Visual Entity Linking

Linus Hammarlund
Lund University
Lund, Sweden
linus.hammarlund@gmail.com

Rebecka Weegar
Lund University
Lund, Sweden
rebecka.weegar@gmail.com

Abstract

We have developed a system for linking entities in captions with segments in their corresponding images. By using part-of-speech tagging, chunking and dependency parsing to extract entities and WordNet similarities we have been able to construct such an entity linking program. It has been successfully tried on the Segmented and Annotated IAPR TC-12 dataset, with a correct assignment rate of 55.48 %.

1 Introduction

Something that goes hand-in-hand with images is their captions. You seldom see an image without a caption. They appear together in almost all of our information channels, ranging from the web to books and newspapers. One could go as far as to say that subtitles are just captions for the frames in a movie.

Therefore it is interesting to study the matching of entities from captions with objects in their images. Possible applications in this field is for example improved image searching where you parse the actual image instead of its caption, classification of images, learning tool for language studies and generation of image descriptions.

The goal of the project is to match entities in a caption with segments in its image using state-of-the-art natural language processing techniques. We are not developing a segment classifier which assigns labels to segment using image features, but rather using one to help us pair up parsed caption-entities with its labelled segments.

We have also done some experiments with relationship between different entities in the captions and the spatial relationships between their matching segments in the images. Previous work that has been done in this area includes (Elliott and Keller, 2013) which uses relations between regions in an image to improve its description. Examples of such relations are *on, besides* and *surrounds*. Our assumption has simply been that related words in the captions should correspond to segments in the image that are close to each other.

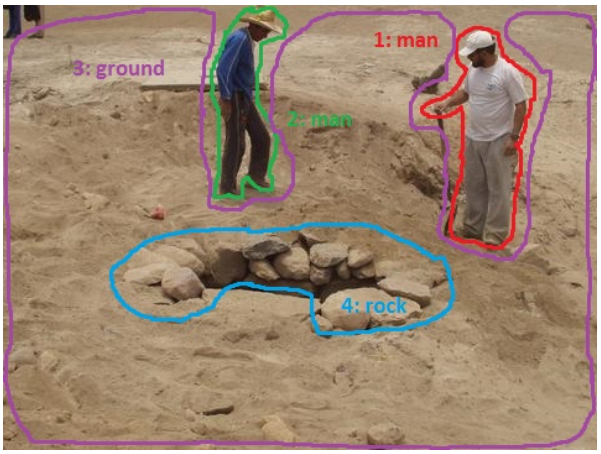

Figure 1: An image with four segments (1: man, 2: man, 3: ground and 4: rock), with the caption *"A rock hole in the ground surrounded by sand. One man in a white tee-shirt, grey pants and a white cap is holding a shovel, the other one is wearing black pants, a white jumper and a traditional hat, is walking towards him."*

We begin this paper by explaining the dataset and the segmentation. After that we continue with a description of the different parts of our algorithm. This is followed by an evaluation, conclusions and suggestions for future improvements.

2 Dataset and ground truths

For this project we have used the *Segmented and Annotated IAPR TC-12 dataset*, SAIAPR TC-12, (Escalante et al, 2009), which contains about 20 000 segmented and annotated images. We have limited ourselves to only using the English annotations. An example from the dataset can be found in figure 1.

The images in the SAIAPR TC-12 dataset are manually segmented and the resulting segments are annotated by using words from a predefined vocabulary. This means that the entities of the caption doesn't necessarily match the segments and vice versa, as they are defined with no correspondence between each other.

For this project we have also used the vocabulary defined by SAIAPR TC-12 as it is used by the classifier (outlined in section 3.1). It consists of 276 words. In this paper the words in the vocabulary are referred to as *labels*. These labels are in turn divided into 13 groups called *cluster labels*: water, sky, vegetation, construction, human, house-objects, ground, animal, vehicle, mountain, road, floor, and fabrics.

