Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution

Casey Kennington
CITEC, Bielefeld University
Universitätsstraße 25
33615 Bielefeld, Germany
`ckennington@cit-ec.`
`uni-bielefeld.de`

David Schlangen
CITEC, Bielefeld University
Universitätsstraße 25
33615 Bielefeld, Germany
`david.schlangen@`
`uni-bielefeld.de`

Abstract

An elementary way of using language is to *refer* to objects. Often, these objects are physically present in the shared environment and reference is done via mention of perceivable properties of the objects. This is a type of language use that is modelled well neither by logical semantics nor by distributional semantics, the former focusing on inferential relations between expressed propositions, the latter on similarity relations between words or phrases. We present an account of word and phrase meaning that is perceptually grounded, trainable, compositional, and 'dialogue-plausible' in that it computes meanings word-by-word. We show that the approach performs well (with an accuracy of 65% on a 1-out-of-32 reference resolution task) on direct descriptions and target/landmark descriptions, even when trained with less than 800 training examples and automatically transcribed utterances.

1 Introduction

The most basic, fundamental site of language use is co-located dialogue (Fillmore, 1975; Clark, 1996) and referring to objects, as in Example (1), is a common occurrence in such a co-located setting.

(1)   *The green book on the left next to the mug.*

Logical semantics (Montague, 1973; Gamut, 1991; Partee et al., 1993) has little to say about this process – its focus is on the construction of syntactically manipulable objects that model inferential relations; here, e.g. the inference that there are (at least) two objects. Vector space approaches to distributional semantics (Turney and Pantel, 2010) similarly focuses on something else, namely semantic similarity relations between words or phrases (e.g. finding closeness for "coloured tome on the right of the cup"). Neither approach by itself says anything about processing; typically, the assumption in applications is that fully presented phrases are being processed.

Lacking in these approaches is a notion of *grounding* of symbols in features of the world (Harnad, 1990).[1] In this paper, we present an account of word and phrase meaning that is (a) perceptually grounded in that it provides a link between words and (computer) vision features of real images, (b) trainable, as that link is learned from examples of language use, (c) compositional in that the meaning of phrases is a function of that of its parts and composition is driven by structural analysis, and (d) 'dialogue-plausible' in that it computes meanings incrementally, word-by-word and can work with noisy input from an automatic speech recogniser (ASR). We show that the approach performs well (with an accuracy of 65% on a reference resolution task out of 32 objects) on direct descriptions as well as target/landmark descriptions, even when trained with little data (less than 800 training examples).

In the following section we will give a background on reference resolution, followed by a description of our model. We will then describe the data we used and explain our evaluations. We finish by giving results, providing some additional analysis, and discussion.

2 Background: Reference Resolution

Reference resolution (RR) is the task of resolving referring expressions (REs; as in Example (1)) to a *referent*, the entity to which they are intended to refer. Following Kennington et al. (2015a), this can be formalised as a function f_{rr} that, given a representation U of the RE and a representation W

---

[1]But see discussion below of recent extensions of these approaches taking this into account.

of the (relevant aspects of the) world, returns I^*, the identifier of one the objects in the world that is the referent of the RE. A number of recent papers have used stochastic models for f_{rr} where, given W and U, a distribution over a specified set of candidate entities in W is obtained and the probability assigned to each entity represents the strength of belief that it is the referent. The referent is then the argmax:

$$I^* = \underset{I}{\operatorname{argmax}} P(I|U, W) \qquad (1)$$

Recently, generative approaches, including our own, have been presented (Funakoshi et al., 2012; Kennington et al., 2013; Kennington et al., 2014; Kennington et al., 2015b; Engonopoulos et al., 2013) which model U as words or ngrams and the world W as a set of objects in a virtual game board, represented as a set properties or concepts (in some cases, extra-linguistic or discourse aspects were also modelled in W, such as deixis). In Matuszek et al. (2014), W was represented as a distribution over properties of tangible objects and U was a Combinatory Categorical Grammar parse. In all of these approaches, the objects are distinct and represented via symbolically specified *properties*, such as colour and shape. The set of properties is either read directly from the world if it is virtual, or computed (i.e., discretised) from the real world objects.

In this paper, we learn a mapping from W to U directly, without mediating symbolic properties; such a mapping is a kind of perceptual grounding of meaning between W and U. Situated RR is a convenient setting for learning perceptually-grounded meaning, as objects that are referred to are physically present, are described by the RE, and have visual features that can be computationally extracted and represented.

