Visual Entity Linking: A Preliminary Study

Rebecka Weegar, Linus Hammarlund,
Agnes Tegen, Magnus Oskarsson, Kalle Åström, Pierre Nugues
Lund University
rebecka.weegar@gmail.com, linus.hammarlund@gmail.com, agnes.tegen@gmail.com,
magnuso@maths.lth.se, kalle@maths.lth.se, pierre.nugues@cs.lth.se

Abstract

In this paper, we describe a system that jointly extracts entities appearing in images and mentioned in their accompanying captions. As input, the entity linking program takes a segmented image together with its caption. It consists of a sequence of processing steps: part-of-speech tagging, dependency parsing, and coreference resolution that enables us to identify the entities as well as possible textual relations from the captions. The program uses the image regions labelled with a set of predefined categories and computes WordNet similarities between these labels and the entity names. Finally, the program links the entities it detected across the text and the images. We applied our system on the Segmented and Annotated IAPR TC-12 dataset that we enriched with entity annotations and we obtained a correct assignment rate of 55.48%

Introduction

Many images that appear in information media, ranging from the web to books and newspapers, are associated with short textual descriptions or captions. Although it is possible to extract a set of entity mentions from a caption, this set would refer to the image as a whole, not to a specific segment or region that would correspond to one specific entity.

To the best of our knowledge, very little work has been done on linking entity mentions extracted from a text to their segmented counterpart in images. Elliott and Keller (2013) is an exception that uses relations between regions in an image to improve its description. Examples of such relations are *on*, *besides*, and *surrounds*.

We developed a program to match the entities mentioned in a caption with segments in the image it accompanies. We focused on the language processing part and we set aside the segmentation step, which detects and labels regions in an image. We used segmented and labeled images as input.

We also carried out experiments to connect the textual relationships between entities in the captions, and the spatial ones between their matching segments in the images. Our assumption is that the words in the captions related through prepositions, for instance, should correspond to segments in the image that are close to each other.

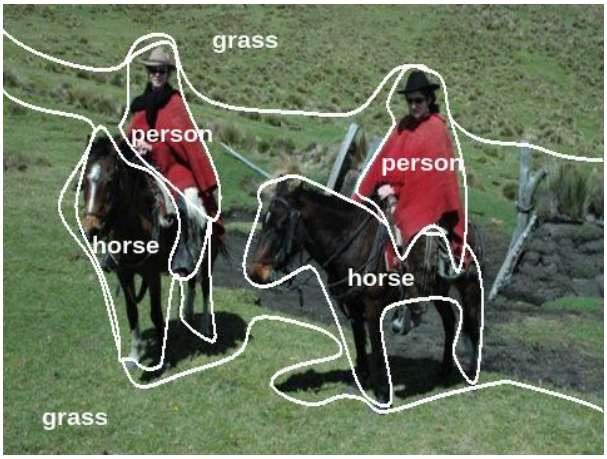

Figure 1: An image from the SAIAPR TC-12 dataset with six segments labeled as 1: horse, 2: grass, 3: horse, 4: person, 5: person, 6: grass with the caption *Two people with red capes are riding on dark brown horses on a green meadow.*

Possible applications of a visual entity linker include improved image search and image classification, learning tools for language studies, and the generation of image descriptions.

Dataset and Annotation

We used the *Segmented and Annotated IAPR TC-12 dataset* (SAIAPR TC-12) (Escalantea et al. 2010), which contains about 20,000 segmented images. Each image is annotated with a short text. Figure 1 shows an example from this dataset.

The images in the SAIAPR TC-12 dataset are manually segmented and the resulting segments are annotated using a set of 275 predefined categories such as man, sky, sky-blue, viola, etc. These categories and the captions were created independently with no correspondence between each other. This means that the entities in a caption do not necessarily reflect or match the segments and *vice versa* and that the captions do not borrow the vocabulary used to name the categories.

