A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution

Casey Kennington
CITEC, Bielefeld University
ckennington@cit-ec.uni-bielefeld.de

Livia Dia
University of York
lad507@york.ac.uk

David Schlangen
Bielefeld University
david.schlangen@uni-bielefeld.de

Abstract

A large part of human communication involves referring to entities in the world, and often these entities are objects that are visually present for the interlocutors. A computer system that aims to resolve such references needs to tackle a complex task: objects and their visual features must be determined, the referring expressions must be recognised, extra-linguistic information such as eye gaze or pointing gestures must be incorporated — and the intended connection between words and world must be reconstructed. In this paper, we introduce a discriminative model of reference resolution that processes incrementally (i.e., word for word), is perceptually-grounded, and improves when interpolated with information from gaze and pointing gestures. We evaluated our model and found that it performed robustly in a realistic reference resolution task, when compared to a generative model.

1 Introduction

Reference to entities in the world via definite description makes up a large part of human communication (Poesio and Vieira, 1997). In task-oriented situations, these references are often to entities that are visible in the shared environment. In such co-located settings, interlocutors can make use of extra-linguistic cues such as gaze or pointing gestures. Furthermore, listeners resolve references as they unfold, often identifying the referred entity before the end of the reference (as found, *inter alia*, by Tanenhaus and Spivey-Knowlton (1995); Spivey (2002)). Computational research on reference resolution, however, has mostly focused on offline processing of full, completed referring expressions, and not attempted to model this online nature of human reference resolution. On a more technical level, most of the models making use of stochastic information (see discussion below) have been generative models; even though such models are known to often have certain disadvantages compared to discriminative models.[1]

In this paper, we introduce a discriminative model of reference resolution that is *incremental* in that it does not wait until the end of an utterance to process, rather it updates its interpretation at each word. Moreover, the semantics of each word is *perceptually grounded* in visual information from the world. We evaluated our model and found that it works robustly when compared to a similar generative approach.

In the following section we explain the task of reference resolution and discuss related work. That is followed by an explanation of our model and evaluation experiment. We end with some analyses of the model's strengths and areas of improvement.

2 Background and Related Work on Reference Resolution

Reference resolution (RR) is the task of resolving referring expressions (henceforth REs; e.g., *the red one on the left*) to the entity to which they are intended to refer; the *referent*. This can be formalised as a function f_{rr} that, given a representation U of the RE and a representation W of the (relevant aspects

[1] But see the nuanced discussion by Ng A.Y. & Jordan M. I. (2002).

Proceedings of the 11th International Conference on Computational Semantics, pages 195–205,
London, UK, April 15-17 2015.

of the) world (which can include aspects of the discourse context), returns I^*, the identifier of one the objects in the world that is the intended referent of the RE.

$$I^* = f_{rr}(U, W) \tag{1}$$

This function f_{rr} can be specified in a variety of ways. A number of recent papers have used stochastic models using the following approach: given W and U, the goal of RR is to obtain a distribution over a specified set of candidate entities in that world, where the probability assigned to each entity represents the strength of belief that it is the referred one. The referred object is then the argmax of that distribution:

$$I^* = \underset{I}{\operatorname{argmax}}\, P(I|U, W) \tag{2}$$

We have worked in this area before. Kennington and Schlangen (2013) we applied Markov Logic Networks (Richardson and Domingos, 2006) to the task of computing the distribution over I. The world W, a virtual game board of puzzle pieces, was represented symbolically (e.g., objects were represented by their properties such as colour and shape). The utterance U was represented by its words. We repeated the experiments later in Kennington et al. (2013) where the utterance and world were represented in the same way, but the model that produced the distribution over the candidate objects was generative; it modeled the joint distribution over the objects and their properties, and the words in the utterance. (This model will be further discussed and used as a baseline for comparison below.) In Kennington et al. (2014); Hough et al. (2015) we used that same generative model and representation of W, but U was represented as a semantic abstraction.

Funakoshi et al. (2012) used a Bayesian network approach. The world W (in their case, a set of tangram puzzle pieces) was represented as a set of *concepts* (e.g., shape type), and U was represented by the words in the REs, in an interctive human-human setting. The Bayesian network was used to learn a mapping between concepts and U. The model could handle various types of REs, namely definite references, exophoric pronoun references, and deictic (pointing) references to objects. Similar data was used in Iida et al. (2011), but the mapping between U and W was done with a support vector machine classifier. We recently applied our generative model to this data, with improved results in some areas Kennington et al. (2015).