Further comparison to related work will be discussed in Section 5.

3 Modelling Reference to Visible Objects

Overview As a representative of the kind of model explained above with formula (1), we want our model to compute a probability distribution over candidate objects, given a RE (or rather, possibly just a prefix of it). We break this task down into components: The basis of our model is a model of word meaning as a function from perceptual features of a given object to a judgement about how well a word and that object "fit together". (See Section 5 for discussion of prior uses of this "words as classifiers"-approach.) This can (loosely) be seen as corresponding to the *intension* of a word, which for example in Montague's approach is similarly modelled as a function, but from possible worlds to extensions (Gamut, 1991). We model two different types of words / word meanings: those picking out properties of single objects (e.g., "green" in "the green book"), following Kennington et al. (2015a), and those picking out relations of two objects (e.g., "next to" in (1)), going beyond Kennington et al. (2015a). These word meanings are learned from instances of language use.

The second component then is the application of these word meanings in the context of an actual reference and within a phrase. This application gives the desired result of a probability distribution over candidate objects, where the probability expresses the strength of belief in the object falling in the *extension* of the expression. Here we model two different types of composition, of what we call *simple references* and *relational references*. These applications are strictly compositional in the sense that the meanings of the more complex constructions are a function of those of their parts.

Word Meanings The first type of word (or rather, word meaning) we model picks out a single object via its visual properties. (At least, this is what we use here; any type of feature could be used.) To model this, we train for each word w from our corpus of REs a binary logistic regression classifier that takes a representation of a candidate object via visual features ($\mathbf{x}$) and returns a probability p_w for it being a good fit to the word (where $\mathbf{w}$ is the weight vector that is learned and σ is the logistic function):

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x} + b) \qquad (2)$$

Formalising the correspondence mentioned above, the *intension* of a word can in this approach then be seen as the classifier itself, a function from a representation of an object to a probability:

$$[\![w]\!]_{obj} = \lambda\mathbf{x}.p_w(\mathbf{x}) \qquad (3)$$

(Where $[\![w]\!]$ denotes the meaning of w, and $\mathbf{x}$ is of the type of feature given by f_{obj}, the function computing a feature representation for a given object.)

We train these classifiers using a corpus of REs (further described in Section 4), coupled with representations of the scenes in which they were used and an annotation of the referent of that scene. The setting was restricted to reference to single objects. To get positive training examples, we pair each word of a RE with the features of the referent. To get negative training examples, we pair the word with features of (randomly picked) other objects present in the same scene, but *not* referred to by it. This selection of negative examples makes the assumption that the words from the RE apply only to the referent. This is wrong as a strict rule, as other objects could have similar visual features as the referent; for this to work, however, this has to be the case only more often than it is not.

The second type of word that we model expresses a relation between objects. Its meaning is trained in a similar fashion, except that it is presented a vector of features of a *pair* of objects, such as their euclidean distance, vertical and horizontal differences, and binary features denoting higher than/lower than and left/right relationships.

Application and Composition The model just described gives us a prediction for a pair of word and object (or pair of objects). What we wanted, however, is a distribution over all candidate objects in a given utterance situation, and not only for individual words, but for (incrementally growing) REs. Again as mentioned above, we model two types of application and composition. First, what we call '**simple references**'—which roughly corresponds to simple NPs—that refer only by mentioning properties of the referent (e.g. "the red cross on the left"). To get a distribution for a single word, we apply the word classifier (the *intension*) to all candidate objects and normalise; this can then be seen as the *extension* of the word in a given (here, visual) discourse universe W, which provides the candidate objects ($\mathbf{x}_i$ is the feature vector for object i, $normalize()$ vectorized normalisation, and I a random variable ranging over the candidates):

$$
\begin{aligned}
[\![w]\!]_{obj}^{W} = \\
normalize(([\![w]\!]_{obj}(\mathbf{x}_1), \ldots, [\![w]\!]_{obj}(\mathbf{x}_k))) = \\
normalize((p_w(\mathbf{x}_1), \ldots, p_w(\mathbf{x}_k))) = P(I|w) \quad (4)
\end{aligned}
$$

