CS671: Natural Language Processing

Homework 3 - Paper Review

Name: PRABUDDHA CHAKRABORTY
Roll Number: 15111027

1st year, M.Tech (Computer Science and Engineering)
IIT Kanpur

Paper under discussion

Title:

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

Authors:

- Casey Kennington
- David Schlangen

Aim of the paper

To present a model of reference resolution

- that learns a perceptually-grounded meaning of
  - words,
  - including relational words.

Given: Image + Referring Expr ---Find the OBJ in the Image

Aim of the paper (continued)

The Word and Phrase Meaning presented is –

- Perceptually Grounded
  - Links Words and Visual Features
- Trainable
  - That Link learned from Examples of Language Use
- Compositional
  - Meaning of phrase = f (Parts of the phrase)
- Dialogue Plausible
  - Incremental Computation
  - Works for Noisy Input

Model Presented in this paper

W = Set of Objects in the World

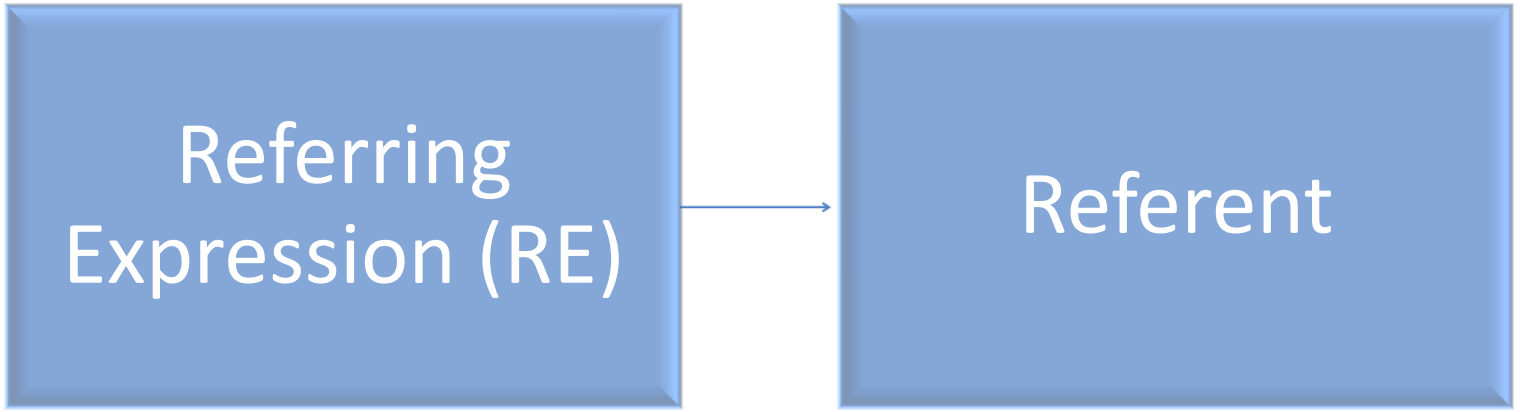
U = Referring Expression Set

Types of Composition

- Simple References
  - picking out properties of single objects
  - (e.g., “green” in “the green book”) [1]

- Relational References
  - Picking out relations of two objects
  - “The green book to the left of the mug.” [1]

Reference Resolution



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

- I^* -> Object which is Intended Referent
- U -> A Representation of RE
- W -> Representation of the Relevant World

