

Semantics for anaphora resolution in a microworld using visual inputs

Manas Agarwal

`manas@cse.iitk.ac.in`

Under the guidance of Dr. Amitabha Mukerjee

September 30, 2009

1 Motivation

Anaphora resolution which most commonly appears as pronoun resolution is the problem of resolving references to earlier or later items in the discourse. These items are usually noun phrases representing objects in the real world called referents but can also be verb phrases, whole sentences or paragraphs. Anaphora resolution has been an area of active research in the realm of Natural Language Processing and is classically recognized as a very difficult problem. Most of the work is focussed on pronoun and noun-phrase resolution. The traditional resolution techniques for these type of problems are based on:

- Matching certain attributes or constraints. These may include gender, plurality or semantic consistency. [2]
- Giving weighted preference to various factors such as proximity, centre of attention [1] or syntax.

The above however cannot resolve many types of ambiguous referents. For example, There are two balls on the table. Ball A is at rest and Ball B is moving towards A. It moved past another. Here in the last sentence, whether it refers to A or B, this cannot be resolved by present anaphora resolution methods. But what if we have been provided a video of this scene? By the movements of the balls, we can relate the pronoun with the correct ball. Thus, this project aims to resolve such type of anaphora by looking into a simple 2D video (preferably 2-body interactions) and a describing commentary on it. This scenario is kept in view of a more extended practical example: given a soccer match video from top and its commentary. But we aim to start from a much elementary examples first.

2 Past Work

Work done by [3] provides an insight to deriving linguistic structures from visual inputs. It claims that semantics of certain actions may be learned prior to language, that is, we can form a crude linguistic description of a scene just by the visual inputs only. These actions were prominently change of positions of agents in a multi-agent setups, and the result were semantics describing the relative motions amongst them. They also incorporated centre of attention feature since they claim that a commentator is more likely to talk about things that are in attentive focus. The results of the work also supported their claims. By restricting the constituents participating in an action using computational model of visual attention, they have achieved better correlation with the commentary. Since, my work also aims to acquire some linguistic constructs involving primarily motion in a visual sequence, this work is important.

3 Methodology

- First we have to define our microworld. We are considering a 2D rectangular region filled with circular balls. The microworld will only allow motions of balls (moving in the plane, mutual and wall collisions) within the region. We will start from the interaction of only two balls in the beginning. We would extend to multi-ball sequence depending upon the results.
- The feature vectors for different frame would consist of combination of position and velocity vectors of each ball.
- Using these feature vectors, we would like to categorize the relationship between different balls as moving towards each other, moving away(both or one-static),one approaching another(other also moving or static), moving left or right or up or down. In the above mentioned paper, Merge Neural Gas Algorithm is used. A study of other algorithms will also be done to decide the better one suited for this problem.
- A study of the effect of attentive focus will also be done.[4]
- Commentary for the video sequence will be taken. The sentence corresponding to correct action will be determined. This can be done by identifying verb in the sentence corresponding to the action schema that has been determined through the algorithm. The agents in the sentence then can easily be identified, hence resolving any pronouns occurring in it. If verb cannot be identified with acceptable accuracy, then the most attentive action taking place can be assumed.
- To find the accuracy and amount of correct correlation and finding ways to improve it.

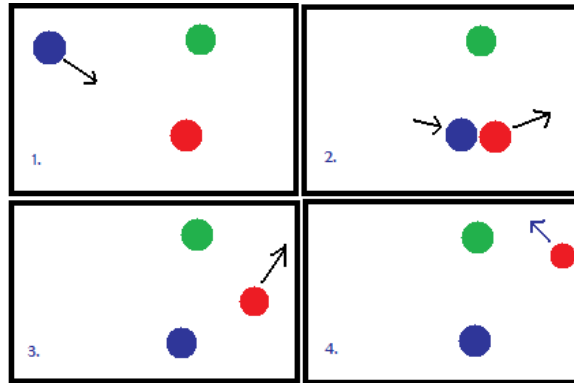


Figure 1: Frames of a microworld video sequence. Arrows shows the velocity vector direction

4 Objective

- To find an acceptable algorithm for finding action schema.
- To resolve such anaphora to an acceptable extent, and study the challenges posed by the problem.
- To study the relationship of accuracy versus number of agents in the microworld.

5 Example

Figure 1 is an example of a video of a 3 ball system. Given is the following commentary: *There is one red, green and blue ball. The blue ball is moving. The red ball is approached by it. They both collides. Now, it is moving away. It collides the wall. It is turning left, towards green ball.*

- In sentence 3, *it* can only be identified as the blue ball, since it is moving towards red (ie, chasing it). Thus, if algorithm clusters this action schema to moving towards each other (one static), then this can be determined.
- In sentence 4, *They* is again taken as bot red and blue balls, as they are in state of contact.
- In sentence 5, *It* is the red ball, since its motion is approaching walls.
- In sentence 6, *It* is the red ball, since its moving owards green ball

In all the above sentences, if our algorithm correctly clusters the action schema, then the resolution is unambiguous.

References

- [1] B. J. GROSZ, A. K. J., AND WEINSTEIN, S. Centering: A framework for modeling the local co-herence of discourse. *Computational Linguistics* 21(2) (1995), 202–225.
- [2] HOBBS, J. R. Resolving pronoun references. *Lingua* 44 (1978), 311–338.
- [3] SATISH, G., AND MUKERJEE, A. Acquiring linguistic argument structure from multimodal input using attentive focus. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on* (2008), pp. 43–48.
- [4] SINGH, V. K., M. S., AND MUKERJEE, A.