In effect, this combines the individual classifiers into something like a multi-class logistic regression / maximum entropy model—but, *nota bene*, only for application. The training regime did not need to make any assumptions about the number of objects present, as it trained classifiers for a 2-class problem (how well does this given object fit to the word?). The multi-class nature is also indicated in Figure 1, which shows multiple applications of the logistic regression network for a word, and a normalisation layer on top.
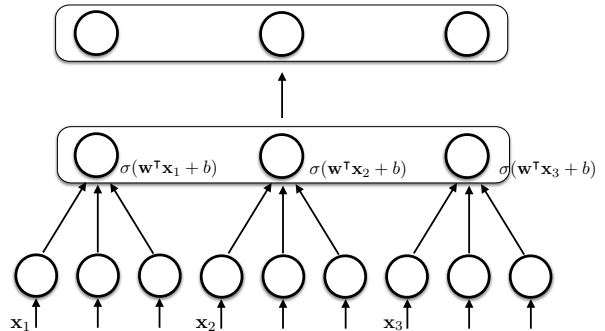

Figure 1: Representation as network with normalisation layer.

To compose the evidence from individual words $w_1, \ldots, w_k$ into a prediction for a 'simple' RE $[_{sr} w_1, \ldots, w_k]$ (where the bracketing indicates the structural assumption that the words belong to one, possibly incomplete, 'simple reference'), we average the contributions of its constituent words. The averaging function $avg()$ over distributions then is the contribution of the construction 'simple reference (phrase)', sr, and the meaning of the whole phrase is the application of the meaning of the construction to the meaning of the words:

$$
\begin{aligned}
[\![[_{sr} w_1, \ldots, w_k]]\!]^{W} = [\![sr]\!]^{W} [\![w_1, \ldots, w_k]\!]^{W} = \\
avg([\![w_1]\!]^{W}, \ldots, [\![w_k]\!]^{W}) \quad (5)
\end{aligned}
$$

where $avg()$ is defined as

$$
\begin{aligned}
avg([\![w_1]\!]^{W}, [\![w_2]\!]^{W}) = P_{avg}(I|w_1, w_2) \\
\text{with } P_{avg}(I = i|w_1, w_2) = \\
\frac{1}{2}(P(I = i|w_1) + P(I = i|w_2)) \text{ for } i \in I \quad (6)
\end{aligned}
$$

The averaging function is inherently incremental, in the sense that $avg(a, b, c) = avg(avg(a, b), c)$ and hence it can be extended "on the right". This represents an incremental model where new information from the current increment is added to what is already known, resulting in an intersective way of composing the meaning of the phrase. This cannot account for all constructions (such as negation or generally quantification), of course; we leave exploring other constructions that could occur even in our 'simple references' to future work.

Relational references such as in Example (1) from the introduction have a more complex structure, being a relation between a (simple) reference to a *landmark* and a (simple) reference to a *target*. This structure is indicated abstractly in the following 'parse': $[_{rel}[_{sr}w_1, \ldots, w_k][_r r_1, \ldots, r_n][_{sr}w'_1, \ldots, w'_m]]$, where the w are the target words, r the relational expression words, and w' the landmark words.

As mentioned above, the relational expression similarly is treated as a classifier (in fact, technically we contract expressions such as "to the left of" into a single token and learn one classifier for it), but expressing a judgement for pairs of objects. It can be applied to a specific scene with a set of candidate objects (and hence, candidate pairs) in a similar way by applying the classifier to all pairs and normalising, resulting in a distribution over pairs:

$$\llbracket r \rrbracket^W = P(R_1, R_2 | r) \tag{7}$$

We expect the meaning of the phrase to be a function of the meaning of the constituent parts (the simple references, the relation expression, and the construction), that is:

$$\llbracket [_{rel}[_{sr}w_1, \ldots, w_k][_r r][_{sr}w'_1, \ldots, w'_m]] \rrbracket =$$
$$\llbracket rel \rrbracket (\llbracket sr \rrbracket \llbracket w_1 \ldots w_k \rrbracket, \llbracket r \rrbracket, \llbracket sr \rrbracket \llbracket w'_1 \ldots w'_m \rrbracket) \tag{8}$$

(dropping the indicator for concrete application, ^{W} on $\llbracket \ \rrbracket$, for reasons of space and readability).