Engonopoulos et al. (2013) also used a generative approach; W was modeled as an *observation model* (i.e., a set of features over the objects in a 3D scene), and U was a *semantic* model that abstracted over the referring expression.

In Matuszek et al. (2014), W was represented as a distribution over properties (e.g., color and shape) of real-world objects (small wooden blocks of various shapes and colors) as represented by computer-vision output. U was represented as a semantic abstraction in the form of a Combinatory Categorical Grammar parse. Resolving I amounted to generatively computing a joint distribution over the representation of U and W.

In all of these approaches, the objects are distinct and have specific visual *properties*, such as color, shape, and spatial placement. The set of properties is defined and either read directly from the world if it is virtual, or computed (i.e., discretised) from the real world objects. In this paper, we take a different approach to representing the world W and how the distribution for I is computed. Instead of representing the world as a set of discrete properties or concepts, we represent the world with a set of more low-level visual features (e.g., color-values) and compute the distribution over objects discriminatively, as will be explained in Section 3. This represents a kind of perceptually-grounded learning of how visual features connect to words in the RE, where the meaning of the word is represented by a classifier, as explained below.

Treating words as classifiers working with perceptual information has been explored before. Steels and Belpaeme (2005) used neural networks to connect language with colour terms. Their model learned the way colours were used by interacting with humans. Larsson (2013) addressed integrating perceptual

meaning into a formal semantics by a model that was tested in a game where participants described a simple object as being on the *left* or *right* side of the game board. Both of these approaches focused on a very limited lexicon of words that are used to described visual objects, namely colours or left/right, and these approaches only took limited perceptual information into account. In this paper, we don't limit the lexicon to a certain class of words, rather we attempt to learn a perceptually-grounded meaning of all the words in a corpus (described below).

Situated RR is a convenient setting for learning perceptually-grounded meaning, as objects that are referred to physically exist, are described by the RE, and have visual features that can be computationally extracted and represented. Kelleher et al. (2005) approached RR using perceptually-grounded models, focusing on saliency and discourse context. The task of RR was also used 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*. However, unlike our approach, these approaches did not model the meaning of words directly, nor are they incremental.

3 A Model of Reference to Visible Objects

Our model interpolates information from the referring expression proper and from other, multimodal information sources such as gaze and pointing gestures. We will describe the parts of the model in turn.

3.1 A Discriminative Model of Linguistic Evidence

Overview As explained with formula (2), we want our model to give us a probability distribution over the candidate objects, given a referring expression (or, to be more precise, given possibly just a prefix of the expression). Instead of trying to learn a mapping between full referring expressions and objects, we break the problem down into one of learning a mapping between individual words and objects, and of composition of the evidence into a prediction for the full expression (or expression prefix).

The Word Model At the basis of the model then is a prediction for a given word and a given object of how well the object fits the word. To compute this, we trained for each word w from our corpus of referring expressions 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):

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\intercal \mathbf{x} + b) \tag{3}$$

This model is shown in Figure 1 as a one-layer neural network.

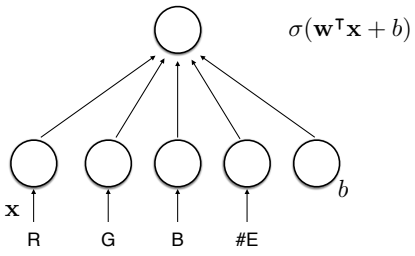

Figure 1: Representation as 1-layer NN

hue, saturation, value
red, green, blue
number of detected edges
x,y coordinates
horizontal skewness
vertical skewness
orientation value

Figure 2: List of features used in our model (total of 12 features).

We train these classifiers using a corpus of REs (further described below), 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 process is

shown in Algorithm 1. 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. This is so for our domain, and in general seems a plausible thing to assume that often words used in REs do indeed uniquely single out their referent.