Representing f_{rr}

- Stochastic Model for f_{rr}

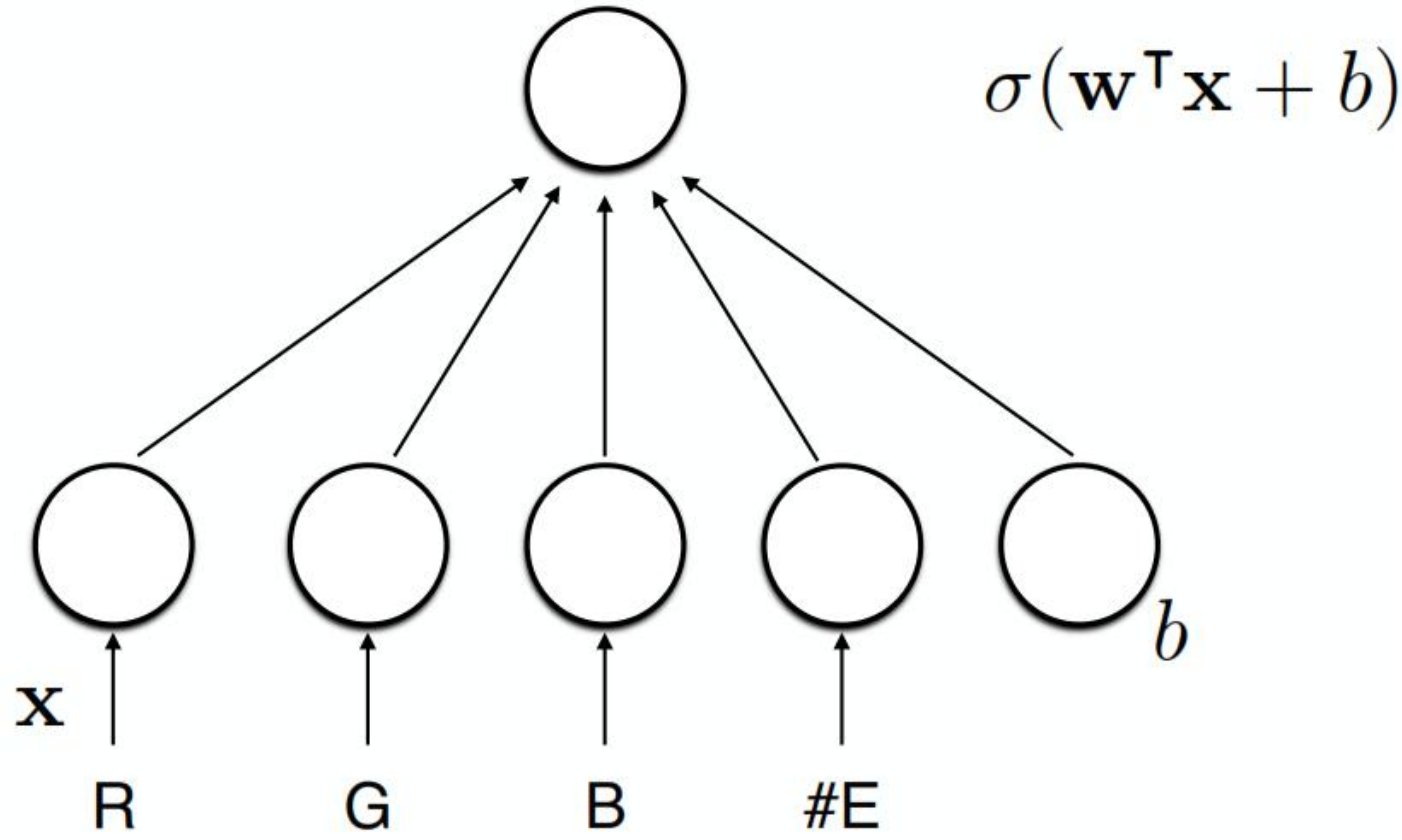
$$I^* = \text{argmax } P(I \mid U,W)$$

P -> Probability assigned to each element ‘I’ represents strength of belief.

We calculate how likely each object is linked with the given U and W. Then take the max.

Classifier for one Word in the corpus

[2]



The ouput is P (obj|w).

Calculation of P

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

- x -> Representation of the object by visual feature.
- w -> Weight vector that us learned.
- ώ -> One word from the training corpus
- P_w (x) -> Probability that 'x' is a good fit with 'ώ'.
- b -> bias
- σ -> Logistic Function

Training Algorithm for our Model

For each **RE** in the Corpus
{
    For each **Word** in RE
    {
        Pair 'Word' with features of the object referred to by the RE; //Adds to positive Example

        Pair word with features of 'n' objects not referred to by RE; //Add to negative examples
    }
}
For each **Word** in Vocabulary
{
    Train word classifier.
}

