Natural Language Semantics With Pictures: Some Language & Vision Datasets and Potential Uses for Computational Semantics

David Schlangen
Department of Linguistics, University of Potsdam, Germany*
david.schlangen@uni-potsdam.de

Abstract

Propelling, and propelled by, the "deep learning revolution", recent years have seen the introduction of ever larger corpora of images annotated with natural language expressions. We survey some of these corpora, taking a perspective that reverses the usual directionality, as it were, by viewing the *images* as semantic annotation of the natural language expressions. We discuss datasets that can be derived from the corpora, and tasks of potential interest for computational semanticists that can be defined on those. In this, we make use of relations provided by the corpora (namely, the link between expression and image, and that between two expressions linked to the same image) and relations that we can add (similarity relations between expressions, or between images). Specifically, we show that in this way we can create data that can be used to learn and evaluate lexical and compositional grounded semantics, and we show that the "linked to same image" relation tracks a semantic implicature relation that is recognisable to annotators even in the absence of the linking image as evidence. Finally, as an example of possible benefits of this approach, we show that an exemplar-model-based approach to implicature beats a (simple) distributional space-based one on some derived datasets, while lending itself to *explainability*.

1 Introduction

In model-theoretic formal semantics, the central semantic notion "truth" is explicated as a relation between a sentence and a mathematical structure, its *model*. Semantics textbooks are surprisingly evasive about what exactly this structure is meant to be, other than hinting at that it in some way represents the general "situation", or "world", that the sentence is taken to be talking about. In any case, what the model as a mathematical structure does is to provide a collection of *individuals* about which the sentence could be talking, and an *interpretation* of the non-logical lexical items occurring in the sentence, in terms of sets of individuals (or tuples of individuals). The collection of individuals is typically called the *domain* D, and the set of interpretations I, so that a model $M = \langle D, I \rangle$.

It is this intended relation with the world that allows us to see an analogy between these structures and photographic images. A photograph is a frozen moment in time, a representation of how the world was (or looked like) at a certain moment, at a certain place and from a certain perspective. And just as a sentence in formal semantics is evaluated relative to a model, a sentence describing a situation can be seen as true *relative to an image* — if (and only if) the image *depicts* a situation of the described type. Hence, in a slight reversal of our usual way of talking, we can say that a given *image* does (or does not) make a given sentence true (instead of saying that the sentence is a true description of the image), and we can see the image as a model of the sentence.[1]

What does this sleight of hand buy us? A very large amount of data to play with! The field of computer vision has as one of its central aims to find meaning in pixels – see e.g., Davies (2012), Marr

*Work done while author was at Bielefeld University.

[1]There are interesting subtleties here. In our everyday language, we are quite good at ignoring the image layer, and say things like "the woman is using a computer", instead of "the image shows a woman using a computer", or "this is a computer", instead of "this is an image of a computer". This also seems to carry over to tense, where we can say "is using", instead of "was using at the time when the picture was taken". There are however contexts in which talk about the image *as* image is relevant, and this can happen in large corpora such as discussed here. So this is something to keep in mind.

(1982) – and a convenient way of representing meaning is with natural language. It is also a field that has been data-driven for a long time, and so there is a large number of data sets available that in some way pair images with natural language expressions.[2] Recent years have specifically seen the creation of large scale corpora where images are paired with ever more detailed language (e.g., single sentence or even full paragraph captions describing the image content; facts about the image spread over question and answer; detailed descriptions of parts of the image in terms of agents and patients; see references below).[3] Given the understanding that all these expressions are meant to "fit" to the images that they are paired with, and using the slight conceptual inversion of treating the images as "truthmakers" (Fine, 2017) for the sentences, this gives us an unprecedentedly large set of language expressions that are "semantically annotated."[4,5] As we will show, this gives us material to learn about the lexical and compositional semantics that underlies the use of the expressions.

The **contributions of this paper** are as follows: 1.) To make explicit a perspective that so far has been taken only implictly in the literature, which is to view images as *models* of natural language expressions; 2.) to show by example that taking this perspective opens up interesting data sets for computational semantics questions; 3.) specifically, to look at how grounded interpretation functions could be learned from and tested on this data, and; 4.) how data can be derived that expresses various implicature relationships; and 5.) to show how exemplar-based model building can be used to predict some of those relations. Our code for working with the corpora mentioned here (and some others) is available at `http://purl.com/cl-potsdam/sempix`.

2 Background

2.1 The Approach: Learning Semantics From Relations in Corpora

Our general approach will be to look at relations that are expressed in the data or can be added using computational methods, and then to ask what these can tell us about *semantic* relations like truth and entailment, and in turn what these tell us about the meanings of expressions. Figure 1 illustrates the idea. The corpora provide us with an "annotates" relation between images and expressions; in the Figure, holding between I_1 and e_1 and e_2, and I_2 and e_3, where the expressions for example could be captions. Implicitly, there is also an "annotates same image" relation that holds between expressions; here, e_1 and e_2, as alternative captions of the same image. Standard natural language processing and computer vision techniques (see below) allow us to compute similarity relations between pairs of images (e.g., I_1 and I_2) and between pairs of expressions (e_1, e_3). The question then is whether these relations can tell us something about *satisfaction / denotation* ($\models, [\![\cdot]\!]$) and *entailment* ($\models, \vdash$).