What is the contribution of the relational construction, $\llbracket rel \rrbracket$? Intuitively, what we want to express here is that the belief in an object being the intended referent should combine the evidence from the simple reference to the landmark object (e.g., "the mug" in (1)), from the simple (but presumably deficient) reference to the target object ("the green book on the left"), and that for the relation between them ("next to"). Instead of averaging (that is, combining additively), as for sr, we combine this evidence multiplicatively here: If the target constituent contributes $P(I_t | w_1, \ldots, w_k)$, the landmark constituent $P(I_l | w'_1, \ldots, w'_m)$, and the relation expression $P(R_1, R_2 | r)$, with I_l, I_t, R_1 and R_2 all having the same domain, the set of all candidate objects, then the combination is

$$P(R_1 | w_1, \ldots, w_k, r, w'_1, \ldots, w'_m) =$$
$$\sum_{R_2} \sum_{I_l} \sum_{I_t} P(R_1, R_2 | r) * P(I_l | w'_1, \ldots, w'_m) *$$
$$P(I_t | w_1, \ldots, w_k) * P(R_1 | I_t) * P(R_2 | I_l) \tag{9}$$

The last two factors force identity on the elements of the pair and target and landmark, respectively (they are not learnt, but rather set to be 0 unless the values of R and I are equal), and so effectively reduce the summations so that all pairs need to be evaluated only once. The contribution of the construction then is this multiplication of the contributions of the parts, together with the factors enforcing that the pairs being evaluated by the relation expression consist of the objects evaluated by target and landmark expression, respectively.

In the following section, we will explain the data we collected and used to evaluate our model, the evaluation procedure, and the results.
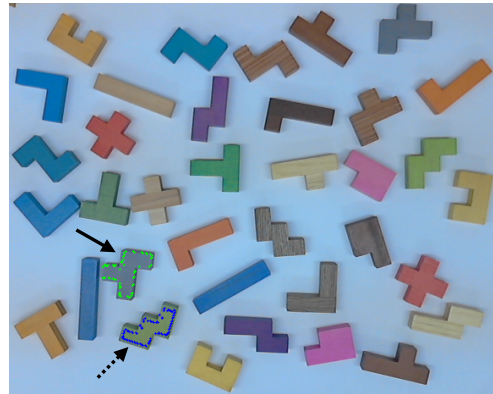
4 Experiments

Figure 2: Example episode for *phase-2* where the target is outlined in green (solid arrow added here for presentation), the landmark outlined in blue (dashed arrow).

Data We evaluated our model using data we collected in a Wizard-of-Oz setting (that is, a human/computer interaction setting where parts of the functionality of the computer system were provided by a human experimentor). Participants were seated in front of a table with 36 Pentomino puzzle pieces that were randomly placed with some space between them, as shown in Figure 2. Above the table was a camera that recorded a video feed of the objects, processed using OpenCV (Pulli et al., 2012) to segment the objects (see below for details); of those, one (or one pair) was chosen randomly by the experiment software. The video image was presented to the participant on a display placed behind the table, but with the randomly selected piece (or pair of pieces) indicated by an overlay.

The task of the participant was to refer to that object using only speech, as if identifying it for a friend sitting next to the participant. The wizard

295

(experimentor) had an identical screen depicting the scene but not the selected object. The wizard listened to the participant's RE and clicked on the object she thought was being referred on her screen. If it was the target object, a tone sounded and a new object was randomly chosen. This constituted a single *episode*. If a wrong object was clicked, a different tone sounded, the episode was flagged, and a new episode began. At varied intervals, the participant was instructed to "shuffle" the board between episodes by moving around the pieces.

The first half of the allotted time constituted phase-1. After phase-1 was complete, instructions for phase-2 were explained: the screen showed the target and also a landmark object, outlined in blue, near the target (again, see Figure 2). The participant was to refer to the target using the landmark. (In the instructions, the concepts of landmark and target were explained in general terms.) All other instructions remained the same as phase-1. The target's identifier, which was always known beforehand, was always recorded. For phase-2, the landmark's identifier was also recorded.

Nine participants (6 female, 3 male; avg. age of 22) took part in the study; the language of the study was German. Phase-1 for one participant and phase-2 for another participant were not used due to misunderstanding and a technical difficulty. This produced a corpus of 870 non-flagged episodes in total. Even though each episode had 36 objects in the scene, all objects were not always recognised by the computer vision processing. On average, 32 objects were recognized.