Simple References

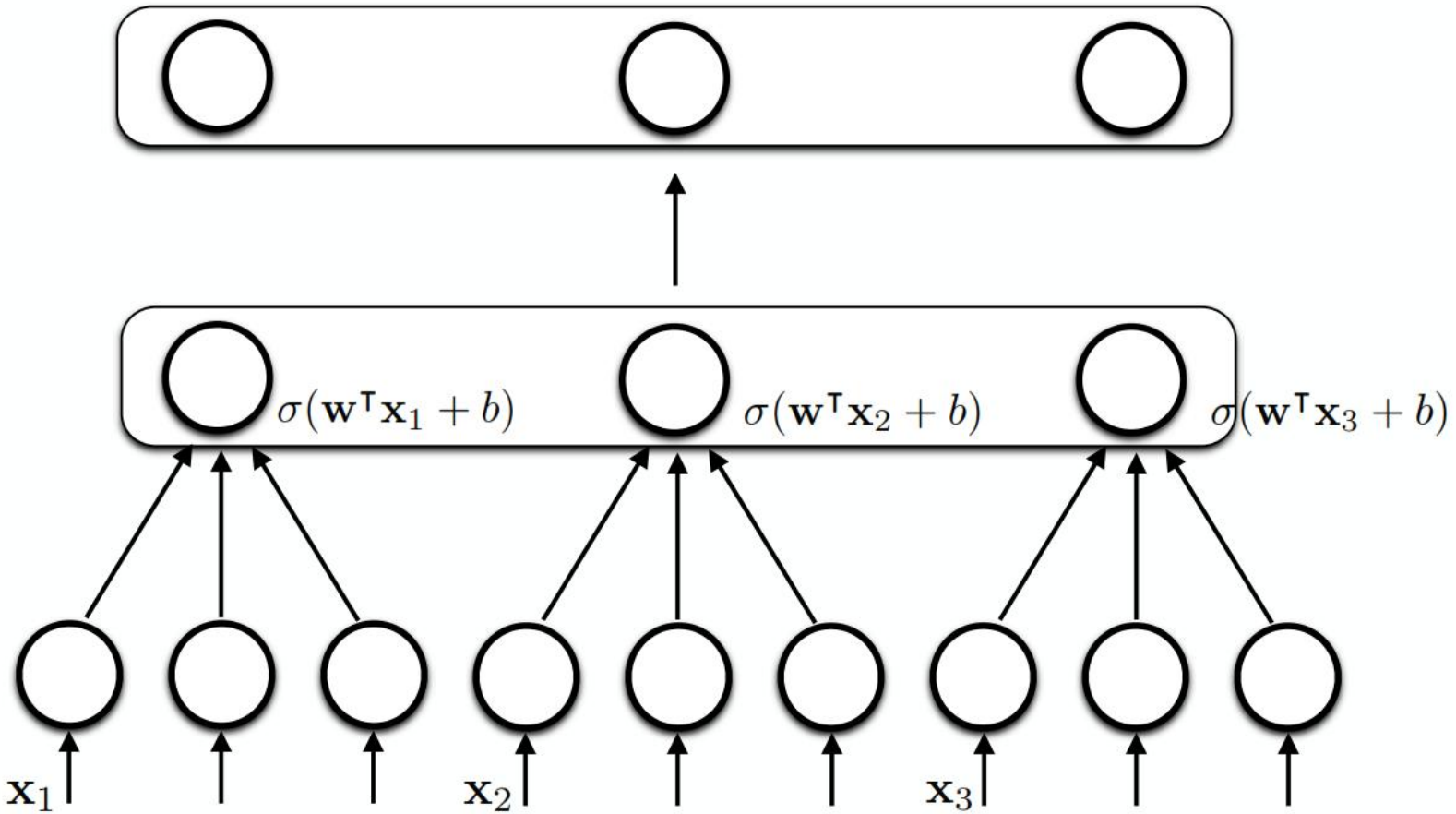
We have seen how to get the probability for one Object for a given word. So now we shall compute the probabilities for all the objects for a given word say w.