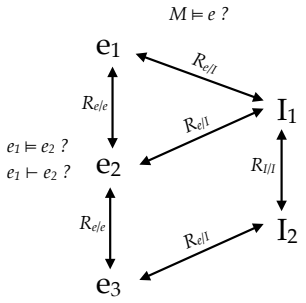

Figure 1: Relations in corpora & to be derived

2.2 Corpora Used Here

We make use of data from the following corpora:

• **MSCOCO / RefCoco / GoogleREX:** The "Microsoft Common Objects in Context (COCO)" collection (Lin et al., 2014) contains over 300k images with object segmentations (of objects from 80 prespecified categories), object labels, and nearly 400,000 image captions. It was augmented with 280,000

[2] See for example the (incomplete) lists at `http://www.cvpapers.com/datasets.html` and `https://riemenschneider.hayko.at/vision/dataset/`.

[3] While there are by now some non-English or even multi-lingual corpora, the majority provide *English* language annotations, including all of those that we discuss here.

[4] The corpora we discuss here provide almost 8 million distinct natural language expressions (with many more that can be derived from them). In comparison, the largest "classical" semantics resource, the Groningen Meaning Bank (Bos et al., 2017), provides some 10,000 annotated sentences, and the Parallel Meaning Bank (Abzianidze et al., 2017) another 15,000. There is no competition here, though: the Meaning Bank annotations are obviously much deeper and much more detailed; the proposal in this paper is to view the image corpora discussed here as complementary.

[5] The relation between images and models is implicit in (Young et al., 2014), from where we took inspiration, but not further developed there in the way that we are attempting here. Hürlimann and Bos (2016) make an explicit connection between image and models, but only look at denotations; as do Schlangen et al. (2016).

referring expressions by Yu et al. (2016), using the ReferitGame where one player needs to get another to identify a predetermined object in the image, with the players getting feedback on their success. Mao et al. (2016) also provide expressions for COCO objects, but collected monologically with the instructions to provide an expression that uniquely describes the target object.

• **Flickr30k / Flickr30kEntities:** Flickr30k (Young et al., 2014) is a collection of 30,000 images from a public image website which were augmented with 160,000 captions; Plummer et al. (2015) annotated these captions with positions of the objects in the images that they mention (Flickr30kEntities).

• **Visual Genome:** This dataset by Krishna et al. (2016) combines images from COCO and another data set (yielding around 100k images), and augments them with 2 million "region descriptions", which are statements true about a part of an image, and resolved for the entities mentioned and their relations. These descriptions are parsed into object names and attributes, and normalised by reference to the WordNet ontology (Fellbaum, 1998). Krause et al. (2017) added 20,000 image description paragraphs (i.e., extended, multi-sentence captions) for some of the images.

All these data sets give us images paired with natural language expressions; in most of them, the relation between image and expression is annotated more fine-grainedly by linking regions within the image to (parts of) the expressions.[6] Also, some corpora provide an additional layer that could be seen as corresponding to the logical form of the expression, for example by normalising nouns to a resource like WordNet (Fellbaum, 1998) or by annotating the predicate / argument structure.

3 Expressions and Denotations

3.1 Images as Semantic Models: An Example

Above, we have introduced our analogy between semantic models and images. An example shall make it clearer. Figure 2 shows an image (from COCO) with *object segmentations* (rectangular patches indicating the position of an object in the image) and identifiers, as provided by the corpus. We can directly treat this as the *domain* provided by the model, so that here $D = \{o_{92839}, o_{93793}, o_{387589}, o_{387727}, o_{505664}, o_{510191}, o_{660005}, o_{1168354}, o_{1587273}, o_{1716887}, o_{1863940}, o_{1864058}, o_{1864291}\}$.

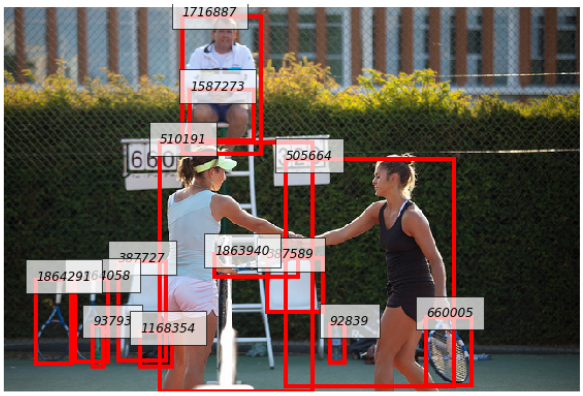

Figure 2: A Segmented Image from COCO

The corpus also provides natural language annotations for these objects, for example "the woman in white" and "the woman in black" (for o_{505664}, o_{510191}, respectively). We can use this to "reverse engineer" the interpretation functions covering these words, and in particular derive that $I(woman) \subseteq \{o_{505664}, o_{510191}\}$. If we make an additional *exhaustivity* assumption over the set of annotations, we can strengthen this to $I(woman) = \{o_{505664}, o_{510191}\}$; that is, make the assumption that these are the *only* objects (in this image / the set of segmented objects from that image) to which this term can be applied. We will need to make this assumption when we want to generate *negative instances* used in machine learning, but need to keep in mind that in general, this assumption is unwarranted, as exhaustivity was not a goal when creating the corpora.

