Visual disambiguation of prepositional phrase attachments: multimodal machine learning for syntactic analysis correction*

Sebastien Delecraz[1], Leonor Becerra-Bonache[2], Alexis Nasr[1], Frederic Bechet[1], and Benoit Favre[1]

[1] Aix-Marseille Univ, Université de Toulon, CNRS, LIS, UMR 7020, Marseille, France
sebastien.delecraz@univ-amu.fr, alexis.nasr@univ-amu.fr, frederic.bechet@univ-amu.fr, benoit.favre@univ-amu.fr

[2] Univ Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien, UMR 5516, Saint-Étienne, France
leonor.becerra@univ-st-etienne.fr

Abstract. Prepositional phrase attachments are known to be an important source of errors in parsing natural language. In some cases, pure syntactic features cannot be used for prepositional phrase attachment disambiguation while visual features could help. In this work, we are interested in the impact of the integration of such features in a parsing system. We propose a correction strategy pipeline for prepositional attachments using visual information, trained on a multimodal corpus of images and captions. The evaluation of the system shows us that using visual features allows, in certain cases, to correct the errors of a parser. It also helps to identify the most difficult aspects of such integration.

Keywords: Prepositional Phrase Attachments · Multimodality · Correction Strategy

1 Introduction

Natural languages are intrinsically ambiguous. Some of these ambiguities can be solved by using only syntactic information, but many others require access to the context in which they have been produced. For instance, the ambiguity involved in the famous example of "John saw a man with a telescope" could be easily solved if we could have access to that scene.

Prepositional phrase (PP) attachments are a common source of ambiguity. They constitute one of the most difficult constructions to deal with [20], representing around 20% of errors in parsing natural languages [18]. The main difficulty lies in the fact that this kind of ambiguity can often not be solved using only linguistic cues (as shown in the previous example). Even if it can

* The work of Leonor Becerra-Bonache has been performed during her teaching leave granted by the CNRS (French National Center for Scientific Research) in Laboratoire d'Informatique et Systèmes of Aix-Marseille University.

be overcame, a parser often does not have semantic constraints to rule out incorrect attachments. However, adding information from a visual source to the analysis of a sentence is a difficult task because it is necessary to extract relevant information from that source and to be able to relate this information to the sentence.

The work that we present in this article has a twofold objective. First, to propose a method for resolving PP-attachments based on the use of visual cues. Second, to analyze the impact of adding visual information on the task of syntactic analysis. To do this, we use a corpus made up of pairs of an image and a caption describing it. This corpus has been annotated manually at different levels. At the image level, rectangles (which we will call boxes) have been identified and a semantic category has been associated with each of them. At the text level, some noun phrases have been identified, as well as some PP-attachments for a subset of frequent prepositions. In addition, boxes corresponding to noun phrases were aligned to the latter. The fact of having simultaneously the analysis of the image (via boxes) and the text (through certain PP-attachments), as well as the alignment between boxes and noun phrases makes it possible to establish a link between the two modalities and to use information from the image to process the text.

The system we propose is based on an attachment error detector, that offers an alternative attachment if it detects an error. The originality of this detector is that it allows to take lexical, but also visual and conceptual clues as input. For example, in the noun phrase *a ball in_front_of a dog with a red collar*, the decision to attach *with* to *dog* rather than *ball* may be based on obvious lexical evidence, but could also be based on visual clues by studying, for example, the relative positions of the boxes corresponding to the words *ball*, *dog* and *collar*.

The paper is organized as follows. Related work is presented in Section 2. We describe our model in Section 3, which is composed of two different modules: automatic multimodal alignment and PP-attachment detection/correction. Finally, Section 4 presents our experiments and a discussion of the results obtained.

2 Related work

The problem of finding the correct PP-attachment has attracted the attention of many researchers in the field of Natural Language Processing. It constitutes an important and challenging problem in parsing natural languages. Many different methods and sources of information have been proposed for the resolution of PP-attachment ambiguities.

Two kind of resources have mainly been used in the literature to solve the PP-attachment problem: semantic knowledge bases [1, 7], and corpora [2, 19, 23]. To our knowledge, there are not too many works using *multimodal information* to deal with this problem. The most relevant work to us is [4]; their approach consists in simultaneously perform object segmentation and PP-attachment resolution for captioned images. In order to do that, they produce a set of possible hypothesis for both tasks, and then they jointly rerank them to select the most

consistent pair. The main difference between their work and ours is that we produce a unique syntactic analysis and it is corrected according to visual informations. Moreover, we perform experiments with a much bigger number of images/caption pairs (22,800 vs. 1,822).