Since we for this project want to match entities from the captions (the annotated descriptions) with a corresponding segment, we have manually annotated two ground truth sets for our evaluation. One to evaluate the noun extraction and one to evaluate the entity linking. For this we have used the first 40 images of the dataset (the test set).

2.1 Entity extraction ground truth

For each image-caption pair in the test set we have selected the nouns that describes an object in the image, without regard to the predefined segments. We have defined an entity as a single noun. In figure 1 the caption reads

> *a rock hole in the ground surrounded by sand; one man in a white tee-shirt, grey pants and a white cap is holding a shovel, the other one is wearing black pants, a white jumper and a traditional hat, is walking towards him;*

where our manually selected nouns are *hole, ground, sand, man, tee-shirt, pants, cap, shovel, pants, jumper*, and *hat*. Note that the number of nouns differs from the number of segments. Here we have eleven nouns but there are only four segments in the image.

In total there were 287 nouns found for the first 40 images.

2.2 Entity linking ground truth

For each segment in each image in the test set a noun in the caption for that image was selected. The noun selected was the one that had the most correspondence according to us.

This means that some segments did not get a matched noun as the segments in the dataset were not made in accordance to the captions. Also multiple segments could be matched to the same word. Another restriction we chose to make was to only allow one single word to be match to a segment, so if a segment would match "man and woman" we only chose "man" as it was the first noun of that noun chunk.

So for figure 1 we got *1: man, 2: man, 3: ground*, and *4: hole* as our ground truth. In total, 301 segments were annotated for the test set.

3 System and algorithm

Our system is divided into three parts. First classify the segments using an external classifier as described in section 3.1. Then we extract the nouns (the entities) from the caption of an image in section 3.2. The third part of the system is linking the extracted nouns with the segments labelled by a classifier.

3.1 Classification of the segments

The segments of the images are classified with an external classifier using image features, such as color and shape. This was done using an early version of a classifier from the Centre for Mathematical Sciences, Lund University. For each segment in an image from the dataset, a list of probabilities for each of the cluster labels were generated.

As this project has been done in cooperation with the classifier-project, many of the steps described in this report were also used to generate data for the classifier (Tegen et al., 2014). This resulted in the other project developing a new version of the classifier that generated segments as well (as described in their paper). Hence we ended up using an older version for evalutation as we wanted the original segments from the SAIAPR TC-12 dataset.

We can see an example of the classification in figure 2 and in table 1. In the figure we can see a boy in front of a wall, where one of the segments was the wall. The probabilities for the cluster labels for that segment can be found in table.

3.2 Extracting entities from captions

The caption of an image in the dataset usually consists of one to three sentences, each separated by a semicolon. Since the goal of this project was to link entities in the caption to segments in the im-

Figure 2: An image of a boy where the wall in the background is one of the original segments from the SAIAPR TC-12 dataset.

Cluster label	Probability
construction	53,764868%
fabrics	30,169785%
house-objects	7,653020%
animal	5,584522%
sky	1,070917%
water	0,804009%
floor	0,374003%
mountain	0,215209%
vegetation	0,210343%
road	0,075258%
human	0,031529%
vehicle	0,029420%
ground	0,017118%

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

age, we needed to find words in the caption that could represent some object in the image.

This was done by dividing the captions into noun and verb chunks using the *Illinois Chunker* (Punyakanok and Roth, 2014) and the *Illinois Part of Speech Tagger* (Roth and Zelenko, 2014)

From the noun chunks the rightmost noun of each chunk has been extracted, assuming that that word would be the the main word of the chunk.

A typical sentence from an image caption is

> *a dark-skinned boy wearing a blue cap and a yellow jumper is standing outside a house, having his hands in his pockets*

That sentence gets divided into the following chunks:

> *[NP a dark-skinned boy] [VP wearing] [NP a blue cap] and [NP a yellow jumper] [VP is standing] [PP outside] [NP a house] , [VP having] [NP his hands] [PP in] [NP his pockets]*

And the rightmost noun of each noun chunk (annotated NP) would then give the following entities/nouns to be linked with the segments in the image: *boy, cap, jumper, house, hands*, and *pockets*.