For this work, we used two vocabularies to name the segment categories: The original one defined by the SAIAPR TC-12 dataset that consists of the 275 words and that we refer to as the *labels*. In addition, we clustered 100 of the most frequent of these labels into 13 groups that we called the *cluster labels* and that define a second vocabulary: water, sky, vegetation, construction, human, house objects, ground, animal, vehicle, mountain, road, floor, and fabrics (Tegen et al. 2014).

As the SAIAPR TC-12 dataset does not provide a gold standard on links between the entities in the captions and the segments in the image, we manually annotated two sets for the evaluation: One to evaluate the noun extraction and the other to evaluate the entity linking. We used the first 40 images of the dataset to build this test set.

System and Algorithms

The entity linking system consists of three parts:

1. A mention detector that extracts the entities from the caption of an image (Sect. *Extracting Entities from Captions*).
2. A segment extractor. In this experiment, we considered that the entities on the image corresponded to the manually annotated segments provided by the SAIAPR TC-12 dataset. We carried out the entity linking evaluation with these segments. For an evaluation of an automatic segment extraction, see Tegen et al. (2014).
3. An entity linker that relates the extracted nouns with the labelled segments.

Extracting Entities from Captions

The caption of an image in the dataset usually consists of a sequence of one to three sentences, each separated by a semicolon. To link entities in the caption to segments in the image, we first extracted the entity mentions from the captions. We restricted these mentions to nouns that we identified using the Stanford CoreNLP program (Raghunathan et al. 2010; Lee et al. 2011) and, as entity identifier, we used their lemmatized form.

As an example, given the caption input:

> a small tower made of grey stones on a grey, rocky bank in the foreground; a light blue lake and a dark brown mountain behind it; a blue sky in the background.

we extracted the following entities: *lake*, *tower*, *stone*, *bank*, *foreground*, *mountain*, *sky*, and *background*.

When the caption contained noun sequences and these sequences were present as an entry in the WordNet dictionary, we combined them into a single string. This means that we merged the sequence *palm* followed by *tree* into the string "palm tree."

When the Stanford CoreNLP program found a coreference between entity mentions, we used, for each entity in the caption, the most representative mention provided by the program to identify it.

Extracting Pairs

Once the entities detected, we investigated their relationships in the captions. Prepositions often indicate some kind of spatial organization and we assumed that the words they link in a sentence could have a relationship in the picture too: Segments representing these words could, for example, be close to each other.

We extracted the pairs of nouns that were linked by a preposition using the Stanford CoreNLP dependency parser (de Marneffe, MacCartney, and Manning 2006). Figure 2 shows the dependency output for the sentence *Two people with red capes are riding on dark brown horses on a green meadow*, where we have the pairs: ***people*** *with* ***cape*** and ***horse*** *on* ***meadow***.

Similarity between Words

As, most of the time, the labels and the mentions used in the captions are different, we created a lexical distance to measure their compatibility using the WordNet ontology (Miller 1995).

WordNet is a lexical database built on the concept of synsets, where different synsets are linked to each other by semantic and lexical relations. We used this structure to measure the similarity between two words. We extracted the first common ancestor of the two words – the first common inherited hypernym – to compute a distance between a pair consisting of a label and a mention. As an example, the subgraph linking the words *boy* and *human* is shown below:

Tree for boy: **boy** → male → person → **organism**.

Tree for human: **human** → hominid → primate → placental → mammal → vertebrate → chordate → animal → **organism**.

The first common ancestor for the words *human* and *boy* is *organism*. The node distance between these two words is 12, resulting in a normalized distance of 1/12 = 0.0833.

We used the *WordNet Similarity* (Pedersen, Patwardhan, and Michelizzi 2004) to generate a matrix with distances between the 275 original segment annotation labels and the 2,934 unique nouns extracted from the SAIAPR TC-12 dataset. Table 1 shows the normalized distance between some of the words found in the dataset.

Label	Noun	Normalized distance
human	kid	0.1000
human	boy	0.0833
human	shoe	0.0667
construction	building	0.3333
construction	tower	0.5000
construction	kid	0.1250

Table 1: Distances for some of the nouns and segment labels in the dataset.