$o_{505664} =$

Figure 3: Invidual

Continuing with the discussion, we can think more about what this view on the corpora offers for doing semantics. Our domain D is now populated not just with identifiers or symbols from a vocabulary, but rather with objects that have an internal structure. In the example above we were able simply to read off the interpretation function for the word from the annotation. But we can try to use

[6] This makes working with the images easier, as it allows us to assume that the task of *object recognition* (detecting contiguous regions of pixels that belong to the same object) has already been successfully performed. This is not a strict requirement for working with images these days, however, as high-performing models are available that do this job (Redmon and Farhadi, 2018), (He et al., 2017), but these still add noise from which one might want to abstract for the purposes discussed here.

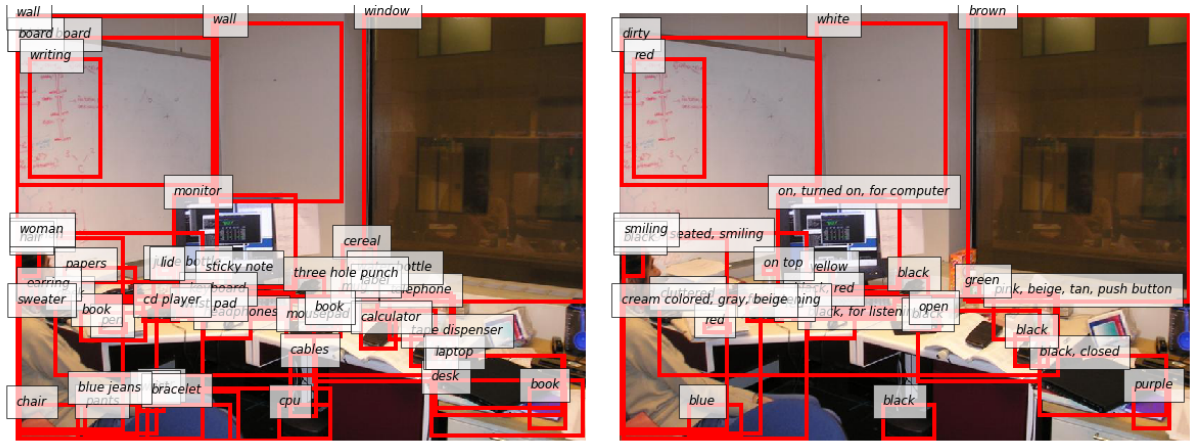

Figure 4: Object "names" (left) and "attributes" (right) from Visual Genome, for an example image

instances like this to *generalise* this function. That is, we can try to turn I into a "constructive" function that not just records a fact ("object o is in the denotation of predicate ϕ"), but rather produces a *judgement*, given a (structured, visually represented) object; we may write $I_{woman}(D)$ to make this explicit.[7]

With this in hand, we can now explore how the available data could help us learn and evaluate lexical interpretation functions and their composition. We will look at the available expression types in order of increasing syntactic complexity. The relation that we will make use of first is the "annotates" relation between expressions e and (parts of) images I. This relation gives us the value of $[\![e]\!]^{\langle D,I\rangle}$ (which is an entity or a truth value); the interest is in learning about $[\![\cdot]\!]^{\langle \cdot,I\rangle}$ as it covers the constituents of e and their composition.

3.2 Expression Types Found in Corpora

3.2.1 Sub-Sentential Expressions

Of the corpora discussed here, only Visual Genome provides open-class **single word** annotations. Objects in the corpus are associated with "names" (typically **nouns**), and "attributes" (typically **adjectives**), which were semi-automatically segmented out of larger expressions provided by annotators (to be discussed below). Figure 4 shows this for one image from the set. (It illustrates at the same time how fine-grainedly the corpus is segmented—on average it provides 36 object bounding boxes per image.)

The Visual Genome annotation provides over 105,000 word form types, of which about 10,500 have at least 10 instances. Using the normalisation to WordNet synsets in the corpus, this reduces to roughly 8,000 types, of which 3,500 occur at least 10 times. The distribution (not shown here) is roughly Zipfian—and reveals a certain bias in the data, with "man" occuring twice as often as "woman", for example. This is a sizeable vocabulary for which interpretation functions can be learned from this data.