Algorithm 1 Training algorithm for our model, each word classifier receives a set of positive and negative training examples, then maximum likelihood is computed for each word.

```
1: procedure TRAIN(frame)
2:     for each RE in corpus do
3:         for each word in RE do
4:             pair word with features of object referred to by RE; add to positive examples
5:             pair word with features of n objects not referred to by RE; add to negative examples
6:         end for
7:     end for
8:     for each word in vocabulary do
9:         train word binary classifier
10:    end for
11: end procedure
```

Application and Composition This model gives us a prediction for a pair of word and object. What we wanted, however, is a distribution over all candidate objects, and not only for individual words, but for (incrementally growing) full REs. To get the former, we apply the word/percept classifier to each candidate object, and normalise (where $\mathbf{x}_i$ is the feature vector for object i):

$$P(I = i|U = w, W) = \frac{p_w(\mathbf{x}_i)}{\sum_{k \in I} p_w(\mathbf{x}_k)} \tag{4}$$

In effect, this turns this into a multi-class logistic regression / maximum entropy model—but 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 1-class problem (how well does this given object fit to the word?). The multi-class nature is also indicated in Figure 3, which shows multiple applications of the network from Figure 1, with a normalisation layer on top.

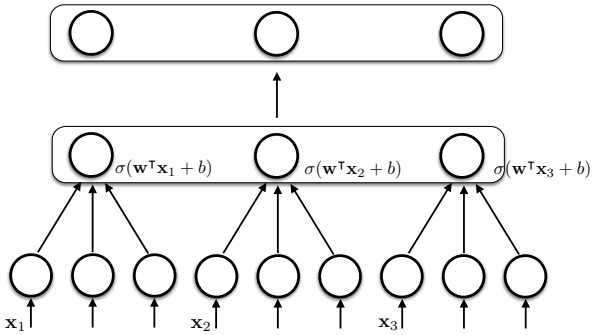

Figure 3: Representation as network with normalisation layer.

To compose the evidence from individual words into a prediction for the whole (prefix of the) referring expression, we chose the simplest possible approach, namely simply to average the previous distribution with the new one. This represents an incremental model where new information from the current increment is added to what is already known. Such a simple model represents an intersective way of composing the words (or, rather, their denotations). More sophisticated approaches can be imagined (e.g., using syntactic structure); we leave exploring them to future work.

$$P(I = i|U_1^k, W) = [P(I = i|U_1^{k-1}, W) + P(I = i|U^k, W)] * \frac{1}{2} \tag{5}$$

3.2 Evidence from Gaze and Deixis

The full model combines the evidence from linguistic information with evidence from other information sources such as the speaker's gaze and pointing gestures. For each, we calculate a reference point (R) on the scene: for gaze, the fixated point as provided by an eye tracker; for deixis, the point on the scene that was pointed at based on a vector calculated from the shoulder to the hand (as described in Kousidis et al. (2013), using the Microsoft Kinect). The centroids of all the objects (I) can then be compared to that reference point to yield a probability of that object being 'referred' by that modality (i.e., gazed at or pointed at) by introducing a Gaussian window over the location of the point:

$$p_{distance}(R_i, I_j; \sigma) = \exp -\frac{(x_i - x_j)^2}{2 * \sigma^2} * \exp -\frac{(y_i - y_j)^2}{2 * \sigma^2} \tag{6}$$

where the mean is R and σ is set by calculating the standard deviation of all the object centroids and the reference point. This can then be normalised over all the $p_{distance}$ scores to produce a distribution over I for each modality where the closer the object is to the reference point, the higher its probability. (We implicitly make the somewhat naive assumption here that the referred object will be looked at by the speaker most of the time during and around the RE. This is in general not true (Griffin and Bock, 2000), but works out here.)

Our final model of RR fuses the the three described modalities of speech, gaze, and deixis using a linear interpolation, where the α parameters are learned from held-out data by ranging over values such that the α values sum to one, and computing the average rank (metric explained below), retaining the α values that produced the best score for that set:

$$P(I|S) = P(I|S_1)\alpha_1 + P(I|S_2)\alpha_2 + P(I|S_3)(1 - \alpha_1 - \alpha_2) \tag{7}$$

4 Evaluation

We will now explain our evaluation experiment, including the data we used, the pre-processing performed on it, a generative model that we will compare to, and the metrics that we will use in our evaluation. We end this section with the results.