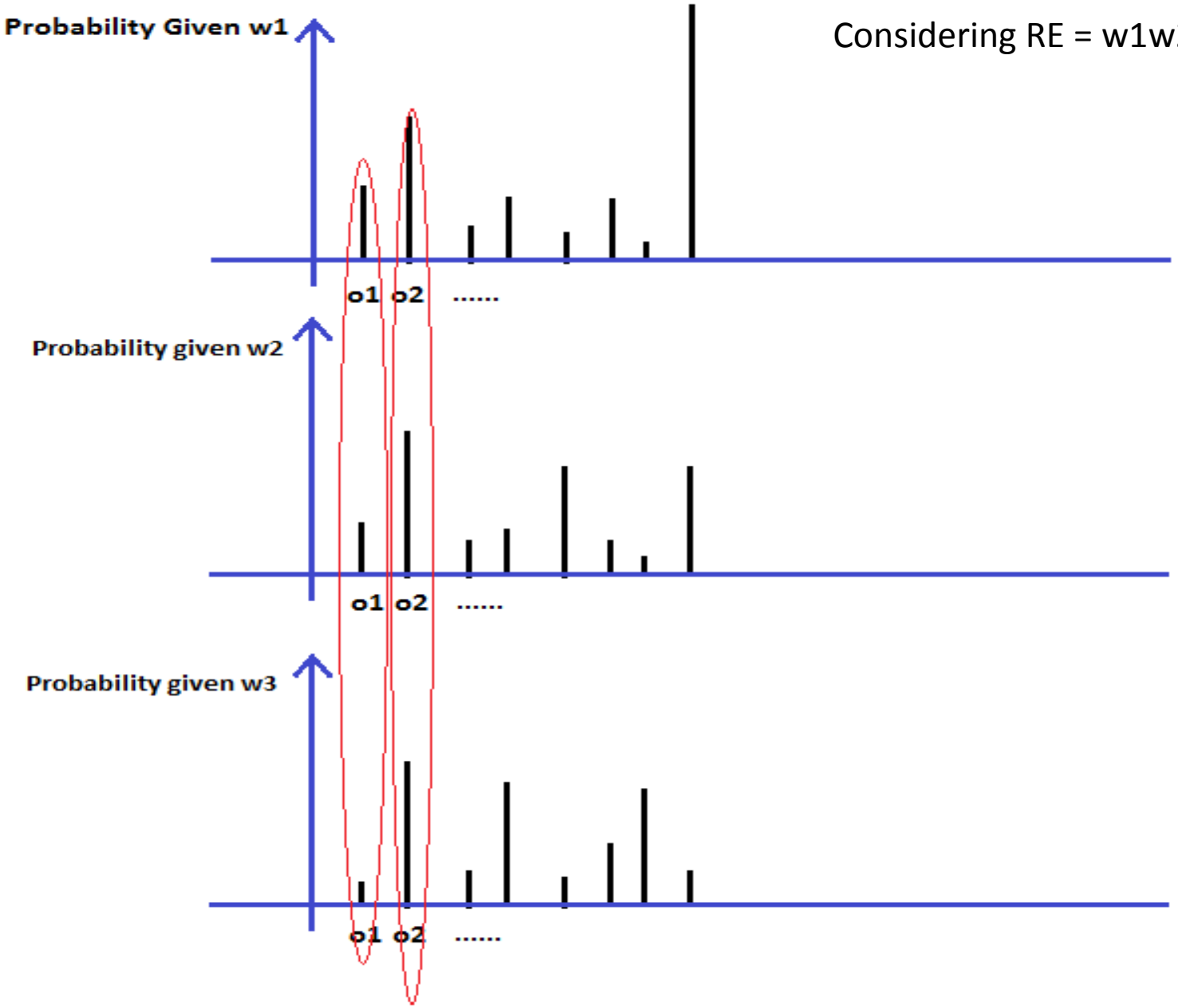
$$[\![w]\!]^W_{obj} =$$
$$normalize(([\![w]\!]_{obj}(\mathbf{x}_1), \ldots, [\![w]\!]_{obj}(\mathbf{x}_k))) =$$
$$normalize((p_w(\mathbf{x}_1), \ldots, p_w(\mathbf{x}_k))) = P(I|w)$$

To compose the evidence from individual words, into a prediction for a 'simple' RE. We average the contributions of its constituent words.

$$[\![[_{sr} w_1, \ldots, w_k]]\!]^W = [\![sr]\!]^W [\![w_1, \ldots, w_k]\!]^W =$$
$$avg([\![w_1]\!]^W, \ldots, [\![w_k]\!]^W)$$

Normalization Layer





Considering RE = w1w2w3

I* = oi , if AVG.oi = MAX (AVG.o1, AVG.o2, ……., AVG.on)

Relational References

It is a relation between

- a(simple) reference to a landmark
- and a (simple) reference to a target.

Example—The green book to the left of the mug.

This structure is indicated abstractly in the following 'parse':

$$[_{rel}[_{sr}w_1, \ldots, w_k][_r r_1, \ldots, r_n][_{sr}w'_1, \ldots, w'_m]]$$

Relational References (Continued)

- We generally contract expressions such as “to the left of” into a single token
  - And learn one classifier for it.
  - Expressing a judgement for a pair of objects.

- We apply the classifier to all Pairs
  - And normalize

$$[\![r]\!]^W = P(R_1, R_2|r)$$

Relational References (Continued)

- The belief in an object being the intended referent should combine the **evidences** from—
  - Simple reference to the landmark object
  - simple reference to the target object
  - The relation between the target object and the landmark object.

- Instead of averaging as for SR,
  - We Combine the Evidences multiplicatively.