3.3 Extracting pairs

When the entities in the captions had been found, it was also possible to investigate the relationships between the them.

To do this, the captions for each picture were first tagged with their part of speech and divided into sentences. After that they were parsed with the dependency parser *MaltParser* (Hall et al., 2014). The output of the parser can be seen in figure 3.

1	a	DT	3	det
2	rock	NN	3	nn
3	hole	NN	0	null
4	in	IN	3	prep
5	the	DT	6	det
6	ground	NN	4	pobj
7	surrounded	VBN	6	partmod
8	by	IN	7	prep
9	sand	NN	8	pobj

Figure 3: Output of the dependency parsing. The columns are: Word counter, word, part of speech, head, dependency relationship to head

In the parsed sentences patterns where two nouns were linked by a preposition were extracted. Examples of such patterns were: *wall with gate*, *hole in ground* and *man in tee-shirt*. An example of this can be seen in figure 3

The reason for doing this was the idea that words that are linked in text might have a relationship in the picture also. Prepositions indicates some kind of spatial relationships and segments representing theses words could for example be close to each other in the picture.

The assumption that the found pairs corresponds to segments that are close to each other,

makes it necessary to decide how to determine which segments that are close. To do this, a bounding box for each segment was calculated, and distances between segments were measured as the euclidean distance between the gravity centres of each of the bounding boxes.

3.4 Similarity between words

The approach we used to see if two words corresponded to each other was using distances in the WordNet database (Princeton University, 2010). This was needed to match a noun against a label from the classifier.

WordNet is a lexical database built around the concept of synsets where different synsets are linked to each other by semantical and lexical relations. This structure can be used to find a measurement of similarity between two words. By finding the first common ancestor – the first common inherited hypernym – of the two words a distance can be computed for each pair of two words. An example of this can be found in table 2.

Tree for human	Tree for boy
organism	organism
↑	↑
animal	person
↑	↑
chordate	male
↑	↑
vertebrate	boy
↑	
mammal	
↑	
placental	
↑	
primate	
↑	
hominid	
↑	
human	

Table 2: First common ancestor for the two words human and boy. The distance between these two words is 12, resulting in the normalized distance $1/12 = 0.0833$.

For this project we used an implementation of the algorithm described above known as *WordNet Similarity* (Pedersen et al., 2004). This allowed us to generate a matrix with distances between the 276 labels from the classifier and the 2934 unique nouns extracted from the SAIAPR TC-12 dataset.

In table 3 we see the normalized distance between some of the words found in the dataset.

Label	Noun	Normalized dist.
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 3: Distances for some of the nouns and labels in the dataset.

3.5 Linking entities with segments

The final part of our system assigned a noun (or a `NULL`-token) to each segment of an image. It produces pairs of segment-noun for an image using the algorithm described below.

For each segment in an image, do the follow steps.

1. Get the cluster label assigned to the segment by the classifier (take the one with the highest probability).
2. Assign a cluster label to each noun extracted from the caption for the image.
3. Select the nouns that have the same cluster label as the segment.
4. Assign the noun with the highest distance score to the segment.

The usage of noun pairs did not make it into the final version of our algorithm due to time constraints but is used for the classifier as mentioned in section 3.1.

4 Results and evaluation

For the evaluation of the project we used the 40 first images of the SAIAPR TC-12 dataset as described in section 2, using the ground truths described in section 2.1 and 2.2.

4.1 Evaluation of word extraction

The extracted nouns were compared to the nouns of the ground truth for noun extraction.

The automatic noun extraction part of our system found 364 words for the first 40 images used in the ground truth. On average this means that

the system found 9.1 words per image whereas the ground truth has 7.175 words per image.

The bigger part of the extra words found are words that describe *where* something is located in the image. The two most common extra words are *background* (18 occurrences) and *foreground* (11 occurrences), as in *a blue sky in the background*, or *rails in front of a tunnel*, where front gets tagged as a noun. 63 (72 %) of the extra words falls into this category.

