2014) to compute the semantic similarity of the region labels and the entity identifiers. WS4J comes with a number of metrics that approximate similarity as distances between WordNet synsets: PATH, WUP (Wu and Palmer, 1994), RES, (Resnik, 1995), JCN (Jiang and Conrath, 1997), HSO (Hirst and St-Onge, 1998), LIN (Lin, 1998), LCH (Leacock and Chodorow, 1998), and LESK (Banerjee and Banerjee, 2002).

We manually lemmatized and simplified the image labels and the entity mentions so that they are compatible with WordNet entries. It resulted in a smaller set of labels: 250 instead of the 275 original labels. We also simplified the named entities from the captions. When a person or location was not present in WordNet, we used its named entity type as identifier. In some cases, it was not possible to find an entity identifier in WordNet, mostly due to misspellings in the caption, like *buldings*, or *buidling*, or because of POS-tagging errors. We chose to identify these entities with the word *entity*. The normalization reduced the 3,742 entity identifiers to 2,216 unique ones.

Finally, we computed a 250×2216 matrix containing the similarity scores for each (image label, entity identifier) pair for each of the WS4J semantic similarity metrics.

5.2 Statistical Associations

We used three functions to reflect the statistical association between an image label and an entity identifier:

- Co-occurrence counts, i.e. the frequencies of the region labels and entity identifiers that occur together in the pictures of the training set;
- Pointwise mutual information (PMI) (Fano, 1961) that compares the joint probability of the occurrence of a (image label, entity identifier) pair to the independent probability of the region label and the caption entity occurring by themselves; and finally
- The simplified Student's t-score as described in Church and Mercer (1993).

As with the semantic similarity scores, we used matrices to hold the scores for all the (image label, entity identifier) pairs for the three association metrics.

5.3 The Mapping Algorithm

To associate the region labels of an image to the entities in its caption, we mapped the label L_i to the caption entity E_j that had the highest score with respect to L_i. We did this for the three association scores and the eight semantic metrics. Note that a region label is not systematically paired with the same caption entity, since each caption contains different sets of entities.

Background and *foreground* are two of the most frequent words in the captions and they were frequently assigned to image regions. Since they rarely represent entities, but merely tell *where* the entities are located, we included them in a list of stop words, as well as *middle*, *left*, *right*, and *front* that we removed from the identifiers.

We applied the linking algorithm to the annotated set. We formed the Cartesian product of the image labels and the entity identifiers and, for each image region, we ranked the caption entities using the individual scoring functions. This results in an ordered list of entity candidates for each region. Table 2 shows the average ranks of the correct candidate for each of the scoring functions and the total number of correct candidates at different ranks.

6 Reranking

The algorithm in Sect. 5.3 determines the relationship holding between a pair of entities, where one element in the pair comes from the image and the other from the caption. The entities on each side are considered in isolation. We extended their description with relationships inside the image and the caption. Weegar et al. (2014) showed that pairs of entities in a text that were linked by the prepositions *on*, *at*, *with*, or *in*, often corresponded to pairs of segments that were close to each other. We further investigated the idea that spatial relationships in the image relate to syntactical relationships in the captions and we implemented it in the form of a reranker.

For each label-identifier pair, we included the relationship between the image segment in the pair and the closest segment in the image. As in Weegar et al. (2014), we defined the closeness as the Euclidean distance between the gravity centers of the bounding boxes of the segments. We also added the relationship between the caption entity in the label-identifier pair and the entity mentions which were the closest in the caption. We parsed the captions and we measured the distance as the number of edges between the two entities in the dependency graph.

6.1 Spatial Features

The Segmented and Annotated IAPR TC-12 data set comes with annotations for three different types of spatial relationships holding between the segment pairs in each image: Topological, horizontal, and vertical (Hernández-Gracidas and Sucar, 2007). The possible values are *adjacent* or *disjoint* for the topological category, *beside* or *horizontally aligned* for the horizontal one, and finally *above*, *below*, or *vertically aligned* for the vertical one.

6.2 Syntactic Features

The syntactic features are all based on the structure of the sentences' dependency graphs. We followed the graph from the caption-entity in the pair to extract its closest ancestors and descendants. We only considered children to the right of the candidate. We also included all the prepositions between the entity and these ancestor and descendant.