The disambiguation of PP-attachments by using visual information is also related to *visual relation learning*. The most related work to us is [25], in which the authors developed new visual descriptors for representing object relations in an image. Their model relies on a multimodal representation of object configurations for each relation, and it is able to learn classifiers for object relations from image-level supervision only (i.e., from image-level annotations such as "person on bike", without annotating the objects involved in the relation). While we could use their spatial relation classifiers, the focus of our work is different. We deal with the problem of disambiguating PP-attachments. We use similar visual features for representing the spatial configuration of objects, but objects are detected and represented in a different way.

Our system also aligns fragments of sentences (more concretely, noun phrases) with boxes in the image in order to be able to use multimodal information, without this being our main goal. A first step in this task is to detect objects in an image [27, 14, 13]. This requires two things: i) to find the position of the object in the image, often by calculating the coordinates of the rectangle that surrounds the object; ii) to predict a semantic class to the object. Many works have tried to learn correspondences between a part of a sentence and a part of an image, with different kind of applications in mind, such as caption generation [33, 11, 16] and image retrieval [6, 3].

Many researchers in psycholinguistics and cognitive psychology have also studied the interaction between vision and ambiguous language in human sentence processing, such as [32, 5]. These works provide evidence of the relevance of visual information for humans to solve linguistic ambiguity. This information is also of great relevance during the first stages of children's language acquisition, since much of the sentences received by children are linked to their immediate visual environment [31, 30]. Our work is inspired by these ideas and address the problem of disambiguating PP-attachments by an artificial system that uses, among other cues, the visual information linked to a concrete ambiguous sentence.

3 Our model

The model that we propose in this paper takes as input an image/sentence pair, and provides a syntactic analysis of the sentence by performing two different tasks. First, an automatic alignment of boxes detected in the image and noun phrases in the corresponding captions. Second, detection and correction of incorrect PP-attachments. We explain them in detail in the next sections.

3.1 Automatic multimodal alignment

This task is divided into three steps: detection of boxes in the image, detection of noun phrases in a sentence, and, finally, their alignment.

Detection of boxes The task of detecting boxes in an image consists of predicting the presence or absence of an object in an image given a list of objects that the system is able to recognize. When an object is recognized, the coordinates of the box containing it are produced. We have used here the real-time object detection model based on neural networks called YOLOv2 [28], which produces, for a given image, a list of boxes associated with semantic labels.

This system is broken down as follows: it takes an image as input and then cuts it into a grid. For each cell of the grid the system predicts a fixed number of bounding boxes, a confidence score for each box, and a probability for each semantic category. The final predictions are made by multiplying the confidence scores by the probability of the semantic categories.

Detection of noun phrases Although the detection of noun phrases is a widely studied task, the target phrases in our work correspond to visual objects and may differ in nature from the typical noun phrases resulting from syntactic analysis.

It consists of a simple detector of the beginning and the end of noun phrases, which associates to any word in the sentence a label in the form B (*begin*), I (*inside*) and O (*outside*) depending on whether the word starts a noun phrase, is within a noun phrase without being the first word or is outside a noun phrase. The prediction is made using an average perceptron based on the words of the sentence and their parts-of-speech. An evaluation by using our test corpus indicates an error rate of 2.2% per word.

Alignment The alignment problem consists of determining for each noun phrase which is its corresponding object detected in the image. For example, given the caption *Someone is holding out a punctured ball in front of a brown dog with a red collar*, it is necessary to find among the boxes corresponding to the objects detected in the corresponding image (e.g., the balloon, the arm, the dog, the collar) which noun phrases they correspond to (*someone*, *a punctured ball*, *a brown dog*, *a red collar*). It is a difficult artificial vision problem because of the very different nature of the aligned objects: on the one hand, the pixels of the image and, on the other hand, a sequence of words. The problem become even more difficult when some noun phrases may correspond to several objects in the image (e.g. *children playing soccer*), some objects are only partially represented in the image (*people standing in a train*), and the object detector may have detected objects not represented in the caption.