4.1 Data

We used data from the Pentomino puzzle domain as described by Kousidis et al. (2013). In this Wizard-of-Oz study, the participant was confronted with a game board containing 15 randomly selected Pentomino puzzle pieces (out of a repertoire of 12 shapes, and 6 colors). The positions of the pieces were randomly determined, but in such a way that the pieces grouped in the four corners of the screen, an example is shown in Figure 4. The participants were seated at a table in front of the screen. Their gaze was then calibrated with an eye tracker (*Seeingmachines FaceLab*) placed above the screen and their arm movements (captured by a Microsoft Kinect, also above the screen) were also calibrated. They were then given task instructions: (silently) choose a Pentomino tile on the screen and then instruct the computer system to select this piece by describing and pointing to it. When a piece was selected (by the wizard), the participant had to utter a confirmation (or give negative feedback) and a new board was generated and the process repeated. In this paper, we denote each of these instances as an *episode*. The utterances, board states, arm movements, and gaze information were recorded in a similar fashion as described in Kousidis et al. (2012). The wizard was instructed to elicit pointing gestures by waiting to select the participant-referred piece by several seconds, unless a pointing action by the participant had already occurred. When the wizard misunderstood, or a technical problem arose, the wizard had an option to flag the episode. In total, 1214 episodes were recorded from 8 participants (all university students). All but one were native speakers of German; the non-native spoke proficient German.

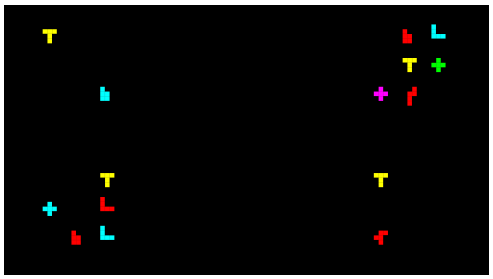

Figure 4: Example Pentomino board for gaze and deixis experiment; the yellow T in the top-right quadrant is the referred object.

Figure 5: Pentomino Board that has been distorted from its original form (Figure 4); all objects have distorted shapes and colors.

For each episode, the corpus contains two transcriptions of the utterance (one performed by expert transcribers, as well as one created by Google Web Speech; WER 49.8% when compared to expert transcription), deixis and gaze information, as well as an image of what the board looked like to the participant, see Figure 4. Removing episodes that were flagged by the Wizard yielded 1049 episodes with corresponding data. We also have a version of each board image that has been processed in a more involved way, which will now be explained.

Scene Processing We want our model to work with images of real objects as input, even though for our particular data the scenes are represented symbolically (that is, we know without uncertainty each piece's shape, color, and position). Using the images that were generated from these symbolic descriptions and performing computer vision on them does not introduce much uncertainty, as there is no variation in color or appearance of individual shapes, and so the data cannot serve to form generalisations. To get closer to conditions as they would hold when working with camera images (e.g., variations of color due to variations in lighting, distortion of shapes due to camera angles, etc.), we pre-processed these images:[2] We shifted the color spectrum as follows: the hue channel by a random number between -15 and 15 and the saturation and value channels by a random number between -50 and 50. For the object shapes, we apply affine transformations defined by two randomly generated triangles and warp the image using that transform. This generates more complex shapes that retain some notion of their original form. Figure 5 shows a game board that has been distorted from its original in Figure 4.

Using these distorted images, we processed each image using the Canny Edge Detector (Canny, 1986) and used mathematical morphology to find closed contours of the objects, thereby segmenting the objects from each other. We acquired the boundary of the objects (always 15 of them), following the inner contours as identified by a border tracing algorithm (Suzuki and Abe, 1985). For each individual object we then extract the number of edges, RGB (red, green, blue) values, HSV (hue saturation value), and from the object's *moments*: its centroid, horizontal and vertical skewness (third order moments measuring the distortion in symmetry around the x and y axis), and the orientation value representing the direction of the principal axis (combination of second order moments). Taken together, this set of features represents a single object, which can be used for the word-object classifiers described earlier.

Procedure Using 1000 episodes, we evaluate our model across 10 folds (900 episodes for training, 100 for evaluation). Our baseline model is a generative model of RR that will be described below (random baseline is 7%). We also incorporate gaze and deixis by treating them as individual RR models and interpolating their distributions with the distribution given by the model. We ran the experiments twice, once with hand-transcribed utterances as basis for U, and once with automatic speech recogniser (ASR) output. The α weights (Equation 7) for hand-transcribed data were for speech, deixis and gaze: 0.72, 0.16, and 0.12 respectively, and for ASR 0.53, 0.23, 0.24, respectively (note that for ASR, more weight was given to the non-speech models).