Linking Entities with Segments

The final step assigns a noun, possibly a `null`-token, to each segment of an image. It produces these pairs of segments and nouns using the algorithm below.

Using an input consisting of a segmented image and a caption, for each segment in the image, do the following steps:

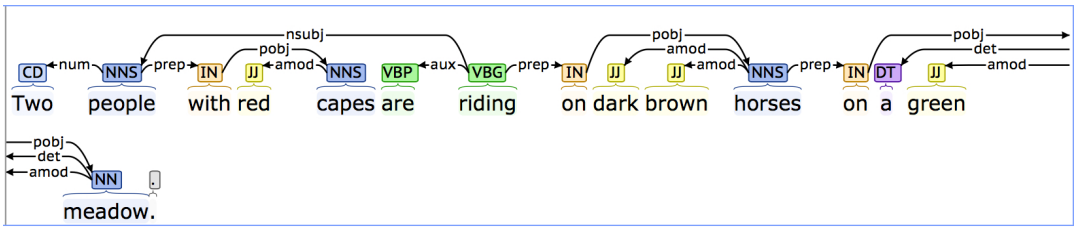

Figure 2: Output of the dependency parsing.

1. Extract the segment labels from the image and convert them into a cluster category.
2. Assign a cluster label to each mention extracted from the caption using the WordNet similarity.
3. Select the mentions that have the same cluster label as the segment.
4. In the case of multiple mentions with the same cluster category, assign the mention with the highest distance score to the segment.

Results and Evaluation

Building a Test Set

We used the 40 first images of the SAIAPR TC-12 dataset to evaluate our program. For each image-caption pair in the test set, we selected the nouns that describe an object in the image, without regard to the predefined image segments. In the case of a coreference chain, we associated the entities with the most representative mention. In Figure 1, we selected four nouns from the caption: *people*, *cape*, *horse*, and *meadow*. Note that the number of nouns differs from the number of segments as the image contains six annotated segments. In total, we extracted 289 nouns from the test set.

Entity Linking Annotation

For each segment in each image of the test set, we selected the noun in the caption that we judged the most relevant for this segment. In some cases, we could not find a matching noun and we left the segment unassigned. In addition, a same word could match multiple segments. We chose to allow one single word per segment.

In Figure 1, we annotated the segments with the following mentions: segment 1: *horse*, segment 2: *meadow*, segment 3: *horse*, segment 4: *people*, segment 5: *people*, segment 6: *meadow*. In total, we annotated 301 segments in the test set.

Evaluation of the Mention Extraction

We first compared the nouns automatically extracted by the parser with our manual annotation of the entity mentions. This extraction step detected 389 mentions in the captions of the test set. On average, this means that the system found 9.725 mentions per image whereas the manual annotation produced 7.3 mentions per image (292 in total). We computed the precision and recall for the whole set and we obtained a recall of 0.986 and a precision of 0.740 yielding a harmonic mean F_1 of 0.846.

The relatively low precision comes from extra words describing *where* something is located in the image. The two most common extra words are *background* (19 occurrences) and *foreground* (13 occurrences), or *front* as in or *rails in front of a tunnel*. Of the extra words, 61 (60.4%) fall into this category.

Another category of extra words is related to the photographic process. Someone in the picture may be *waving at the camera*.

Evaluation of Entity Linking

We evaluated the entity linking from a segment-oriented perspective. The algorithm uses gold segments as input and assigns each segment a mention. To evaluate the entity linking step, we manually assigned the segment labels using the cluster label categories and the original categories.

As a baseline for the evaluation, we used a random mention assignment, where each of the segments in the test set was assigned one of the extracted mentions, or a `null`-token, as that was a valid assignment as well. This resulted in a baseline score of 13.29%.

Our algorithm resulted in a score of 55.48% (167 correctly assigned mentions for 301 segments).

We also ran the algorithm using the 275 labels instead of the 13 cluster labels to see if the amount of labels used in the segment categorization would have any influence on the score. The test resulted in a score of 52.49% (158 correctly assigned mentions for 301 segments).