Another category of the extra words is words related to to photography. Someone in the picture may be *waving at the camera*. Another example is *a close up picture*. Neither the camera or the picture is visible in the actual image.

Also, colours have been tagged as nouns three times and 11 of the extra words are not falling into any specific category.

The automatic extraction failed to find 10 of the words (3.5%) that the are present in the ground truth. The main reason for missing a word is that it is mistakenly not tagged as a noun. Examples of this are the words reed, streetlamps and skirts, which were tagged as verbs.

Figure 5 shows a comparison of number of found nouns in the manual and automatic extraction for each picture in the test set.

Even if some words were missed, the automatic extraction found all the words in the ground truth for 80 % of the pictures in the test set.

4.2 Evaluation of entity linking

The entity linking were evaluated from the segment-perspective. Each segment is assigned a noun by our algorithm, thus assuming that we have the correct segments given for each image.

As a baseline for the evaluation we chose random noun assignment. Each of the segments in the test set were assigned one of the extracted nouns (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 nouns for 301 segments). For the purpose of this paper the evaluation was done using a "manual classifier" where each segment was manually assigned one of the cluster label, thus simulating the classifier mentioned in section 3.1. This was done to rule out the variances that appeared due to faults in the classifier.

We also ran the algorithm using the 276 labels instead of the 13 cluster labels to see if the amount of labels used by the classifier had any result on the score. The test resulted in a score of 52.49% (158 correctly assigned nouns for 301 segments), thus indicating that a better classifier will not generate a better result with our current algorithm.

The results are shown in detail in figure 4.

4.3 Evaluation of noun pairs

The evaluation of the noun pairs have been done against the original segments in the SAIAPR TC-12 data set.

By using the prepositions *on, at, with*, and *in* and looking at twenty of the images in the data set, 61 noun pairs were found. This gave an average of 3.05 pairs per image.

As discussed previously, some of the nouns that were extracted from the captions did not represent any actual object in the image. The 61 pairs were checked and pairs that contained such words were removed. The removed words were of the same types as described in section 3.2

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

To evalute if the assumption that these pairs of words corresponds 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. This was the case for six of the pairs. Of those six pairs, three had corresponding segments that were closest according to the euclidean distance between the gravity centres of their bounding boxes. Two of them had segments that were not considered the closest, but they were still adjacent to each other, and one pair was neither closest to each other nor adjacent.

Since there were so few noun pairs where there existed segments corresponding to both words, we also looked at how many of the noun pairs that were covered by the same segment. An example of this can be seen in Figure 1. The nouns *man* and *tee-shirt* forms a pair. But the segment covering the man man wearing the tee-shirt also covers the tee-shirt, the shirt has no segment of its own. 21 of the 31 pairs are covered by the same segment.

For five of the 31 pairs, one or both of the words did not have a matching segment.

Table 4 and table 5 shows a summary of these results.

Description	Amount	Percent
Pairs in test set	61	100%
Both words in pair are visible objects	31	51%
Both words are visible and have a corresponding segment in the image	26	42%

Table 4: Pairs found in test set.

Description of segment relationships	Amount	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 5: 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.

5 Conclusions and future work

In this paper we have shown that it is possible to link nouns with pre-existing segments with a 55.40 % hitrate. This has been done using chunking, lexical distance between words and dependency parsing. The algorithm and system developed can be used on the entire SAIAPR TC-12 dataset and can easily be extended to other image sets.

There is still room for a lot of improvement on our algorithm. For example the noun pairs did not make it into the final version. The results of the experiments indicates that there is truth to the assumption that words related by prepositions corresponds to spatially related segments in the image. In the test set, 83% of the pairs that are related to objects in the picture, are adjacent to each other, closer to each other than to other segments, or even covered by the same segment.