To tackle this problem, we divide this task into two sub-tasks: the first is to calculate an association score between each visual object and each noun phrase,

the second is to decide which of these potential associations will be retained for the future. We are not addressing the problem of multiple associations.

The association score between a visual object and a sequence of words is calculated by projecting the pixels of the image and the words of the caption towards the same representation space. Each visual object and each textual sequence is represented by vectors in this common space, which makes it possible to calculate a similarity between the vectors to obtain an association score. This projection in a common space is carried out using neural networks. The parameters of this network can be driven from known image/text pairs using a method based on *visual semantic embeddings* [10], which take advantage of a convolutional neural network to create image representations and recurrent neural network to create word sequence representations.

Once the alignment score is obtained for each image/text pair, it is necessary to determine a global association taking into account that it is neither injective nor surjective (some elements are not associated, others have multiple associations). This association is achieved using the following heuristic: pairs with the highest score are iteratively selected, each box can be assigned to no more than one noun phrase. Only pairs of scores greater than 0.3 are considered (threshold determined on a development corpus).

3.2 Detection and correction of PP-attachments

The automatic alignment between the image and its caption allows the integration of visual features for the task of correcting PP-attachments produced by a parser. This section presents the correction module, which is divided into two steps. A first step that detects the attachment errors produced by the parser using a classifier. Then the correction strategy in which candidate governors at the target preposition are selected and then evaluated again with a classifier. The governor with the highest score is selected.

Detection of errors in the attachment The detection of attachment errors is carried out using the AdaBoost algorithm [12]. To train this classifier we used two types of features: lexical and visual. These features concern to the p preposition, its governor G and its object O. When the governor is a verb, the subject of the verb serves as G. So, in the sentence "Jean eats with gloves on.", we get $G = Jean$, $p = with$ and $O = gloves$.

For the *lexical features*, starting from the dependent tree produced by a parser, we use: (a) the lemma and the grammatical category of the governor and the object; (b) the distance between the preposition and its governor. A detailed description of this feature is presented in previous works [8].

Visual features are calculated from the bounding boxes that the alignment system has associated to the governor and the object of the preposition. We distinguish two types of visual features:

– Conceptual features: person, body part, animal, clothing, instrument, vehicle and other. They are predicted when objects are detected in the

image. If the alignment module has not selected any boxes for one of the two selected noun phrases (governor or object), the UNK value is used to represent the concept of this noun phrase and none of the spatial features are calculated.

– Spatial features: given the governor's box $b_G = [x_g, y_g, w_g, w_g, h_g]$ and the object box $b_O = [x_d, y_d, w_d, h_d, h_d]$ of the preposition, where (x, y) are the coordinates of the box center, and (w, h) are the box height and width, we use the features proposed by [25]:

$$V_{S1} = \frac{x_d - x_g}{\sqrt{w_g h_g}} \; V_{S3} = \sqrt{\frac{w_d h_d}{w_g h_g}} \; V_{S2} = \frac{y_d - y_g}{\sqrt{w_g h_g}} \; V_{S4} = \frac{b_g \cap b_d}{b_g \cup b_d}$$

Features V_{S1} and V_{S2} represent the horizontal and vertical relative positions between the two boxes, respectively. V_{S3} is the ratio of box sizes, V_{S4} the overlap between boxes, and $V_{S5} = \frac{w_g}{h_g}$, $V_{S6} = \frac{w_d}{h_d}$ the aspect ratio of each box, respectively.

Based on all these features, the classifier checks whether the alignment proposed by the parser is correct or not.

Correction strategy In order to increase the accuracy of the parser, we use a correction strategy that consists in changing the attachment proposed by the parser using an error corrector [8]. When a connection is detected as incorrect by the classifier, we apply rules to the syntax tree generated by the analyzer to obtain a set of alternative connections. These possible new attachments are given to the classifier to make a final decision by selecting the one with the highest probability of attachment.

4 Experiments

4.1 Dataset

There is a lack of datasets that provide not only paired sentences and images, but detailed information about the correspondence between regions in images and phrases in captions. In this paper we focus in a multimodal corpus called *Flickr30K Entities* (F30kE) that provides such type of annotations [26]. It constitutes an extension of the original Flickr30k dataset [34], which is a well-known benchmark for sentence-based image description.