The Combination

$$P(R_1|w_1,\ldots,w_k,r,w'_1,\ldots,w'_m) =$$

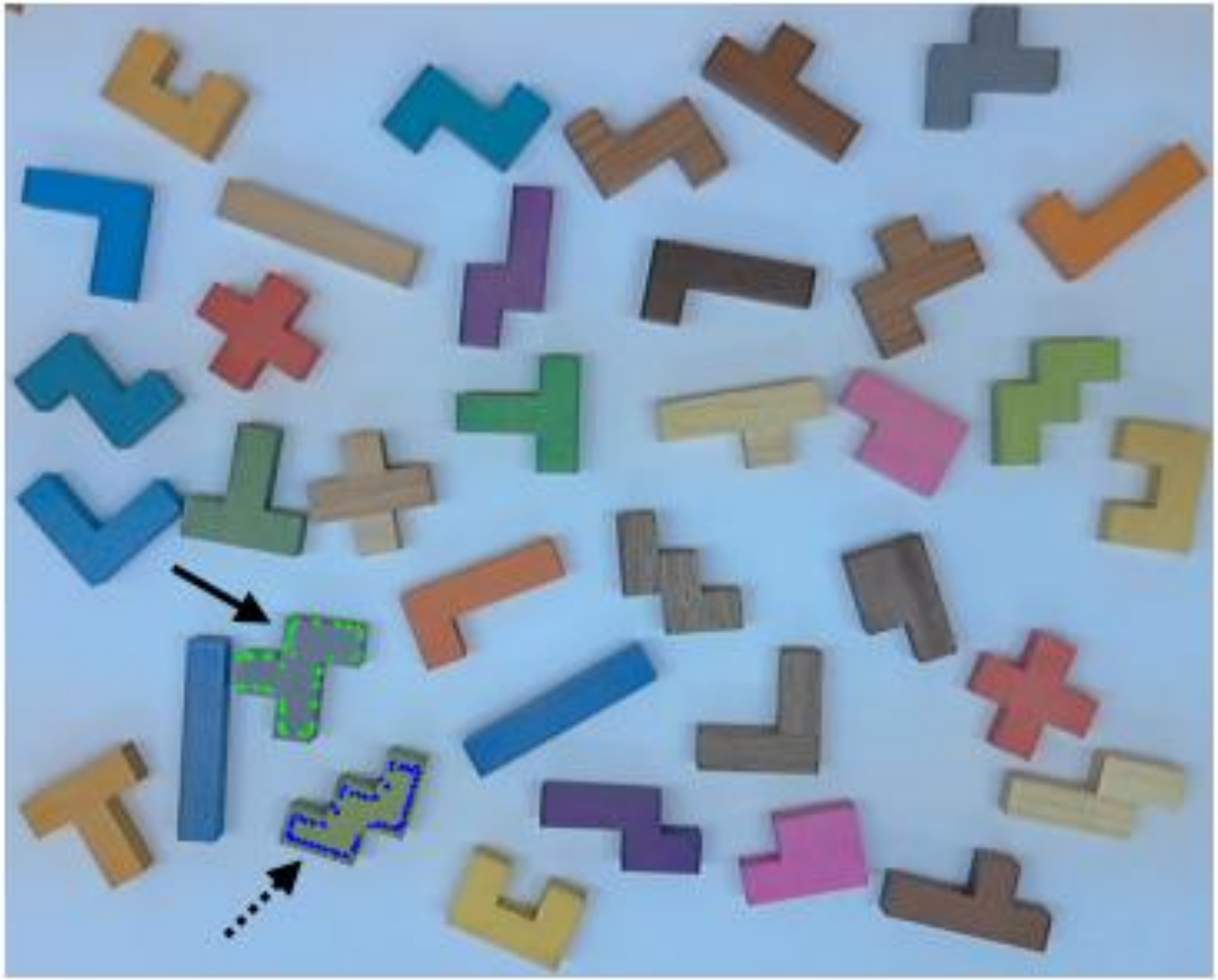
$$\sum_{R_2}\sum_{I_l}\sum_{I_t} P(R_1,R_2|r) * P(I_l|w'_1,\ldots,w'_m) *$$

$$P(I_t|w_1,\ldots,w_k) * P(R_1|I_t) * P(R_2|I_l)$$

Experiment

- 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

Experiment (Continued)



Experiment (Phase-1)

- ES (Experiment Software) Choose one Obj.
- Participant Refers to the “Target Obj” using only speech.
- Wizard (Experimentor) Sees the Table full of Obj and hears “Participant’s” RE and identifies the object.
- If Wrong-> Flag
- Repeat