This tells us that it is likely that the pairs could be used to improve classification of segments and the linking of entities in text to the segments . It could be done by looking for pairs in the captions and trying to find segments in the image that are close to each other that matches the words in the pairs.

For this project, only the prepositions *in, on, at* and *with* have used. They only tell that two entities have some kind of relationship. It is of course also possible to find prepositions like *above, below, over, under* etc., that also tell how the objects in the picture are placed with relation to each other.

With a more detailed segmentation the number of matching pairs could perhaps be higher since some of the words in the pairs related to an object in the picture, but that object was not covered by a segment and could therefore not be matched

By extending the parsing patterns other types of relationships can also be found. For example, by allowing for patterns where two nouns are linked by longer chains of prepositions, relationships such as *in middle of*, *at top of* and *in front of* can also be found.

Also, there is more information in the captions that could be used for classification of segments and for ng segments to entities in the caption. An example of this is colour which is often present in the captions in the data set. Examples of this is *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*

Using the same data that is used in this paper one could also restructure the algorithm using the distance score differently, or use another distance algortihm as there are other ones available in WordNet Similarity.

One limitation found in the current version is that you can only map nouns to segments. Changing it to map segments to nouns (or a combination of them both) would give you access to better reranking options as well as more flexibility in what nouns and segments you choose.

Acknowledgments

We would like to thank our supervisor Pierre Nugues for his support and inputs during the project, and Dennis Medved at the Department of Computer Science for valuable insights and discussions.

We would also like to thank the classifer-group consisting of Agnes Tegen, Kalle Åström, Magnus Oskarsson, Jiang Fangyuan and the two people mentioned above, at the Centre for Mathematical Sciences for great discussions, inspiration and inputs on our system

References

Elliott, D., and Keller, F.. *Image Description using Visual Dependency Representations*. 2013. School of

Informatics, University of Edinburgh.

Escalante, H. J., Hernndez, C., Gonzalez, J.., Lpez, A., Montes, M., Morales, E., Sucar, E., , L., Grubinger, M.. *The Segmented and Annotated IAPR TC-12 Benchmark*. Computer Vision and Image Understanding, doi:10.1016/j.cviu.2009.03.008, 2009.

Hall, J., Nilsson, J., and Nivre, J.. *MaltParser - a data-driven dependency parser (Version 1.7.1).*. (Computer program) Available at: http://www.maltparser.org. (Accessed 14 November 2013).

Pedersen, T., Patwardhan, S., and Michelizzi, J.. *WordNet::Similarity - Measuring the Relatedness of Concepts.* 2004. Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2004).

Princeton University. *"About WordNet."* WordNet. 2010. Princeton University. http://wordnet.princeton.edu

Punyakanok, V., and Roth, D.. *Illinois Chunker*. (Computer program). Available at: http://cogcomp.cs.illinois.edu/page/software. (Accessed 2 November 2013).

Roth, D., and Zelenko, D.. *Illinois Part of Speech Tagger.*. (Computer program) Available at: http://cogcomp.cs.illinois.edu/page/software. (Accessed 2 November 2013).

Tegen, A., Weegar, R., Hammarlund, L., Oskarsson, M., Fangyuan, J., Medved, D., Nugues, P., and strm, K.. *Image Segmentation and Labeling Using Free-form Semantic Annotation*. 2014. Centre for Mathematical Sciences and Department of Computer Science, Lund University.