F30kE is composed of almost 32K images with five captions per image. Their annotation consists of identifying which mentions among the captions of the same image refer to the same set of entities (a total of 244k co-reference chains were annotated), and associating them with bounding boxes localizing those entities (a total of 276K bounding boxes were manually annotated). Each mention in the captions is categorized, using manually constructed dictionaries, into the

following eight types: people, body parts, animals, clothing, instruments, vehicles, scene, and other.

In order to use this corpus for our task, we enriched it with a manual attachment of 29,068 prepositions to their governor. The attachment correction was made by a single annotator, who had only access to the caption, the target preposition and the corresponding image. More details can be found in [9]. For our experiments, we subdivided this corpus into three sets: learning (23,254 prepositions), development (2,907 prepositions) and test (2,907 prepositions).

4.2 Setup

Alignment module

Image processing: We re-trained the YOLOv2 model on the F30kE corpus using as initialization the weights provided by the authors and limiting the number of categories to the eight semantic categories of the F30kE corpus. Only predictions with a confidence score above 0.1 were retained. The system detected 7,110 of the 14,229 boxes of images from our test corpus. An object is considered detected if the ratio of the area of the intersection over union between its ground truth box and the predicted box is greater than 0.5.

The detector achieves a recall of 0.49 and an accuracy of 0.29 on the test set. If we take into account the semantic categories, these performances fall down to 0.25 and 0.15 respectively. These results show us that this is a difficult task and that automatic image processing in this detection task represents a first barrier to the use of visual information.

Afterwards, the content of each box is resized to 224 by 224 pixels, then passed to the input of a ResNet-152 [15] network pre-trained for the image classification task in thousand scene categories from the ImageNet Large Scale Visual Recognition Challenge [29]. The last layer of the network is replaced by a dense layer (i.e., a linear transformation) that projects the representations to a vector size 1,024.

Text processing: The words of the noun phrases are first projected into a 300 size representation space that provides inputs to a GRU (*gated recurrent unit*) recurrent layer whose hidden representation is of size 1,024. The hidden representation of the recurrent network, after reading the words of a noun phrase, is used as a representation for the textual modality.

Alignment: A ℓ_2-normalization is applied to neural networks activations of both modalities in order to be compared using the scalar product; this is equivalent to calculating the cosine similarity between the two vectors. Learning is performed on batches (*batches*) of size 48 for 30 periods using the Adam [17] optimization method. This is a model of *triplet ranking* whose learning is performed by calculating the similarity between an image/text pair existing in the learning data and a random pair with one of the two members in common and modifying the model so that the score of the valid pair is higher than that of the invalid pair.

Table 1 shows the performance of the alignment system between boxes and noun phrases. The error rate is calculated as follows: for each noun phrase, the association is considered correct if the box with the highest similarity to this noun phrase is the one corresponding to it in the ground truth data. The results are calculated according to two methods, VSE and VSE+++, which differ by the cost function used for learning [10]. According to the model used, those provided with the VSEpp tool were trained on the Flickr30k and Microsoft Common Objects in Context [21] corpus, or the model was re-trained on the data of our task (F30kE). The models available with the VSEpp implementation were trained on complete images and complete description sentences. Their performance falls on boxes containing only one object in our corpus, doubling the number of association errors, compared to the same model re-trained on the target data (from 21 % to 37 %) which demonstrates the importance of re-training the model under the same conditions as the test.

Table 1. Alignment error rate on our test set by comparing ground truth boxes and ground truth noun phrases, depending on the model (VSE, VSE+++) and training corpus (Flickr30k and Microsoft COCO are the models provided with the tool, trained on complete sentence/image pairs rather than noun phrases and box contents)

Approaches	Trainning corpus	Error rate
VSE++	Flickr30k	42.07%
VSE++	MS-COCO	38.90%
VSE	MS-COCO	37.17%
VSE	Fine-tuning	21.47%

PP-attachment detector module In order to identify PP-attachments in captions we used a standard transition parser [24] trained on the Penn Treebank corpus [22]. The PP-attachment error detector is trained on our learning corpus. The classifier parameters were adjusted according to its performance on the development corpus.

We also used the development corpus to evaluate the performance of the rules we used to find potential governors G_p. At the output of the parser the rules allow us to retrieve the correct manually annotated governor for the preposition in the set G_p in 92.28% of cases. This score therefore represents an upper bound for PP-attachments.