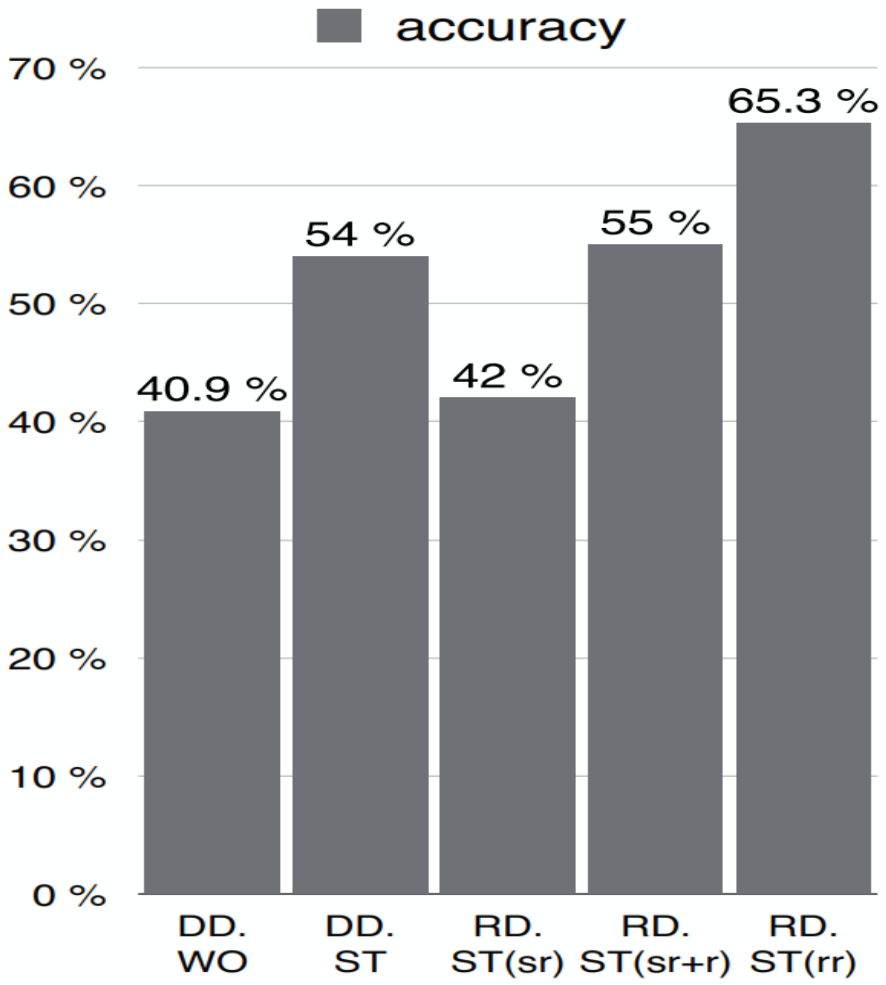
Experiment (Phase-2)

- ES Picks 2 Obj
- Participant Refers to the “Target Obj” using “Landmark obj” with only speech.
- Wizard (Experimentor) Sees the Table full of Obj and hears “Participant’s” RE and identifies the target object.
- If Wrong-> Flag
- Repeat