We have briefly mentioned the problem of getting *negative instances* of word denotations, as required by typical machine learning methods. One method is to sample from the set of objects in a given image that are *not* annotated with a word; but this requires making the aforementioned (non-warranted) exhaustivity assumption. Schlangen et al. (2016) have shown this to be unproblematic for the data that they used; establishing to what degree it would be here we leave to future work. It is likely to be more of a problem for adjectives, where the choice of what to mention is governed much more by the context than the choice of which name to use for an object.

This data can be assembled into simple nominal phrases (ADJ + N; e.g. "brown window" for top right of Figure 4). Semantically, these would be **indefinite noun phrases**, as all that is guaranteed is that they are appropriate for the object that they apply to (but there may be others of that type in a given image). With the denotation being known, this can be used to evaluate the semantic composition.

[7] This perspective has previously been taken by Schlangen et al. (2016) and developed for simple expressions; the present section builds on that work.

More interesting and complex are the noun phrases found in the **referring expression** corpora (the ReferIt variants; see above). These expressions were produced in the form that they are recorded in the corpora (unlike the single word expressions discussed above), and in an actual context of use, namely with the aim to single out an object to a present interlocutor. This makes this data set also interesting from a pragmatic point of view, as one can ask how the context (in the image, but also in the production situation) may have influenced the linguistic choices. The following shows the referring expressions available for the tennis player on the right in the image from Figure 2 above; also shown is the annotation from the GoogleREX corpus:

(1) a. RefCoco: lady in black on right | girl in black | woman in black
b. RefCoco+: black shirt | girl in black | player in black
c. GoogleREX: woman in black tank top and shorts holding tennis racket | woman in black outfit shaking other tennis player hand

Contrasting the GoogleREX expressions illustrates the influence of the context of use on the shape of the expressions. The GoogleREX annotators did not have interlocutors and were just tasked with producing expressions that describe the object uniquely. The ReferIt expressions do this as well, but additionally, they do this in the most efficient and effective way, as the players had an incentive to be as fast as possible, while ensuring referential success. This shows: The average length of RefCOCO expressions is 3.5 token, that of GoogleREX 8.3.[8]

We also find **relational expressions** like the following in these corpora, which identify the *target* object by relating it to another one (the *landmark*):

(2) woman under suitcase | laptop above cellphone right | black van in front of cab

To learn the interpretation of such relational items (here, "under", "above", "in front of"), it would be good to have grounding information also about the landmark. The corpora mentioned so far do not give us this,[9] and so we turn to Visual Genome, and away from referring expressions.

Visual Genome was collected with the explicit purpose of providing material for learning "interactions and relationships between objects in an image" (Krishna et al., 2016). The starting point of the annotation was the marking of a region of interest in the image, and the annotation of that region with a "region description", ie. an expression that is true of that region. Note the difference to referring expressions: no stipulation is made about whether it is or is not true of *other* objects in the image. Annotators were encouraged to provide region descriptions that are relational, and these then form the basis of an abstracted representation of that relation. Figure 5 shows an example of such a region description; the corresponding annotation is shown in (3), slightly re-arranged to make clearer its similarity to classical logical forms (LFs).[10]

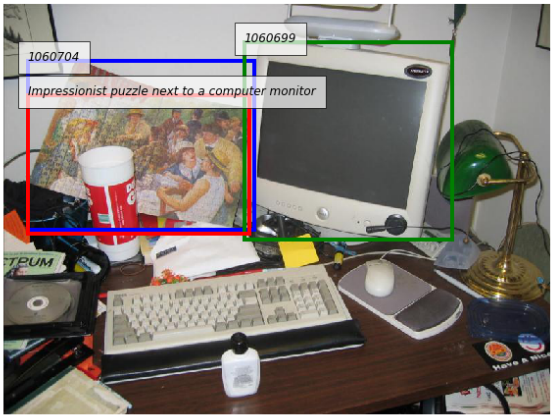

Figure 5: A region description from Visual Genome

(3) `"next to a":be.v.01(1060704:puzzle.n.01, 1060699:computer_monitor.n.01)`

[8]These corpora have been used by Kazemzadeh et al. (2014), Yu et al. (2016), Mao et al. (2016), Schlangen et al. (2016), Cirik et al. (2018) to train and test models of referring expression resolution.

[9]For a portion of GoogleREX, this was added by Cirik et al. (2018).

[10]What this also illustrates is that the normalisation decisions made in the corpus can occasionally be somewhat questionable. Here, the part "next to a" is normalised to the verb "be"; presumably, the annotator added the elided copula here and rather ignored the spatial relation.