To obtain transcriptions, we used Google Web Speech (with a word error rate of 0.65, as determined by comparing to a hand transcribed sample) This resulted in 1587 distinct words, with 15.53 words on average per episode. The objects were not manipulated in any way during an episode, so the episode was guaranteed to remain static during a RE and a single image is sufficient to represent the layout of one episode's scene. Each scene was processed using computer vision techniques to obtain low-level features for each (detected) object in the scene which were used for the word classifiers.

We annotated each episode's RE with a simple tagging scheme that segmented the RE into words that directly referred to the target, words that directly referred to the landmark (or multiple landmarks, in some cases) and the relation words. For certain word types, additional information about the word was included in the tag if it described colour, shape, or spatial placement (denoted *contributing* REs in the evaluations below). The *direction* of certain relation words was normalised (e.g., *left-of* should always denote a landmark-target relation). This represents a minimal amount of "syntactic" information needed for the application of the classifiers and the composition of the phrase meanings. We leave applying a syntactic parser to future work. An example RE in the original German (as recognised by the ASR), English gloss, and tags for each word is given in (2).

(2)　a.　*grauer stein über dem grünen m unten links*
　　b.　gray block above the green m bottom left
　　c.　tc　ts　r　l　lc　ls tf　　tf

To obtain visual features of each object, we used the same simple computer-vision pipeline of object segmentation and contour reconstruction as used by Kennington et al. (2015a), providing us with RGB representations for the colour and features such as skewness, number of edges etc. for the shapes.

Procedure　We break down our data as follows: episodes where the target was referred directly via a 'simple reference' construction (DD; 410 episodes) and episodes where a target was referred via a landmark relation (RD; 460 episodes). We also test with either knowledge about structure (simple or relational reference) provided (ST) or not (WO, for "words-only"). All results shown are from 10-fold cross validations averaged over 10 runs; where for evaluations labelled RD the training data always includes all of DD plus 9 folds of RD, testing on RD. The sets address the following questions:

- how well does the sr model work on its own with just words? – DD.WO
- how well does the sr model work when it knows about REs? – DD.ST
- how well does the sr model work when it knows about REs, but not about relations? – RD.ST (sr)
- how well does the model learn relation words after it has learned about sr? RD.ST (r)
- how well does the rr model work (together with the sr)? RD.ST with DD.ST (rr)

Words were stemmed using the NLTK (Loper and Bird, 2002) Snowball Stemmer, reducing the

vocabulary size to 1306. Due to sparsity, for relation words with a token count of less than 4 (found by ranging over values in a held-out set) relational features were piped into an UNK relation, which was used for unseen relations during evaluation (we assume the UNK relation would learn a general notion of 'nearness'). For the individual word classifiers, we always paired one negative example with one positive example.

For this evaluation, word classifiers for sr were given the following features: RGB values, HSV values, x and y coordinates of the centroids, euclidean distance of centroid from the center, and number of edges. The relation classifiers received information relating two objects, namely the euclidean distance between them, the vertical and horizontal distances, and two binary features that denoted if the landmark was higher than/lower than or left/right of the target.
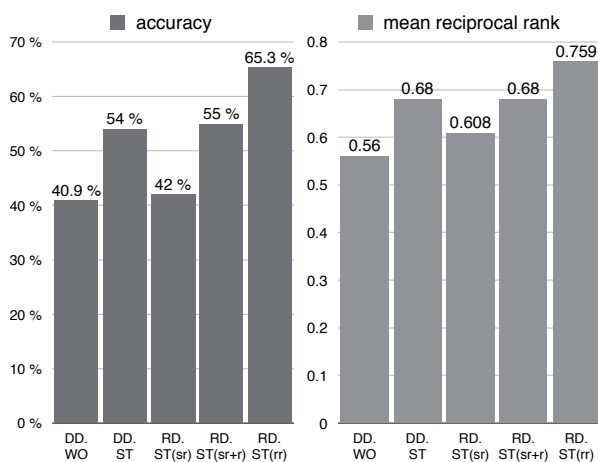

Figure 3: Results of our evaluation.

Metrics for Evaluation To give a picture of the overall performance of the model, we report **accuracy** (how often was the argmax the gold target) and **mean reciprocal rank** (MRR) of the gold target in the distribution over all the objects (like accuracy, higher MRR values are better; values range between 0 and 1). The use of MRR is motivated by the assumption that in general, a good rank for the correct object is desirable, even if it doesn't reach the first position, as when integrated in a dialogue system this information might still be useful to formulate clarification questions.