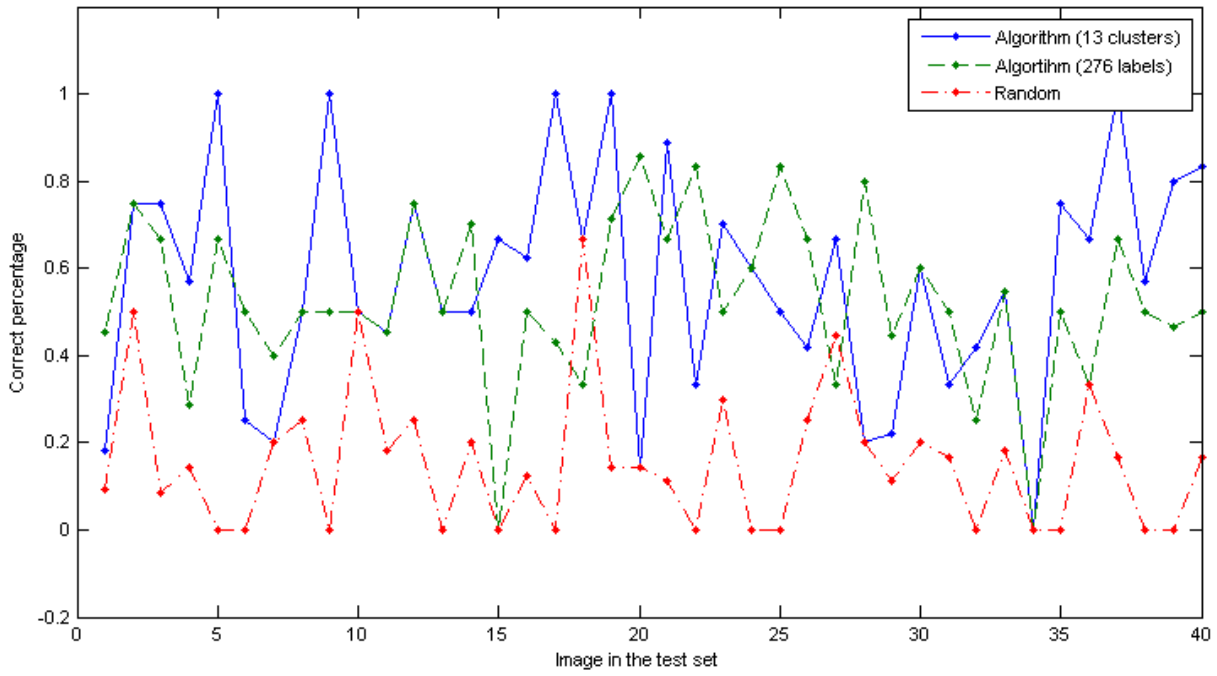

Figure 4: Three graphs showing the result for the evaluation of the entity linking step. One showing the baseline (random), one showing our algorithm and one showing our algorithm with more labels.

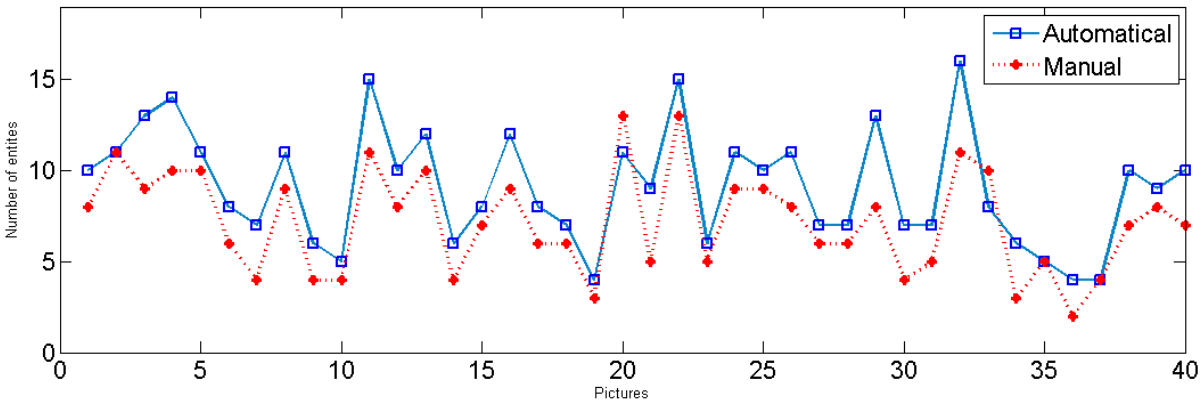

Figure 5: Number of entities found in captions, manual and automatic extraction.