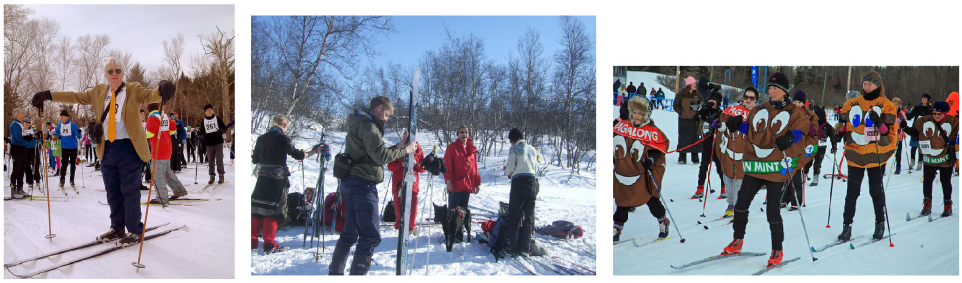

Figure 7: "A man standing in the snow with skis on." (left), and distractors (visual similarity, middle; semantic similarity, right)

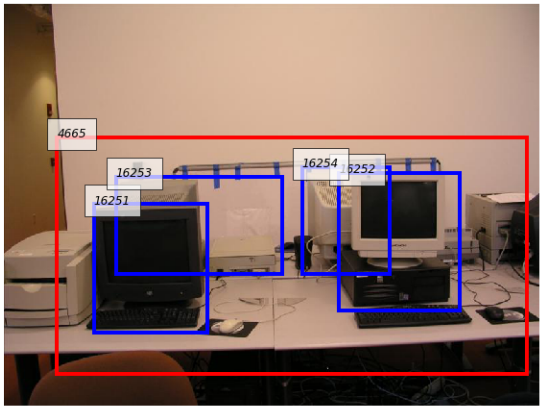

Figure 6: "there are desktop computers on the desk"

There are over 5 million region descriptions in Visual Genome, of which almost 2 million are parsed into this logical form. There are around 37,000 different relational terms in this set, of which around 3,100 occur more than 10 times. From this, a sizable number of relational interpretation functions could be learned.

Before we move on, we note that in about 6.8% of the region descriptions there is more than one object associated with an expression; as in the example in Figure 6, where "desktop computers" is resolved to four different bounding boxes. Such configurations could be used to learn the function of the plural morpheme. Looking at the expressions, there are also more than 1,000 instances each of quantifiers and numerals such as "several", "two", "many", which provides opportunity to learn their meaning.

3.2.2 Sentences

We now turn to expressions that need to be evaluated relative to the image as a whole. Such expressions can be *constructed* for example by plugging the nominal phrases from above into the sentence frame "there is NP_{indef}" (e.g., "there is a brown window" for Figure 4), to yield **existential assertions**. Negative examples (where the constructed sentence is false) can be selected by sampling an image that is not annotated as containing an object of that type, again making use of an exhaustivity assumption.

Constructing examples in this way gives us control over the complexity of the expression, at the cost of a loss in naturalness. Some of the corpora, however, also come with *attested* examples of expressions that are meant to describe the image as a whole; COCO for example provides over 400,000 of such **captions**. Figure 7 (left) gives an example of an image/caption pair.

How can we sample negative instances, where the image is *not* described by the caption? One method is to simply sample an arbitrary image from the corpus: there will be a good chance that it does not fit the caption. Too good a chance, perhaps, in that we are likely to hit an image that does not even contain any of the entity types mentioned in the expression. To make the task harder, we can now make use of one of the derived relations described above, namely a simiarity relation between images.

We looked at two ways of defining such a relation. *Visual* similarity ($sim^{vis}_{I/I}$) is the inverse of the cosine distance in image representation space, using a pre-trained convolutional neural network (we used VGG-19, Simonyan and Zisserman (2014), pre-trained on ImageNet Russakovsky et al. (2015)). We compute *content-based* or *semantic* similarity ($sim^{sem}_{I/I}$) by vectorising the image annotation (in a many-hot representation with the object types as dimensions), using SVD to project the resulting matrix into a lower-dimensional space. Given our analogy between semantic models and images, with this we then have a similarity relation between models, and we can select distractor images / models that are more challenging to refute. Figure 7 shows distractors selected via visual (middle) and semantic (right)

Figure 8: Image described by paragraph (see text; left), and distractors (visual similarity, middle; semantic similarity, right)

similarity. As this illustrates, quite fine-grained resolution abilities are required to recognise these as not fitting the caption. (Man, but skis not on; man with skis on, but not standing.)

Captions from COCO are not finely grounded (no links between objects in the image and parts of the expression). Flickr30kEntities provides this; for reasons of space, we do not show an example here. We also skip over the *wh*-questions (with answers) that are available for COCO and Visual Genome, noting only that they add an interesting generation challenge (if set up as open answer task; if set up as multiple-choice task, this reduces to making a decision for a proposition).

3.2.3 Discourses

Finally, example (4) shows an **image description paragraph** for a Visual Genome image. The associated image and two distractors are shown in Figure 8. The semantic challenge here when evaluating such a paragraph relative to an image, at least when a probabilistic approach is taken, is that a decision must be made on how to combine the uncertain judgements from each constituent sentence.