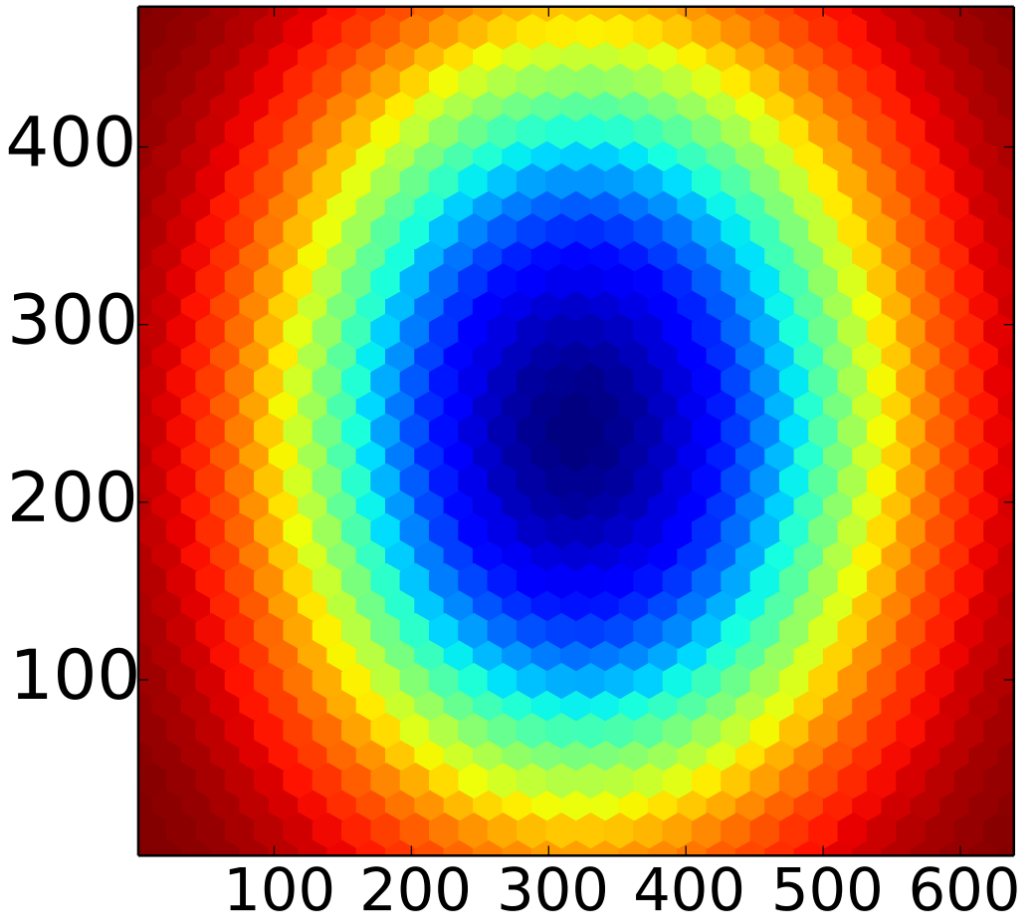
Experiment Result

- DD->Episodes where Target Referred directly via “Simple Reference”
- RD-> Episodes where Target Referred by “Landmark”
- ST-> Structure Knowledge (simple or relational reference)
- WO-> Only Word (NON ST)

Experiment Result (Continued)

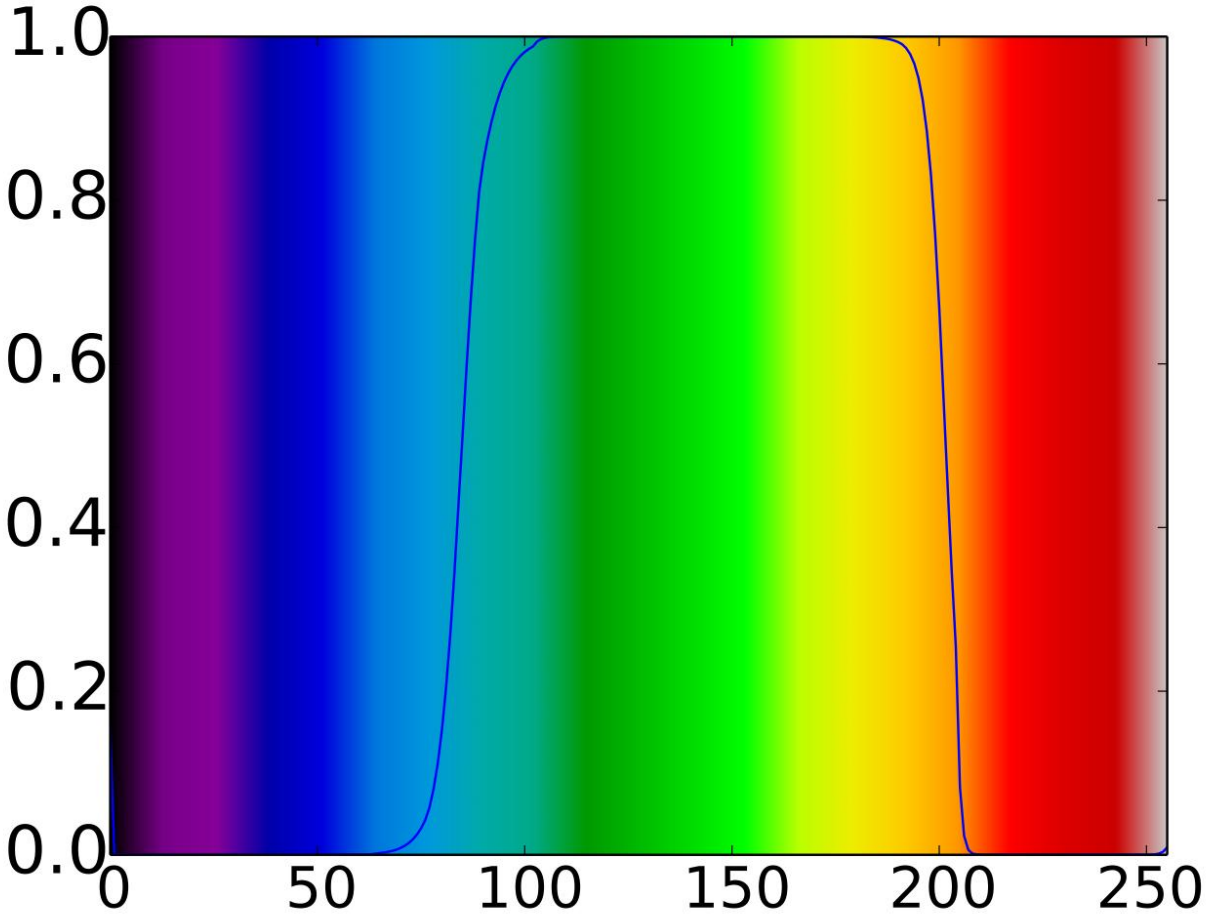


Experiment Result (Continued)