[2]This approach also allows us to keep control over the degree of noisiness and systematically study its effect; this is something, however, that we leave for future work.

Task The task is RR, as described earlier. At each increment, the model returns a distribution over all objects; the probability for each object represents the strength of the belief that it is the referred one. The argmax of the distribution is chosen as the hypothesised referent.

A Stronger Baseline Model for Comparison To be able to judge the performance of our model better, we compare its results to a generative model we developed for the same domain, the *simple incremental update model* (SIUM) described in Kennington et al. (2013, 2014). SIUM is a good candidate to compare with this approach because it is also meant to work incrementally and it can accommodate uncertainty in the world representation. As a generative model SIUM learns the joint distribution between RE U and the world W, and it adds as a latent variable R, the properties of the object. This is formalised as follows (following (Kennington et al., 2013); see there for further details):

$$P(I|U,W) = \frac{1}{P(U)} P(I) \sum_{r \in R} P(U|R) P_W(R|I) \tag{8}$$

Here, $P_W(R|I)$ models the connection between objects and properties that are picked out for verbalisation. In the version described in Kennington et al. (2013, 2014), this model is read off the symbolic representation of the scene: the assumption is made that every property that a given object has is equally likely to be verbalised. This is where uncertainty about properties can be inserted into the model, which we do here. We trained an SVM classifier to classify the objects (pre-processed and segmented as described above) with respect to colors (e.g., classes like `red`, `blue`, etc.) and for shapes (e.g., `X`, `T`, etc.). The spatial placement of objects was determined by rules; the board was segmented into 4 quadrants and an object received a `left`/`right`, and `top`/`bottom` property with a probability of 1 if the object was in that corresponding area of the board. Where in Kennington et al. (2013, 2014) there was exact knowledge about the properties of objects, we now have distributions over properties, and the changed assumption now is that the likelihood of a property being verbalised is proportional to the strength of belief in it having this property. Other than that, SIUM remained unchanged.

4.2 Metrics for Evaluation

To give a picture of the overall performance of the model, we report *accuracy* (how often was the argmax the intended referent) after the full referring expression has been processed and *average rank* (position of intended referent among the 15 candidates on ordered distribution; ideal would be an average rank of 1, which would also correspond to 100% accuracy). Together, these metrics give an impression of how interpretable the full distribution is, beyond just the argmax. We report results for testing the model only given speech information (and no interpolation with the other models), and the other modalities added separately and jointly. We also computed results for SIUM given speech information.

We also look into how the model performs incrementally. For this, we followed previously used metrics (Schlangen et al., 2009; Kennington et al., 2013), where the predicted referent is compared to the gold refererent at each increment:

- **first correct**: how deep into the RE (%) does the model predict the referent for the first time?
- **first final**: if the final prediction is correct, how deep into the RE was it reached and not changed?
- **edit overhead**: how often did the model unnecessarily change its prediction (the only *necessary* prediction change happens when it first makes a correct prediction)?

4.3 Results

As Figures 6 and 7 show, our model performs well above the SIUM baseline, for all settings. (The *random selection baseline* sits at 7%.) We assume that it performs better not only because as a discriminative model it does not need to model the full joint distribution, but also because it directly learns a connection between words and visual features and does not need to go through a set of pre-determined features, which could be considered as a "lossy" compression of information. The Figures also show that using

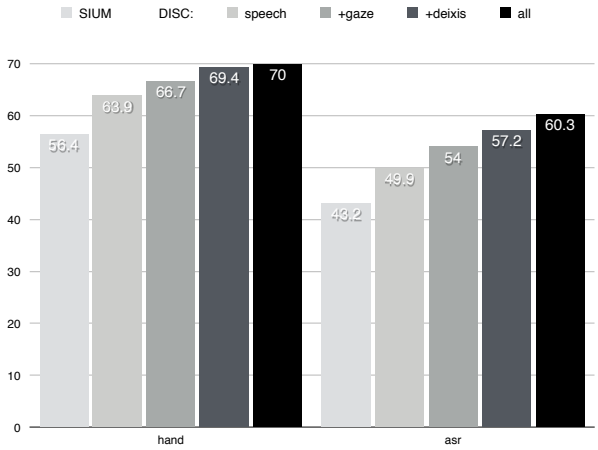

Figure 6: Results of our model in accuracies; higher numbers denote better results.