(4) The baseball player is swinging the bat. The ball is in the air. The dirt on the ground is light brown. The baseball player is wearing blue pants. The other baseball players are watching from The Dugout. The baseball player swinging the bat is wearing a dark-colored baseball hat. He's also wearing a bright red belt.

As this survey has shown, there is plenty of data available for learning grounded interpretation functions for individual words (nouns, adjectives, prepositions), and for evaluating (or even learning) how these functions must be put together to yield interpretations for larger expressions (NPs, sentences, and even discourses).

4 Expressions and Implications

4.1 Images as Implicit Link between Expressions

Besides the question of whether a statement is true of a given situation, an interesting question often is whether a statement *follows* from another one. There are various ways of tying down what exactly "follows" may mean. A very general one is given by Chierchia and McConnell-Ginet (1990), who use "A implies B" for cases where (the statement and acceptance of) A *provides reason* to also accept B.[11] This covers cases where a *proof* can be given that connects B to A (where the relation would be *syntactic consequence*, $\vdash$), cases where an argument can be made that any model that makes A true will also make B true (*semantic consequence*, $\models$), but also cases where A may just make B very plausible, given common sense knowledge (which we might call *common sense implicature*, and denote with $\models_{cs}$).

Here, we look at relations that we can take from the corpora and ask whether these can help us get at these semantic implicature relations. We make use of the fact that for most of the image objects in

[11] This is also how later the influential "recognising textual entailment" challenge (Dagan et al., 2006) would describe the relation, however also starting the tradition in natural language processing to overload the term "entailment" to cover all of what could more generally be called "implication". Young et al. (2014) call their task, defined via images as well and our inspiration for the work described here, with a qualifier as "*approximate* entailment".

the corpora, we have available more than one expression of the same type, e.g., more than one referring expression, or more than one caption. In the following, we take a look at some examples, sorting the discussion by the type of expressions that we pair.[12] We will argue that to predict the presence (or not) of an implicature relation, a different, complementary kind of lexical knowledge is required than for evaluation relative to an image (or situation); cf. Marconi (1997).

4.2 Types of Relations

4.2.1 Same Level / Rephrasing

Example (5) shows referring expressions from RefCOCO (left) paired with another expression referring to the same object (middle) and with one referring to a randomly sampled other object from the corpus. The prediction task is to identify the left/middle pair as standing in an implicature relation, and the left/right pair as not standing in this relation. (To put a practical spin on it, this could be seen as detecting whether the second pair part could be a *reformulation* of the first, perhaps as response to a clarification request.)

(5) a. right girl on floor || lady sitting on right | guy on right
b. woman || left person | pizza on bottom right
c. man trying to help with suitcase || man in jacket | very top zebra

Despite the brevity of the expressions, as this example indicates, this task seems to require quite detailed lexical knowledge, for example detecting incompatiblity between "guy" and "girl" in (5-a), but compatiblity between "lady" and "girl". (If this knowledge were available, perhaps a *natural logic*-type (Moss, 2015) approach could then be taken.) Creating this dataset only requires that several referring expressions are available for the same object, and indeed RefCOCO for example provides on average 7.1 per object, for a total of over 140,000 referring expressions.

We randomly sampled 60 instances of such pairings (balanced pos/neg) and presented each to three workers on Amazon Mechanical Turk, asking for a semantic relatedness judgement (on a 4 step Likert-scale). Using the majority label and binning at the middle of the scale, the accuracy is 0.68. This indicates that while noisy, this method creates a recognisable semantic relation between these expressions.[13]

Example (6) shows similar pairings of captions, with the negative instance (the final part of each sub-example) taken from a distractor image selected for semantic image similarity. As this illustrates, the task only becomes harder, with the caption that is intended to be non-matching occasionally accidentally even intuitively being compatible after all. (Crowd accuracy, henceforth AMT, with same setup as described above: 0.63.)

(6) a. A woman with a painted face riding a skateboard indoors. || A woman with face paint on standing on a skateboard. | There are men who are skateboarding down the trail.
b. Man and woman standing while others are seated looking at a monitor. || A man and woman play a video game while others watch. | Two people standing in a living room with Wii remotes in their hands.

4.2.2 More Specific / Entailment

Since some of the corpora overlap in their base image data, we can intersect the annotation and create derived data sets. (7) shows examples of a caption from COCO (left) paired with an object from Visual Genome (slotted into a "there is (a) __" frame for presentation) taken from the same image (middle), and a randomly sampled object (right) in (7-a) and (7-b), and with region descriptions (also from Visual Genome) in the other examples.

[12] Our inspiration for this approach comes from two sources. As mentioned, Young et al. (2014) used image captions to create their "approximate entailment" data sets; our proposals here can be seen as a generalisation of this to other pairings. Further, the original "natural language inference" dataset by Bowman et al. (2015) used captions as seeds, but had the entailments and contradictions manually generated and not derived via image relations, as we do here.

[13] Note that the task was to judge pairs, not to decide between two hypotheses, which would presumably be a simpler task.