Results Figure 3 shows the results. (Random baseline of 1/32 or 3% not shown in plot.) DD.WO shows how well the sr model performs using the whole utterances and not just the REs. (Note that

all evaluations are on noisy ASR transcriptions.) DD.ST adds structure by only considering words that are part of the actual RE, improving the results further. The remaining sets evaluate the contributions of the rr model. RD.ST (sr) does this indirectly, by including the target and landmark simple references, but not the model for the relations; the task here is to resolve target and landmark SRs as they are. This provides the baseline for the next two evaluations, which include the relation model. In RD.ST ($sr+r$), the model learns SRs from DD data and only relations from RD. The performance is substantially better than the baseline without the relation model. Performance is best finally for RD.ST (rr), where the landmark and target SRs in the training portion of RD also contribute to the word models.

The *mean reciprocal rank* scores follow a similar pattern and show that even though the target object was not the argmax of the distribution, on average it was high in the distribution. For all evaluations, the average standard deviation across the 10 runs was very small (0.01), meaning the model was fairly stable, despite the possibility of one run having randomly chosen more discriminating negative examples. Our conclusion from these experiments is that despite the small amount of training data and noise from ASR as well as the scene, the model is robust and yields respectable results.
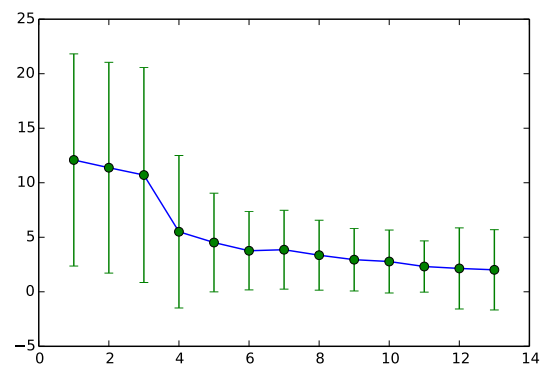

Figure 5: Incremental results: average rank improves over time

Incremental Results Figure 5 shows how our rr model processes incrementally, by giving the *average rank* of the (gold) target at each increment for the REs with the most common length in our data (13 words, of which there were 64 examples). A system that works incrementally would have a monotonically decreasing average rank as the utterance unfolds. The overall trend as shown in that
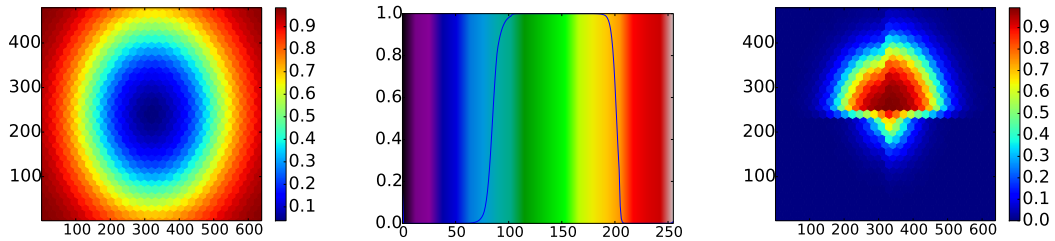
Figure 4: Each plot represents how well selected words fit assumptions about their lexical semantics: the leftmost plot *ecke* (*corner*) yields higher probabilities as objects are closer to the corner; the middle plot *grün* (*green*) yields higher probabilities when the colour spectrum values are nearer to green; the rightmost plot *über* (*above*) yields higher probabilities when targets are nearer to a landmark set in the middle.

Figure is as expected. There is a slight increase between 6-7, though very small (a difference of 0.09). Overall, these results seem to show that our model indeed works *intersectively* and "zooms in" on the intended referent.

4.1 Further Analysis

Analysis of Selected Words We analysed several individual word classifiers to determine how well their predictions match assumptions about their lexical semantics. For example, for the spatial word *Ecke* (*corner*), we would expect its classifier to return high probabilities if features related to an object's position (e.g., x and y coordinates, distance from the center) are near corners of the scene. The leftmost plot in Figure 4 shows that this is indeed the case; by holding all non-position features constant and ranging over all points on the screen, we can see that the classifier gives high probabilities around the edges, particularly in the four corners, and very low probabilities in the middle region. Similarly for the colour word *grün*, the centre plot in Figure 4 (overlaid with a colour spectrum) shows high probabilities are given when presented with the colour green, as expected. Similarly, for the relational word *über* (*above*), by treating the center point as the landmark and ranging over all other points on the plot for the target, the *über* classifier gives high probabilities when directly above the center point, with linear negative growth as the distance from the landmark increases.