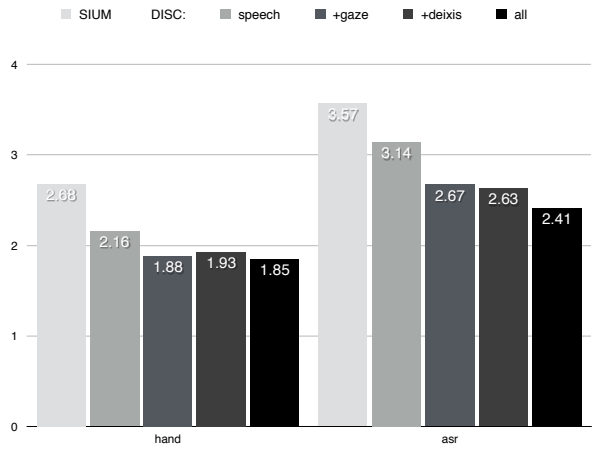

Figure 7: Results of our model in average rank; lower numbers denote better results.

ASR output does have an impact on the performance, as expeced. The speech+deixis models tend to work better than speech+gaze models in terms of accuracy; we speculate that this is due to the (naive) assumption implicit in our setup that participants gaze at the referred object most of the time, where in fact they often look at distractors, etc., making gaze a noisier model of predicting the referent. The story, however, is slightly different for the average rank: speech+deixis is slightly worse than speech+gaze at least for hand-transcribed data; this simply means that though speech+deixis gets the referred object into the argmax position more than speech+gaze, it doesn't always mean a better overall distribution. As expected, the best-scoring average rank is when all models are interpolated, both for hand-transcribed and ASR. Overall, there is about a 6% increase when both modalities are included when using hand-transcribed data. The increase when including both modalities is slightly larger (10%) with ASR. This nicely shows that when given noisier linguistic information, the model can partially recover by taking more benefit from interpolating with other information sources.

4.4 Incremental Results

Figure 8 gives an overview of the incremental performance of our model, compared to that of SIUM. (Results here are for the hand-transcribed utterances.) As the metrics talk about "% into expression", these metrics can of course only be computed when the eventual length of the expression is known, that is, after the fact. Moreover, to make these units comparable (as "10% into the utterance" is very different in terms of words for an utterance that consists of 2 words than for one that is 12 words long), we bin REs into the classes *short*, *normal*, *long* (1-6, 7-8, 9-14 words, respectively, as to group together REs that are of similar length).

Ideally for use downstream in a dialogue system, the reference resolver would make a first correct decision quickly, and this would also be the final decision (that is, in the graph the two bars would be close to each other, and at a low value). As the Figure shows, DISC is somewhat earlier than SIUM and on average that decision is very close to the final decision it makes, but it pays for this in a higher edit-overhead. When looking at first-final, SIUM is not that far away from DISC, but nevertheless, the new model beats the baseline across the board (and, as shown in Table 1, is also correct more often). DISC produces less stable predictions, which is presumably due to its response to the expression being a simple summation of the responses to its component words, whereas SIUM is a proper update model.

5 Further Analysis

We analysed several individual word classifiers to determine how well their predictions match assumptions about their lexical semantics. For example, the classifier for the word *links* (*left*) should yield a high

% edit overhead		
utt length	DISC	SIUM
1-6	11.5	3.8
7-8	19.76	17.2
9-14	41.0	27.5
% never correct		
utt length	DISC	SIUM
all lengths	19.5	32.0