(7) a. A man wearing a black cap leaning against a fence getting ready to play baseball. || there is (a) man | there is (a) cow
b. Rice, broccoli, and other food items sitting beside each other || there is (a) health foods | there is (a) granite
c. A man playing Wii in a room || there is/are (a) a plant that sits on a desk | there is/are (a) field covered in green grass
d. A woman is riding a wave on a surfboard. || there is/are (a) Woman with the surfboard. | there is/are (a) Students sitting at their desks

Judging from these examples, quite detailed knowledge about situations and possible participants seems to be required to predict these relations. (AMT accuracy caption/object: 0.58, caption/region: 0.6.)

4.2.3 More Detailed / Elaboration

Finally, (8) shows examples of a caption (from COCO) paired with a paragraph (from Visual Genome-paragraphs) describing the same (middle) or another, but similar image. The task here is to detect whether the extended description fits with the short description or not, which again seems to require quite detailed knowledge about situations and likely sub-events. (AMT: 0.6.)

(8) two people lying in a bunk bed in a bedroom.
A boy and girl are sitting on bunk beds in a room. The boy is wearing a red shirt and dark pants. The girl is wearing a gray shirt and blue jean pants. There is a green and pink blanket behind the boy on the top bunk. The girl is sitting on a rolled up blanket. She is wearing red glasses on her eyes.
A woman is sitting on a bed beside a little girl. She is wearing a sweater and black bottoms. The woman has eyeglasses on her eyes. The girl is wearing a colorful jacket. The girl is looking at a book that is opened on her lap. The bed is sitting against a white painted wall. There is a red blanket on the bed.

Using this general recipe, further datasets can be created with other combinations, for example pairing sets of region descriptions with further descriptions either from the same or from a different scene, or for the task of predicting the number of distinct entities introduced by a sequence of region descriptions. For reasons of space, we do not show examples here.

5 A Case Study: Model-Building for Predicting Entailment

Entailment tasks, triggered by the aforementioned "natural language inference" dataset (Bowman et al., 2015) have in recent years become a staple NLP task. They are typically tackled with very high-capacity machine learning models that classify distributed representations of the candidate relata, e.g., as in (Devlin et al., 2018). With the perspective developed here, we can liken such approaches to the *syntactic* way of defining entailment ($\vdash$), in that these approaches only take the surface form into account (and implicitly learn and use the required common sense knowledge).

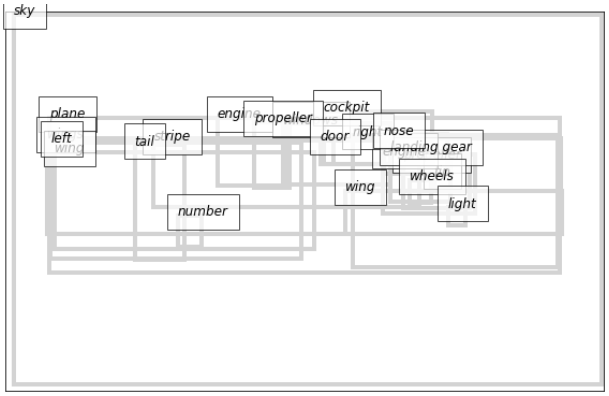

Figure 9: A Retrieved Abstract Exemplar Situation

A *semantic* approach seems possible as well, however. In its brute force form, it would implement the typical way in which *semantic consequence* is defined, by quantification over all models. Here, this would mean testing, along the lines developed in Section 3, whether all images (in a sub-corpus held for that purpose) that make the premise true also make the hypothesis true. We try something else here, which is more like model-building (Bos, 2003), for data of the type illustrated in (7) above.

The idea is as follows. Given the premise (in our case, always a caption), we retrieve a set of images

(other than that from which the caption was taken), via captions that are nearest neighbours in a text embedding space (for which we used the "universal sentence encoder" by Cer et al. (2018)). That is, we make use of a derived expression/expression relation, to create a relation between an expression and a set of retrieved models. (One can think of these as situation exemplars stored in memory and retrieved via their short descriptions.) Figure 9 shows such a retrieved model (abstracted away from the actual image content, which is not used), for the trigger caption "An airplane flying through the sky on a cloudy day." and retrieved via its most similar caption "White and blue airplane flying in a grey sky.".

We then test the candidate expressions (or rather, their "logical forms", as given by Visual Genome) against this set of models. For objects (as in (7-a) and (7-b) above), this checks whether an object of the appropriate type is in the retrieved models; for relations, this additionally checks whether the relation is also present. If the required types are present in all models, this would yield a score of 1. We set a threshold (in our experiments, at 0.2), above which a positive decision is made. As baseline, we use token overlap between premise and hypothesis for objects and intersection over union for the longer region descriptions, and distance in the embedding space. We created 10,000 triples each for the caption/object and the caption/region task.

The results in Table 1 indicate that this rather simple model captures cases that the baselines do not. An example where this is the case is shown in (9); here the retrieved models seem to have provided the entities ("umpire" and "jacket") which are likely to be present in a baseball scene, but aren't literally mentioned in the premise.