Cornor -> Higher chance of Success

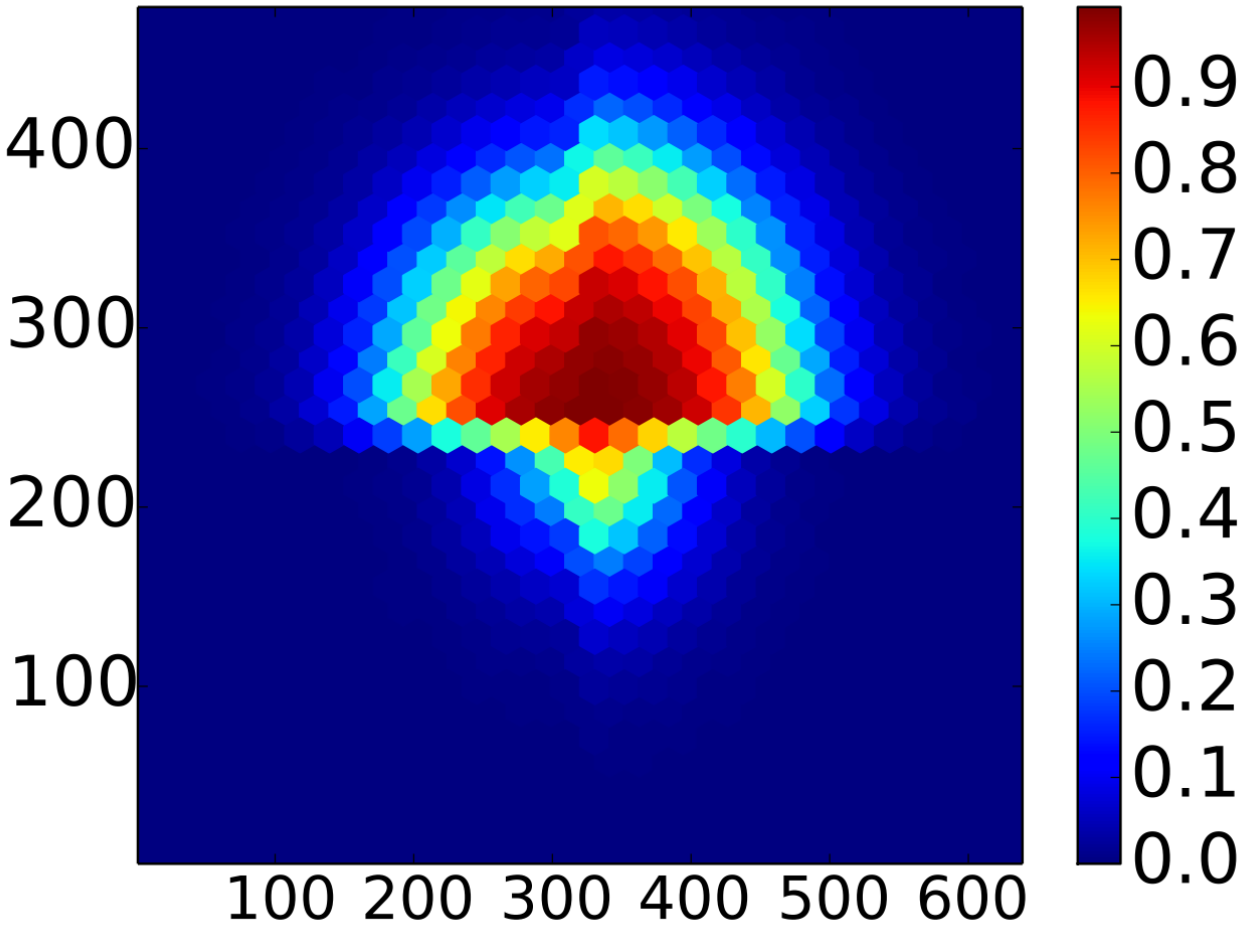
Experiment Result (Continued)



Green Objects -> Higher chance of Success

Experiment Result (Continued)

Target nearer to Landmark set in
Middle-> Higher chance of Success



Conclusion

- The model is
  - simple,
  - compositional,
  - and robust
- Despite
  - low amounts of training data
  - and noisy modalities.
- 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, but with perceptual groundings, is left for future work.

References

- [1] Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. Casey Kennington & David Schlangen
- [2] A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. Casey Kennington, Livia Dia, David Schlangen

Thank you !