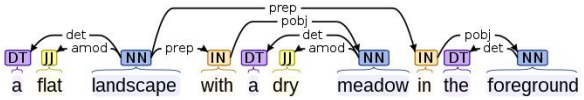

Figure 2: Dependency graph of the sentence *a flat landscape with a dry meadow in the foreground*

Figure 2 shows the dependency graph of the sentence *a flat landscape with a dry meadow in the foreground*. The descendants of the *landscape* entity are *meadow* and *foreground* linked respectively by the prepositions *with* and *in*. Its ancestor is the *root* node and the distance between *landscape* and *meadow* is 2. The syntactic features we extract for the entities in this sentence arranged in the order ancestor, distance to ancestor, preposition, descendant, distance to descendant, and preposition are for *landscape*, (root, 1, null, meadow, 2, with) and (root, 1, null, foreground, 2, in), for *meadow*, (landscape, 2, with, null, –, null), and for *foreground*, (landscape, 2, in, null, –, null). We discard *foreground* as it is part of the stop words.

6.3 Pairing Features

The single features consist of the label, entity identifier, and score of the pair. To take interaction into account, we also paired features characterizing properties across image and text. The list of these features is (Table 3):

1. The label of the image region and the identifier of the caption entity. In Fig 2, we create `grass_meadow` from *(grass, meadow)*.

2. The label of the closest image segment to the ancestor of the caption entity. The closest

Scoring function	Average rank	Rank = 1	Rank ≤ 2	Rank ≤ 3	Rank ≤ 4
co-occurrence	1.58	338	525	609	667
PMI	1.61	340	527	624	673
t-scores	1.59	337	540	623	669
PATH	1.19	559	604	643	668
HSO	1.18	**574**	**637**	**666**	**691**
JCN	1.22	535	580	626	653
LCH	1.19	559	604	643	668
LESK	1.19	560	609	646	670
LIN	1.19	542	581	623	652
RES	**1.17**	559	611	638	665
WUP	1.21	546	599	640	663

Table 2: Average rank of the correct candidate obtained by each scoring function on the 200 annotated images of the test set, and number of correct candidates that are ranked first, first or second, etc. The ceiling is 730

`Label`: Simplified segment label	`Entity`: Identifier for the caption entity	`Label_Entity`: Label and entity features combined
`Score`: Score given by the current scoring function	`Anc_ClosestSeg`: Closest segment label with the ancestor of the caption entity	`Desc_ClosestSeg`: Closest segment label with the descendant of the caption entity
`AncDist`: Distance between the ancestor and the caption entity, and distance between segments	`DescDist`: Distance between the descendant and the caption entity, and distance between the segments	`TopoRel_DescPreps`: Topological relationship between segments and the prepositions linking the caption entity with its descendant
`TopoRel_AncPreps`: Topological relationship between the segments and the prepositions linking the caption entity with its ancestor	`XRel_DescPreps`: Horizontal relationship between segments and the prepositions linking the caption entity with its descendant	`XRel_AncPreps`: Horizontal relationship between segments and the prepositions linking the caption entity with its ancestor
`YRel_DescPreps`: Vertical relationship between segments and the prepositions linking the caption entity with its descendant	`YRel_AncPreps`: Vertical relationship between segments and the prepositions linking the caption entity with its ancestor	`SegmentDist`: Distance (in pixels) between the gravity center of the bounding boxes framing the two closest segments

Table 3: The reranking features using the current segment and its closest segment in the image

segment of the *grass* segment is *river* and the ancestor of *meadow* is *landscape*. This gives the paired feature `meadow_landscape`. The labels of the segments closest to the current segment and the descendant of *meadow* are also paired.

3. The distance between the segment pairs in the image divided into seven intervals with the distance between the caption entities. We measured the distance in pixels since all the images have the same pixel dimensions.

4. The spatial relationships of the closest segments with the prepositions found between their corresponding caption entities. The segments *grass* and *river* in the image are *adjacent* and *horizontally aligned* and *grass* is located *below* the segment labeled *river*. Each of the spatial features is paired with the prepositions for both the ancestor and the descendant.