Task	Model	Strg.Bsln	Embd.Bsln	Task	Model	Strg.Bsln	Embd.Bsln
Captions / Objects	0.67	0.58	0.64	Captions / Regions	0.65	0.54	0.50

Table 1: Results for Predicting Entailment via model retrieval (and baselines)

(9) Baseball batter hitting ball while other players prepare to try and catch it. || jacket worn by umpire | silverware on a napkin

This is clearly not more than a first proof-of-concept. We've included it here to motivate our tentative conclusion that the perspective introduced in this paper might have value not only for deriving interesting data sets, but also for tackling some of the tasks. In future work, we will explore methods that directly predict image layouts [e.g., (Tan et al., 2018)], comparing them to direct prediction approaches and evaluating whether the former methods offer a plus in interpretability through the step of predicting abstract models.

6 Conclusions

Our goal with this paper was to show, with detailed examples and descriptive statistics, that language / vision corpora can be a fertile hunting ground for semanticists interested in grounded lexical semantics. There is data pairing various, ever more complex, kinds of expressions with image objects (either parts of images, or images as a whole). Moreover, using these corpora, data sets can be derived that pair expressions, where a semantic relation holds between the parts that is recognisable to naive annotators (if not alway very clearly). As an example, we've used the perspective of treating images as models to retrieve exemplar models via language descriptions (captions), and probe those for the likely presence of entities and relations in a mentioned situation. It is our hope that this perspective might be useful to other researchers, and with the code released with this paper, we invite everyone to ask their own questions of the data, and to implement ideas on how to learn grounded interpretation.

Acknowledgements This work was done while I was at Bielefeld University and supported by the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277), which is funded by the German Research Foundation (DFG). I thank Sina Zarrieß and the anonymous reviewers for comments.

References

Abzianidze, L., J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D.-D. Nguyen, and J. Bos (2017, April). The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, pp. 242–247. Association for Computational Linguistics.

Bos, J. (2003). Exploring model building for natural language understanding. In *Workshop on Inference in Computational Semantics (ICoS)*.

Bos, J., V. Basile, K. Evang, N. Venhuizen, and J. Bjerva (2017). The groningen meaning bank. In N. Ide and J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation*, Volume 2, pp. 463–496. Springer.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil (2018). Universal Sentence Encoder. *ArXiv*.

Chierchia, G. and S. McConnell-Ginet (1990). *Meaning and Grammar: An Introduction to Semantics*. Cambridge, MA, USA: MIT Press.

Cirik, V., T. Berg-kirkpatrick, and L.-p. Morency (2018). Using Syntax to Ground Referring Expressions in Natural Images. In *AAAI 2018*.

Dagan, I., O. Glickman, and B. Magnini (2006). The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, Berlin, Heidelberg, pp. 177–190. Springer-Verlag.

Davies, E. R. (2012). *Computer and Machine Vision: Theory, Algorithms, Practicalities* (4th ed.). Amsterdam, Boston, Heidelberg, London: Elsevier.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, USA: MIT Press.

Fine, K. (2017). A Theory of Truthmaker Content I: Conjunction, Disjunction and Negation. *Journal of Philosophical Logic 46*(6), 625–674.

He, K., G. Gkioxari, P. Dollár, and R. Girshick (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Hürlimann, M. and J. Bos (2016). Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. In *5th Workshop on Vision and Language*, Berlin, Germany, pp. 10–18.

Kazemzadeh, S., V. Ordonez, M. Matten, and T. L. Berg (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, pp. 787–798.

Krause, J., J. Johnson, R. Krishna, and L. Fei-Fei (2017, January). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick (2014). Microsoft coco: Common objects in context. In *Computer Vision ECCV 2014*, Volume 8693, pp. 740–755. Springer International Publishing.

Mao, J., J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy (2016, June). Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR 2016*, Las Vegas, USA.

Marconi, D. (1997). *Lexical Competence*. Cambridge, Mass., USA: MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, USA: W.H. Freeman.

Moss, L. (2015). Natural logic. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantic Theory 2nd edition*. Wiley-Blackwell.

Plummer, B. A., L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of ICCV*.

Redmon, J. and A. Farhadi (2018). Yolov3: An incremental improvement. *arXiv*.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*(3), 211–252.

Schlangen, D., S. Zarrieß, and C. Kennington (2016, August). Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*, Berlin, Germany.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.

Tan, F., S. Feng, and V. Ordonez (2018). Text2Scene: Generating Abstract Scenes from Textual Descriptions. *ArXiv*.

Young, P., A. Lai, M. Hodosh, and J. Hockenmaier (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2*.

Yu, L., P. Poirson, S. Yang, A. Berg, and B. T.L. (2016). Modeling context in referring expressions. In *Computer Vision ECCV 2016*, Volume 9906 of *Lecture Notes in Computer Science*. Springer.