Evaluation of Noun Pairs

We evaluated the mentions pairs with the original segments in the SAIAPR TC-12 dataset. Considering twenty images from the dataset and using the prepositions *on*, *at*, *with*, and *in*, we extracted 61 mentions pairs. This yielded an average of 3.05 pairs per image.

As discussed previously, some of the mentions extracted from the captions did not represent any actual object in the image. We checked the 61 pairs and those containing such words were removed. The removed words were of the same types as in Sect. *Evaluation of the Mention Extraction*.

By only allowing words that represent something actually visible in the image, 31 of the 61 pairs remained, giving an average of 1.55 per image.

To evaluate the assumption that these mention pairs correspond to segments in the image that are close to each other, it was necessary that both words in the pair actually had a

matching segment in the image and to define a distance between two image segments. We used the Euclidean distance between the gravity centers of their bounding boxes.

Six of the pairs had matching segments in the images and of those six pairs, three had corresponding segments that were the closest according to the distance. Two of them had segments that were not considered the closest, but they were still adjacent to each other, and one pair was neither the closest to each other nor adjacent.

Since only few mention pairs had distinct segments corresponding to both mentions, we looked at the mention pairs that were covered by the same segment. Figure 1 shows an example of this, where the nouns *people* and *cape* form a pair. The segment covering the people also covers the cape while the cape has no segment of its own. We found that 21 of the 31 pairs were covered by the same segment. For five of the 31 pairs, one or both of the words did not have a matching segment.

Tables 2 and 3 show a summary of these results.

Description	**Freq.**	**Percent**
Pairs in test set	61	100
Both mentions are visible objects	31	51
Both mentions are visible and have a corresponding segment in the image	26	42

Table 2: Pairs found in test set.

Segment relationships	**Freq.**	**Percent**
Closest in Euclidean distance	3	11.5
Not closest but adjacent	2	7.7
Both objects covered by same segment	21	80.8

Table 3: Spatial relationships between segments corresponding to the 26 pairs in the test set, where both words have a matching segment or are covered by the same segment.

Conclusions and Future Work

In this paper, we presented a system that links entity mentions to preexisting, manually labeled segments with a 55.48% accuracy. This system uses a dependency parser, a mention detector and coreference solver, and the lexical distance between the mentions.

While we did not integrate the mentions pairs in the final system, the results of the experiments substantiate the assumption that mentions related by prepositions correspond to spatially related segments in the image. In the test set, 42% of the pairs relate to objects that are either adjacent to each other, closer to each other than to other segments, or even covered by a same segment. This hints at a possible improvement of the linking of entities in text to the segments using these pairs.

More information in the captions could be used to automatically classify the segments and link them to entities in the caption. An example of this is the color which is often described in the captions as in:

> a boy with a light blue cap, a red pullover, blue jeans and black shoes is standing in front of a pile of red bricks.

One limitation found in the current version is that it maps segments to nominal mentions. Changing the directions and mapping mentions to segments (or a combination of both) would give access to better processing options as well as more flexibility in how mentions and segments are chosen.

Acknowledgments

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

References

de Marneffe, M.-C.; MacCartney, B.; and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Elliott, D., and Keller, F. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Escalantea, H. J.; Hernándeza, C. A.; Gonzaleza, J. A.; López-Lópeza, A.; Montesa, M.; Moralesa, E. F.; Sucara, L. E.; Villaseñora, L.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding* 114:419–428.

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; and Jurafsky, D. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 28–34.

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

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet::similarity – measuring the relatedness of concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2004)*, 38–41.

Raghunathan, K.; Lee, H.; Rangarajan, S.; Chambers, N.; Surdeanu, M.; Jurafsky, D.; and Manning, C. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 492–501.

Tegen, A.; Weegar, R.; Hammarlund, L.; Oskarsson, M.; Jiang, F.; Medved, D.; Nugues, P.; and Åström, K. 2014. Image segmentation and labeling using free-form semantic annotation. In *Proceedings of the International Conference on Pattern Recognition*.