We trained the reranking models from the pairs of labeled segments and caption entities, where the correct mappings formed the positive examples and the rest, the negative ones. In Fig. 1, the mapping *(grass, meadow)* is marked as correct for the region labeled *grass*, while the mappings *(grass, lagoon)* and *(grass, cloud)* are marked as incorrect. We used the manually annotated images (200 images, Table 1) as training data, a leave-one-out cross-validation, and L2-regularized logistic regression from LIBLINEAR (Fan et al., 2008). We applied a cutoff of 3 for the list of candidates in the reranking and we multiplied the original score of the label-identifier pairs with the reranking probability.

6.4 Reranking Example

Table 4, upper part, shows the two top candidates obtained from the co-occurrence scores for the

Label	Entity 1	Score	Entity 2	Score
cloud	sky	2207	cloud	1096
grass	sky	1489	meadow	887
hill	sky	861	cloud	327
river	sky	655	cloud	250
cloud	cloud	769	sky	422
grass	meadow	699	landscape	176
hill	landscape	113	cloud	28
river	cloud	37	meadow	10

Table 4: An example of an assignment before (upper part) and after (lower part) reranking. The caption entities are ranked according to the number of co-occurrences with the label. We obtain the new score for a label-identifier pair by multiplying the original score by the output of the reranker for this pair

four regions in Fig. 1. The column *Entity 1* shows that the scoring function maps the caption entity *sky* to all of the regions. We created a reranker's feature vector for each of the 8 label-identifier pairs. Table 5 shows two of them corresponding to the pairs *(grass, sky)* and *(grass, meadow)*. The pair *(grass, meadow)* is a correct mapping, but it has a lower co-occurrence score than the incorrect pair *(grass, sky)*.

In the cross-validation evaluation, we applied the classifier to these vectors and we obtained the reranking scores of 0.0244 for *(grass, sky)* and 0.79 for *(grass, meadow)* resulting in the respective final scores of 36 and 699. Table 4, lower part, shows the new rankings, where the highest scores correspond to the associations: *(cloud, cloud)*, *(grass, meadow)*, *(hill, landscape)*, and *(river, cloud)*, which are all correct except the last one.

7 Results

7.1 Individual Scoring Functions

We evaluated the three scoring functions: Co-occurrence, mutual information, and t-score, and the semantic similarity functions. Each labeled segment in the annotated set was assigned the caption-entity that gave the highest scoring label-identifier pair.

To confront the lack of annotated data we also investigated a self-training method. We used the statistical associations we derived from the training set and we applied the mapping procedure in Sect. 5.3 to this set. We repeated this procedure

Feature	*(grass, meadow)*	*(grass, sky)*
Label	grass	grass
Entity	meadow	sky
Label_Entity	grass_meadow	grass_sky
Score	881	1,477
Anc_ClosestSeg	landscape_river	cloud_river
Desc_ClosestSeg	lagoon_river	null_river
AncDist	2_a	2_a
DescDist	1_a	100_a
TopoRel_DescPrep	adj_null	adj_null
TopoRel_AncPrep	adj_with	adj_in
XRel_DescPrep	horiz_null	horiz_null
XRel_AncPrep	horiz_with	horiz_in
YRel_DescPrep	below_null	below_null
YRel_AncPrep	below_with	below_in
SegmentDist	24	24
Classification	correct	incorrect

Table 5: Feature vectors for the pairs *(grass, meadow)* and *(grass, sky)*. The ancestor distance 2_a means that there are two edges in the dependency graph between the words *meadow* and *landscape*, and *a* represents the smallest of the distance intervals, meaning that the two segments *grass* and *river* are less than 50 pixels apart

with the three statistical scoring functions. We counted all the mappings we obtained between the region labels and the caption identifiers and we used these counts to create three new scoring functions denoted with a $\sum$ sign.

Table 6 shows the performance comparison between the different functions. The second column shows how many correct mappings were found by each function. The fourth column shows the improved score when the stop words were removed. The removal of the stop words as entity candidates improved the co-occurrence and t-score scoring functions considerably, but provided only marginal improvement for the scoring functions based on semantic similarity and pointwise mutual information. The percentage of correct mappings is based on the 730 regions that have a matching caption entity in the annotated test set.

The semantic similarity functions – PATH, HSO, JCN, LCH, LESK, LIN, RES and WUP – outperform the statistical one and the self-trained versions of the statistical scoring functions yield better results than the original ones.