Table 1: % edit overhead and never correct

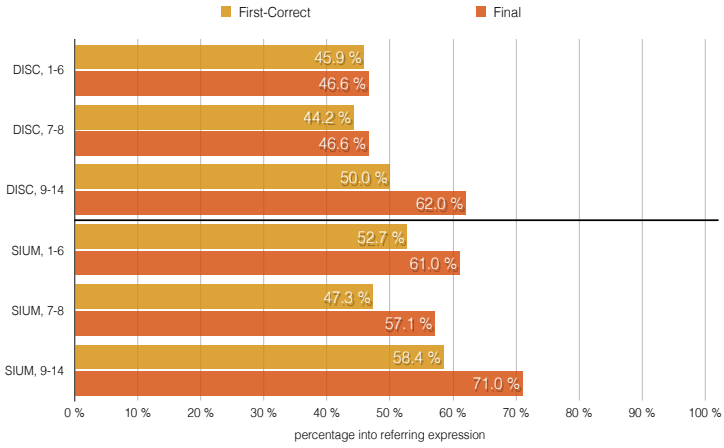

Figure 8: Incremental Performance

probability when given an object representation where the x-coordinate values are small (i.e., on the left of the screen), and lower probabilities for x values that are high. This was indeed the case, as shown in Figure 9. This is a nice feature of the model, as objects that are in the middle of the scene can still be described as *on the left*, albeit with a lower probability. We also tested how well classifiers were learned for colour words. In Figure 10 we show how changing the H S V features (representing colors) across the spectrum, keeping all other object features stable, yielded different responses from the classifier for the word *gelb* (yellow), where the y-axis on the figure represents probability for that particular colour value.[3]

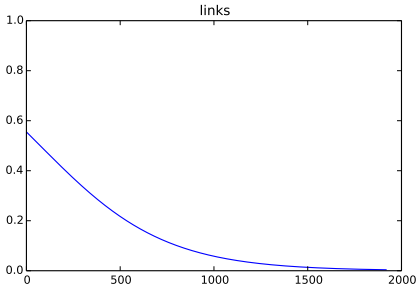

Figure 9: Strength of word *links* (German for *left*) predicting when given different x-coordinate values, the y-axis represents the probability of that point being *left*.

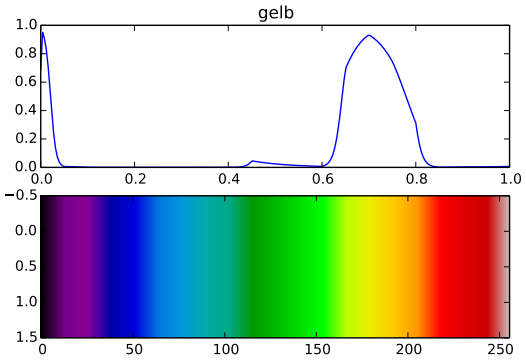

Figure 10: Strength of word *gelb* (German for *yellow*) predicting when given different color (HSV) values.

We further looked into shape words. Figure 11 shows the response of the classifier for *kreuz* (cross) when given object representations where only the shape-related features (number of edges, skewness) were varied across all possible shapes (the x-axis uses here the standard labeling of pentomino pieces with letters whose shapes are similar). Interestingly, the classifier generalised the word to apply not only to objects with the cross shape, but also the Z-shape piece (the red piece in the bottom of the top right group in Figure 4) and others which also intuitively seem to be more similar. For a sanity check, we looked at the responses to change in colour for the word *kreuz*. As Figure 12 shows, this classifier does not pick out any specific color, as it should be. This shows that the word classifier managed to identify those features that are relevant for its core meaning, ignoring the others.

[3]There is also a high probability in the black region. It could be the case that the yellow classifier learned that a low value for B is highly discriminative (black is 0 for all RGB values).

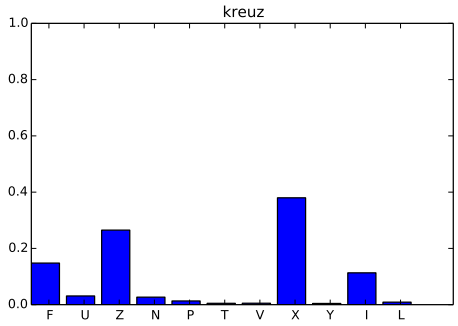

Figure 11: Strength of word *kreuz* (German for *cross*) predicting when given different values of number of edges.

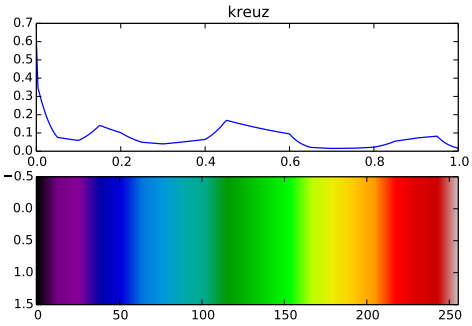

Figure 12: Strength of word *kreuz* (German for *cross*) predicting when given different color (RGB) values.