4.3 Results

The experiments presented here assume a scenario in which lexical information is not available (the semantics of words are not known by the system), in order to highlight the information that can be used in the visual part of the task. The case in which lexical features are used is then seen as an upper bound. In

this context, we are interested in two questions: what is the impact of semantic categories in the visual modality, and what is the impact of spatial information in the same modality?

Table 2 shows the good attachment rate for the 10 most common prepositions of the test corpus. The F30kE corpus being mainly composed of captions that can be understood by a human without seeing the associated image (semantically unambiguous), real ambiguous cases are therefore rare.

Table 2. Correct attachment rate to the test: the number of occurrences is given for each preposition ; the *baseline* is produced by the parser without correction; V_C is obtained after correction by using only visual concepts, V_S only spatial features, V the combination of conceptual and spatial features, L lexical features and $V + L$ is the combination of all features.

Prepositions	Occurrences	Baseline	V_C	V_S	V	L	$V + L$
in	369	0.76	0.76	0.77	0.76	0.85	0.84
with	310	0.65	0.68	0.68	0.70	0.78	0.78
for	168	0.73	0.73	0.73	0.72	0.82	0.83
near	159	0.33	0.53	0.50	0.59	0.84	0.84
through	145	0.95	0.95	0.95	0.95	0.96	0.96
on	143	0.85	0.85	0.85	0.85	0.89	0.87
from	140	0.76	0.76	0.76	0.76	0.86	0.85
next to	137	0.89	0.89	0.89	0.89	0.90	0.89
into	116	0.89	0.89	0.89	0.89	0.92	0.95
over	111	0.66	0.64	0.66	0.68	0.85	0.84

As we can see in Table 2, the overall accuracy of the parser (without correction) is 75%. Note that the results vary a lot depending on the prepositions, ranging from 95% for the preposition *through*, to 33% for the preposition *near*.

Visual concepts (V_C) and spatial information (V_S) provide different improvements depending on the prepositions, but there seems to be mainly a gain with locative prepositions such as *near* (17% points). Since our visual features are focused on spatial information, it is logical that they have an impact on this kind of prepositions. One might think that conceptual categories are sufficient to solve the problem, but it should be noted that the task of visual categorization is difficult (and therefore the classifier is often wrong), and that categories are rather rough and may not remove all ambiguities. The combination of V_C and V_S gives the best results and corrects about 3% of the errors.

Lexical information (L) has a drastic effect since performance increases for most of the prepositions, underlining our hypothesis that the text is unambiguous in the absence of an image. These results are not surprising because it is well known that some ambiguities can be resolved by using only syntactic information. The problem is that this type of information is not always available. The fact that the gain is higher for lexical information than for visual information can also be explained by the unreliability of predictions in the visual modality.

There is no difference on average between the use of lexical information only and the use of all features ($V + L$). However, this result highlights the fact that the fusion of features from text and image modalities takes advantage of the strongest modality without suffering from a modality with lower performance.

We present here some examples of image/text pairs for which the image has allowed, or not, to perform a correct attachment using sets of different features. In all examples the target preposition is in bold, the governor chosen by the analyzer is underlined and the new governor after correction is in square brackets.

Figure 1.a shows a sentence for which the parser has made a bad attachment but the classifier has allowed to correct it by using only visual information. Concretely, the preposition *near* is incorrectly attached to *area* by the parser, and the classifier corrects the attachment by selecting *are* as a governor. This example is the justification for this study: to correct poor connections thanks to visual information.

Figure 1.b shows a sentence for which the use of only visual features did not help to correct a wrong attachment. Concretely, the preposition *on* is incorrectly attached to the word *building*. The alignment system did not find the bounding box for at least one of the two noun phrases. This example shows one of the limitations of this study: the difficulty of the detection and alignment phase between boxes and noun phrases, limits the impact of visual features in correcting erroneous analyses.

Even if lexical features are the most effective, if the learning corpus does not contain enough examples for some entities, visual features may be more effective. Thus Figure 1.c shows a sentence for which the use of only visual features allows to obtain the correct alignment, while the use of only linguistic features produces an error. Concretely, the preposition *with* is incorrectly attached to the word *jeans* instead of the word *wearing*.

a – Two children [are] in a grassy area **near** two horses.

b – Two people sitting in front of an older building **on** a bench.

c – A dog is wearing [jeans] and a blue and yellow shirt **with** a black vehicle in the background.

Fig. 1. Examples of image/text pairs for which have been found, or not, a correct PP-attachment, by using different kind of features.