We applied an ensemble voting procedure with the individual scoring functions, where each function was given a number of votes to place on its preferred label-identifier pair. We counted the votes and the entity that received the majority of the votes was selected as the mapping for the current label. Table 7 shows the results, where

Function	With stop words # correct	%	Without stop words # correct	%
co-oc.	208	28.5	338	46.3
PMI	339	46.4	340	46.6
t-score	241	33.0	337	46.1
$\sum$ co-oc.	226	30.0	387	53.0
$\sum$ PMI	457	62.6	458	62.7
$\sum$ t-score	247	33.8	397	54.4
PATH	552	75.6	559	76.6
HSO	**562**	**77.0**	**574**	**78.6**
JCN	527	72.2	535	73.3
LCH	552	75.6	559	76.6
LESK	549	75.2	560	76.7
LIN	532	72.9	542	74.2
RES	539	73.8	559	76.6
WUP	540	74.0	546	74.8

Table 6: Comparison of the individual scoring functions. This test is performed on the annotated set of 200 images, with 730 possible correct mappings

we reached a maximum 79.45% correct mappings when all the functions were used together with one vote each.

Scoring function	Number of votes		
co-oc.	1	0	1
PMI	1	0	1
t-score	1	0	1
$\sum$ co-oc.	1	0	1
$\sum$ PMI	1	0	1
$\sum$ t-score	1	0	1
PATH	0	1	1
HSO	0	1	1
JCN	0	1	1
LCH	0	1	1
LESK	0	1	1
LIN	0	1	1
RES	0	1	1
WUP	0	1	1
number correct	382	569	580
percent correct	52	78	**79**

Table 7: Results of ensemble voting on the annotated set

7.2 Reranking

We reranked all the scoring functions using the methods described in Sect. 6. We used the three label-identifier pairs with the highest score for each segment and function to build the model and we also reranked the top three label-identifier pairs for each of the assignments. Table 8 shows the results we obtained with the reranker compared to the original scoring functions. The reranking procedure improves the performance of all the scoring functions, especially the statistical ones, where the maximal improvement reaches 58%.

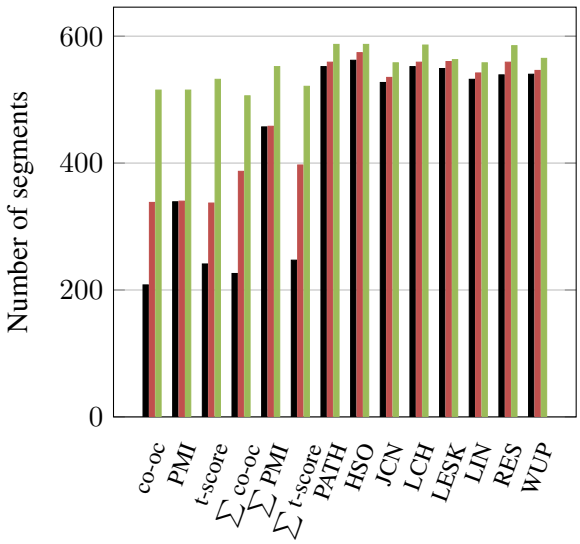

Figure 3: A comparison of the number of correctly assigned labels when using the different scoring functions. The leftmost bars show the results of the original functions, the middle bars show the performance when the stop words are removed, and the rightmost ones show the performance of the reranked functions

Function	correct	correct rerank.	% Improv.
co-oc.	338	515	52.4
PMI	340	515	51.5
t-score	337	532	**57.9**
$\sum$ co-oc.	387	506	30.7
$\sum$ PMI	458	**552**	20.5
$\sum$ t-score	397	521	31.2
PATH	559	**587**	5.0
HSO	574	**587**	2.3
JCN	535	558	4.3
LCH	559	586	4.8
LESK	560	563	0.5
LIN	542	558	3.0
RES	559	585	4.7
WUP	546	565	3.5

Table 8: The performance of the reranked scoring functions compared to the original scoring functions

Figure 3 shows the comparison between the original scoring functions, the scoring functions without stop words, and the reranked versions. There is a total of 928 segments, where 730 have a matching entity in the caption.

We applied an ensemble voting with the reranked functions (Table 9). Reranking yields a significant improvement for the statistical scoring

functions. When they get one vote each in the ensemble voting, the results increase from 52% correct mappings to 75%. When used in an ensemble with the semantic similarity scoring functions, the results improve further.