Note that we selected the type of feature to vary here for presentation; all classifiers get the full feature set and learn automatically to "ignore" the irrelevant features (e.g., that for *grün* does not respond to variations in positional features). They do this wuite well, but we noticed some 'blurring', due to not all combinations of colours and shape being represented in the objects in the training set.

Analysis of Incremental Processing Figure 6 finally shows the interpretation of the RE in Example (2) in the scene from Figure 2. The top row depicts the distribution over objects (true target shown in red) after the relation word *unten* (bottom) is uttered; the second row that for landmark objects, after the landmark description begins (*dem grünen m* / *the green m*). The third row (target objects), ceases to change after the relational word is uttered, but continues again as additional target words are uttered (*unten links* / *bottom left*). While the true target is ranked highly already on the basis of the target SR alone, it is only when the relational information is added (top row) that it becomes argmax.

Discussion We did not explore how well our model could handle generalised quantifiers, such as *all* (e.g., *all the red objects*) or a specific number of objects (e.g., *the two green Ts*). We speculate that one could see as the contribution of words such as *all* or *two* a change to how the distribution is evaluated ("return the n top candidates"). Our model also doesn't yet directly handle more descriptive REs like *the cross in the top-right corner on the left*, as *left* is learned as a global term, or negation (*the cross that's not red*). We leave exploring such constructions to future work.

5 Related Work

Kelleher et al. (2005) approached RR using perceptually-grounded models, focusing on saliency and discourse context. In Gorniak and Roy (2004), descriptions of objects were used to learn a perceptually-grounded meaning with focus on spatial terms such as *on the left*. Steels and Belpaeme (2005) used neural networks to connect language with colour terms by interacting with humans. Larsson (2013) is closest in spirit to what we are attempting here; he provides a detailed
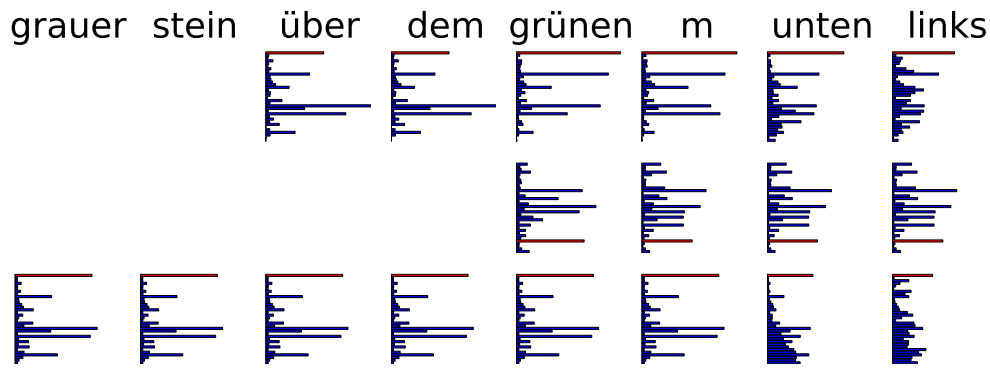
Figure 6: A depiction of the model working incrementally for the RE in Example (2): the distribution over objects for relation is row 1, landmark is row 2, target is row 3.

formal semantics for similarly descriptive terms, where parts of the semantics are modelled by a perceptual classifier. These approaches had limited lexicons (where we attempt to model all words in our corpus), and do not process incrementally, which we do here.

Recent efforts in multimodal distributional semantics have also looked at modelling word meaning based on visual context. Originally, vector space distributional semantics focused words in the context of other words (Turney and Pantel, 2010); recent multimodal approaches also consider low-level features from images. Bruni et al. (2012) and Bruni et al. (2014) for example model word meaning by word and visual context; each modality is represented by a vector, fused by concatenation. Socher et al. (2014) and Kiros et al. (2014) present approaches where words/phrases and images are mapped into the same high-dimensional space. While these approaches similarly provide a link between words and images, they are typically tailored towards a different setting (the words being descriptions of the whole image, and not utterance intended to perform a function within a visual situation). We leave more detailed exploration of similarities and differences to future work and only note for now that our approach, relying on much simpler classifiers (log-linear, basically), works with much smaller data sets and additionally seem to provide an easier interface to more traditional ways of composition (see Section 3 above).