5 Conclusion

This work explores the possibility of using images to disambiguate prepositional phrases attachments in sentences that describe them. Visual features improve the performance by three points on average depending on the prepositions, and sometimes drastically, as in the case of the preposition *near*. However, the difficulty of the problem lies in the detection and categorization of objects, as well as in the alignment between text and images. Indeed, the errors and lack of information resulting from the automation of this step inevitably reduces the overall performance of the attachment corrector.

However, the gain obtained between the output of the parser and the output of the corrector, even when using only visual features, proves two things. First, that information was found at the image level and that an alignment, even partial, was produced. And secondly, that this information could be properly used despite the use of relatively simple descriptors.

A better use of the information from the image is a main direction to explore in order to improve the system with, in particular, the integration of information directly from the pixels, such as the use of the space representation for the image or directly at the level of the bounding boxes.

References

1. Agirre, E., Baldwin, T., Martinez, D.: Improving parsing and PP attachment performance with sense information. In: ACL HLT. pp. 317–325 (2008)
2. Belinkov, Y., Lei, T., Barzilay, R., Globerson, A.: Exploring compositional architectures and word vector representations for prepositional phrase attachment. TACL **2**, 561–572 (2014)
3. Chang, A.X., Monroe, W., Savva, M., Potts, C., Manning, C.D.: Text to 3d scene generation with rich lexical grounding. In: ACL-IJCNLP:2015. pp. 53–62 (2015)
4. Christie, G., Laddha, A., Agrawal, A., et al.: Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In: EMNLP. pp. 1493–1503 (2016)
5. Coco, M.I., Keller, F.: The interaction of visual and linguistic saliency during syntactic ambiguity resolution. QJEP **68**(1), 46 – 74 (2015)
6. Coyne, R., Sproat, R.: Wordseye: an automatic text-to-scene conversion system. In: SIGGRAPH. pp. 487–496 (2001)
7. Dasigi, P., Ammar, W., Dyer, C., Hovy, E.: Ontology-aware token embeddings for prepositional phrase attachment. In: ACL. vol. 1, pp. 2089–2098 (2017)
8. Delecraz, S., Nasr, A., Bechet, F., Favre, B.: Correcting prepositional phrase attachments using multimodal corpora. In: IWPT. pp. 72–77 (2017)
9. Delecraz, S., Nasr, A., Béchet, F., Favre, B.: Adding syntactic annotations to flickr30k entities corpus for multimodal ambiguous prepositional-phrase attachment resolution. In: LREC (2018)
10. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 (2017)
11. Fang, H., Gupta, S., Iandola, F.N., et al.: From captions to visual concepts and back. In: CVPR. pp. 1473–1482 (2015)

12. Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. JSAI **14**(771-780), 1612 (1999)
13. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137 (2015)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. de Kok, D., Hinrichs, E.W.: Transition-based dependency parsing with topological fields. In: ACL. vol. 2: short paper (2016)
19. de Kok, D., Ma, J., Dima, C., Hinrichs, E.: Pp attachment: Where do we stand? In: EACL. vol. 2, pp. 311–317 (2017)
20. Kummerfeld, J.K., Hall, D.L.W., Curran, J.R., Klein, D.: Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In: EMNLP-CoNLL. pp. 1048–1059 (2012)
21. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
22. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational linguistics **19**(2), 313–330 (1993)
23. Mirroshandel, S.A., Nasr, A.: Integrating selectional constraints and subcategorization frames in a dependency parser. Computational Linguistics (2016)
24. Nasr, A., Béchet, F., Rey, J.F., Favre, B., Le Roux, J.: Macaon: An nlp tool suite for processing word lattices. In: ACL HLT. pp. 86–91 (2011)
25. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: ICCV. pp. 5189–5198 (2017)
26. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. ICCV **123**(1), 74–93 (2017)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
28. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: CVPR. pp. 6517–6525. IEEE (2017)
29. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
30. Shaerlaekens, A.: The Two-Word Sentence in Child Language Development: A Study Based on Evidence Provided by Dutch-Speaking Triplets. Mouton, The Hague (1973)
31. Snow, C.E.: Mothers' speech to children learning language. Child Development **43**(2), 549 – 565 (1972)
32. Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M., Sedivy, J.C.: Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. Cognitive Psychology **45**(4), 447 – 481 (2002)
33. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164 (2015)
34. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)