Scoring function	Number of votes		
Reranked co-oc.	1	0	1
Reranked PMI	1	0	1
Reranked t-score	1	0	1
Reranked $\sum$ co-oc.	1	0	1
Reranked $\sum$ PMI	1	0	1
Reranked $\sum$ t-score	1	0	1
Reranked PATH	0	1	1
Reranked HSO	0	1	1
Reranked JCN	0	1	1
Reranked LCH	0	1	1
Reranked LESK	0	1	1
Reranked LIN	0	1	1
Reranked RES	0	1	1
Reranked WUP	0	1	1
number correct	546	594	633
percent correct	75	81	**87**

Table 9: Results of ensemble voting with reranked assignments segments

We also evaluated ensemble voting with different numbers of votes for the different functions. We tested all the permutations of integer weights in the interval {0,3} on the development set. Table 10 shows the best result for both the original assignments and the reranked assignments on the test set. The reranked assignments gave the best results, 88.76% correct mappings, and this is also the best result we have been able to reach.

8 Conclusion and Future Work

The extraction of relations across text and image is a new area for research. We showed in this paper that we could use semantic and statistical functions to link the entities in an image to mentions of the same entities in captions describing this image. We also showed that using the syntactic structure of the caption and the spatial structure of the image improves linking accuracy. Eventually, we managed to map correctly nearly 89% of the image segments in our data set, counting only segments that have a matching entity in the caption.

The semantic similarity functions form the most accurate mapping tool, when using functions in isolation. The statistical functions improve significantly their results when they are used in an ensemble. This shows that it is preferable to use multiple scoring functions, as their different properties contribute to the final score.

	Number of votes	
Scoring function	Original	Reranked
co-oc.	0	0
PMI	2	3
t-score	0	0
$\sum$ co-oc.	0	1
$\sum$ PMI	2	1
$\sum$ t-score	0	1
PATH	1	1
HSO	2	3
JCN	0	0
LCH	0	0
LESK	1	0
LIN	2	0
RES	0	0
WUP	0	0
number correct	298	316
percent correct	83.71	**88.76**

Table 10: Results of weighted ensemble voting.

Including the syntactic structures of the captions and pairing them with the spatial structures of the images is also useful when mapping entities to segments. By training a model on such features and using this model to rerank the assignments, the ordering of entities in the assignments is improved with a better precision for all the scoring functions.

Although we used images manually annotated with segments and labels, we believe the methods we described here can be applied on automatically segmented and labeled images. Using image recognition would then certainly introduce incorrectly classified image regions and thus probably decrease the linking scores.

Acknowledgments

This research was supported by Vetenskapsrådet under grant 621-2010-4800, and the *Det digitaliserade samhället* and eSSENCE programs.

References

Satanjeev Banerjee and Satanjeev Banerjee. 2002. An adapted Lesk algorithm for word sense disambiguation using Wordnet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145.

Kenneth Church and Robert Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA, October. Association for Computational Linguistics.

Hugo Jair Escalantea, Carlos A. Hernándeza, Jesus A. Gonzaleza, A. López-Lópeza, Manuel Montesa, Eduardo F. Moralesa, L. Enrique Sucara, Luis Villaseñora, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114:419–428.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Robert Fano. 1961. *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 363–370, Ann Arbor.

Carlos Arturo Hernández-Gracidas and Luis Enrique Sucar. 2007. Markov random fields and spatial information to improve automatic image annotation. In Domingo Mery and Luis Rueda, editors, *PSIVT*, volume 4872 of *Lecture Notes in Computer Science*, pages 879–892. Springer.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, abs/1406.5679.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT press, Cambridge, MA.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, November.

Iftekhar Naim, Young Chol Song, Qiguang Liu, Henry Kautz, Jiebo Luo, and Daniel Gildea. 2014. Unsupervised alignment of natural language instructions with video segments. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-14)*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*.

Hideki Shima. 2014. WordNet Similarity for Java, February.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the HLT-NAACL*, pages 252–259, Edmonton.

Rebecka Weegar, Linus Hammarlund, Agnes Tegen, Magnus Oskarsson, Kalle Åström, and Pierre Nugues. 2014. Visual entity linking: A preliminary study. In *Proceedings of the AAAI 2014 Workshop on Cognitive Computing for Augmented Human Intelligence*, pages 46–49, Québec, July 27.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.