6 Conclusion

We presented a model of reference resolution that learns perceptually-grounded semantics from relatively low-level, computer vision-based features, rather than from pre-defined sets of object properties or ontologies. We tested the model and found that it worked well for the Pentomino corpus, despite having a very simple notion of compositionality. The model is discriminative and outperforms a generative approach applied the same data. The model fused well with additional modalities, namely gaze and deixis, providing improved results in a reference resolution task. Perhaps best of all, the model is simple: besides the scenes and referring expressions, one only needs to know what object was referred in each scene in order to train the model, which is generally easy to annotate.

For future work, we will apply more principled methods of compositionality. We also plan to apply the model to a more systematic test of how well it performs under varied strengths of image distortion. We further plan on applying the model in a real-time multimodal learning scenario, using video images of real objects.

Acknowledgements We would like to thank the anonymous reviewers for their comments. We also want to thank Spyros Kousidis for his help with the data collection.

References

Canny, J. (1986, June). A computational approach to edge detection. *IEEE transactions on pattern analysis and machine intelligence 8*(6), 679–698.

Engonopoulos, N., M. Villalba, I. Titov, and A. Koller (2013). Predicting the resolution of referring expressions from user behavior. In *Proceedings of EMLNP*, Seattle, Washington, USA, pp. 1354–1359. Association for Computational Linguistics.

Funakoshi, K., M. Nakano, T. Tokunaga, and R. Iida (2012, July). A Unified Probabilistic Approach to Referring Expressions. In *Proceedings of SIGdial*, Seoul, South Korea, pp. 237–246. Association for Computational Linguistics.

Gorniak, P. and D. Roy (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research 21*, 429–470.

Griffin, Z. M. and K. Bock (2000). What the eyes say about speaking. *Psychological science : a journal of the American Psychological Society / APS 11*, 274–279.

Iida, R., M. Yasuhara, and T. Tokunaga (2011). Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues. In *Proceedings of IJCNLP*, Number 2003, pp. 84–92.

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

Kennington, C., R. Iida, T. Tokunaga, and D. Schlangen (2015). Incrementally Tracking Reference in Human/Human Dialogue Using Linguistic and Extra-Linguistic Information. In *Proceedings of NAACL*, Denver, U.S.A. Association for Computational Linguistics.

Kennington, C., S. Kousidis, and D. Schlangen (2013). Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *Proceedings of SIGdial 2013*.

Kennington, C., S. Kousidis, and D. Schlangen (2014). Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model. In *Proceedings of CoLing 2014*.

Kennington, C. and D. Schlangen (2012). Markov Logic Networks for Situated Incremental Natural Language Understanding. In *Proceedings of SIGdial*, Seoul, South Korea, pp. 314–322. Association for Computational Linguistics.

Kousidis, S., C. Kennington, and D. Schlangen (2013). Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *Proceedings of SIGdial 2013*.

Kousidis, S., T. Pfeiffer, Z. Malisz, P. Wagner, and D. Schlangen (2012). Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH 2012 Satellite Workshop*, pp. 39–42.

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Matuszek, C., L. Bo, L. Zettlemoyer, and D. Fox (2014). Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Proceedings of AAAI Conference on Artificial Intelligence*. AAAI Press.

Ng A.Y. & Jordan M. I. (2002, December). On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems. . In *Machine Learning*, Vancouver, Canada.

Poesio, M. and R. Vieira (1997, June). A Corpus-Based Investigation of Definite Description Use. *Comput. Linguist. 24*(2), 47.

Richardson, M. and P. Domingos (2006). Markov logic networks. *Machine Learning 62*(1-2), 107–136.

Schlangen, D., T. Baumann, and M. Atterer (2009). Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of SIGdial*, London, UK, pp. 30–37. Association for Computational Linguistics.

Spivey, M. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology 45*(4), 447–481.

Steels, L. and T. 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.

Suzuki, S. and K. Abe (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing 29*(30), 396.

Tanenhaus, M. K. and M. J. Spivey-Knowlton (1995). Integration of visual and linguistic information in spoken language comprehension. *Science 268*, 1632.