The issue of semantic compositionality is also actively discussed in the distributional semantics literature (see, e.g., (Mitchell and Lapata, 2010; Erk, 2013; Lewis and Steedman, 2013; Paperno

et al., 2014)), investigating how to combine vectors. This could be seen as composition on the level of intensions (if one sees distributional representations as intensions, as is variously hinted at, e.g. Erk (2013)). In our approach, composition is done on the extensional level (by interpolating distributions over candidate objects).

We do not see our approach as being in opposition to these attempts. Rather, we envision a system of semantics that combines traditional symbolic expressions (on which inferences can be modelled via syntactic calculi) with distributed representations (which model conceptual knowledge / semantic networks, as well as encyclopedic knowledge) and with our action-based (namely, identification in the environment via perceptual information) semantics. This line of approach is connected to a number of recent works (e.g., (Erk, 2013; Lewis and Steedman, 2013; Larsson, 2013)); for now, exploring its ramifications is left for future work.

6 Conclusion

In this paper, we presented a model of reference resolution that learns a perceptually-grounded meaning of words, including relational words. The model is simple, compositional, and robust despite low amounts of training data and noisy modalities. Our model is not without limitations; it so far only handles definite descriptions, yet there are other ways to refer to real-world objects, such as via pronouns and deixis. A unified model that can handle all of these, similar in spirit perhaps to Funakoshi et al. (2012), but with perceptual groundings, is left for future work. Our approach could also benefit from improved object segmentation and repre-

sentation.

Our next steps with this model is to handle compositional structures without relying on our closed tag set (e.g., using a syntactic parser). We also plan to test our model in a natural, interactive dialogue system.

References

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 136–145.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Herbert H Clark. 1996. *Using Language*, volume 23. Cambridge University Press.

Nikos Engonopoulos, Martin Villalba, Ivan Titov, and Alexander Koller. 2013. Predicting the resolution of referring expressions from user behavior. In *Proceedings of EMLNP*, pages 1354–1359, Seattle, Washington, USA. Association for Computational Linguistics.

Katrin Erk. 2013. Towards a semantics for distributional representations. In *Proceedings of IWCS*, pages 1–11, Potsdam, Germany.

Charles J Fillmore. 1975. Pragmatics and the description of discourse. *Radical pragmatics*, pages 143–166.

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. 2012. A Unified Probabilistic Approach to Referring Expressions. In *Proceedings of SIGDial*, pages 237–246, Seoul, South Korea, July. Association for Computational Linguistics.

L T F Gamut. 1991. *Logic, Language and Meaning: Intensional Logic and Logical Grammar*, volume 2. Chicago University Press, Chicago.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D*, 42:335–346.

John Kelleher, Fintan Costello, and Jofsef Van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1–2):62–102.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *Proceedings of SIGdial*.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model. In *Proceedings of CoLing*.

Casey Kennington, Livia Dia, and David Schlangen. 2015a. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of IWCS*. Association for Computational Linguistics.

Casey Kennington, Ryu Iida, Takenobu Tokunaga, and David Schlangen. 2015b. Incrementally Tracking Reference in Human/Human Dialogue Using Linguistic and Extra-Linguistic Information. In *NAACL*, Denver, U.S.A. Association for Computational Linguistics.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *Proceedings of NIPS 2014 Deep Learning Workshop*, pages 1–13.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Mike Lewis and Mark Steedman. 2013. Combined Distributional and Logical Semantics. *Transactions of the ACL*, 1:179–192.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *AAAI*. AAAI Press.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, November.

Richard Montague. 1973. The Proper Treatment of Quantifikation in Ordinary English. In J Hintikka, J Moravcsik, and P Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221–242, Dordrecht. Reidel.

Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of ACL*, pages 90–99.

Barbara H Partee, Alice ter Meuelen, and Robert E Wall. 1993. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dordrecht.

Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. 2012. Real-time computer vision with OpenCV. *Communications of the ACM*, 55(6):61–69.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics (TACL)*, 2:207–218.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: a case study for colour. *The Behavioral and brain sciences*, 28(4):469–489; discussion 489–529.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Artificial Intelligence*, 37(1):141–188.