Discovering Coreference Using Image-Grounded Verb Models

Amitabha Mukerjee
Dept. of CS, IITK, India
amit@cse.iitk.ac.in

Kruti Neema
Dept. of CS, IITK, India
krutineema@gmail.com

Sushobhan Nayak
Dept. of EE, IITK, India
snayak@iitk.ac.in

Abstract

Breaking away from traditional attempts at coreference resolution from discourse-only inputs, we try to do the same by constructing rich verb semantics from perceptual data, viz. a 2-D video. Using a bottom-up dynamic attention model and relative-motion-features between agents in the video, transitive verbs, their argument ordering etc. are learned through association with co-occurring adult commentary. This leads to learning of synonymous NP phrases as well as anaphora such as "it","each other" etc. This preliminary demonstration argues for a new approach to developmental NLP, with multi-modal semantics as the basis for computational language learning.

1 Introduction

It is common in discourse to refer to the same object using many phrases. For example, in a shared scene with two square shapes (Figure 1), the larger square may be called "the big box", "the square" or by anaphoric references such as "it", "itself", etc. Resolving the many types of co-reference remains a challenging problem in NLP (Stoyanov et al.(2009)). There are increasing calls for mechanisms with direct semantic interpretation, learned from multimodal input (Roy and Reiter(2005)). This work is posited along such lines; it does not attempt to resolve coreferences, but merely to illustrate how knowledge relating verb argument structure to the visual action schemas may be learned from multi-modal input. The possibility hinted at is that such grounds-up learning driven NLP systems may eventually have a rich enough library of syntacto-semantic structure to handle coreference more fully. Present attempts at analyzing multimodal interfaces (Fang et al.(2009); Steels(2003)) aim to identify the referents in interaction discourse, whereas our objectives are to build a system that can learn the principles of coreference, particularly anaphora. Furthermore, models that consider actions often use prior knowledge for visual parsing of actions (Dominey and Boucher(2005)). With reference to the work on resolving coreference problems, such models typically encode considerable structural knowledge of the linguistic and visual domains. Our work proposes mechanisms whereby these structures may be learned.

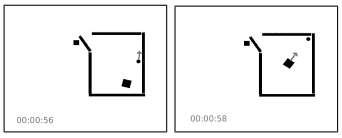


Figure 1: *Multimodal input: 2D video "Chase"*: Three shapes, [big-square], [small-square] and [circle] interact playfully (velocities shown with aroows).

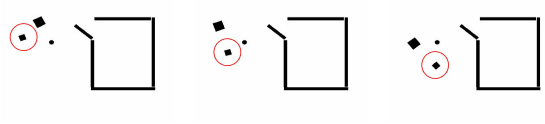


Figure 2: Computed bottom-up attention during the part of the action [chase(big-square,small-square)].

2 Learning Action Models and Argument Structure

Here we consider how an unsupervised process may acquire action structures from simple videos by clustering frequently observed sequences of motions. The perceptual database in the present

Proceedings of Recent Advances in Natural Language Processing, pages 610–615,
Hissar, Bulgaria, 12-14 September 2011.

model is a single 2-D video (from Heider and Simmel(1944)) (Figure 1). Here, the referent objects (a big square, a small square and a circle) are moving around, interacting with each other, and are easily segmented, as opposed to static referents in game-like contexts used in other multimodal co-reference analysis (Fang et al.(2009)). This presents a mechanism for learning events, which is extremely difficult in general contexts. The linguistic database consists of a co-occuring narrative with 36 descriptions of the video. In the 13 from the original Stanford corpus asked the subjects to discriminate actions in a fine and coarse manner. The subsequent 23 collected by us, also from student subjects, were completely unconstratined. Thus, these narratives exhibit a wide range of linguistic variation both in focus (perspective) and on lexical and construction choice.

We consider two-agent spatial interactions, which correspond to verbs with two arguments. The model uses bottom-up dynamic attention (Figure 2) to identify the objects that are related by attention switches (Satish and Mukerjee(2008)). The system considers pairs of objects attended to within a short timespan, and computes two inner-product features a) *pos·velDiff* $[(\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B - \vec{v}_A)]$ and b) *pos·velSum* $[(\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B + \vec{v}_A)]$. The temporal histories of these feature vectors are then clustered using the temporal mining algorithm Merge Neural Gas (Strickert and Hammer(2005)). Four action clusters are discovered, two of which correspond to [come-closer] and [move-away], and two correspond to [chase](Figure 3). Chase has two clusters because it is asymmetric, and the primary attention may be on the chaser (cluster 3) or on the chased (cluster 4). By computing the feature vectors with the referents switched, the system can by itself determine this alternation.

These learned models or *visual schemas* are acquired prior to language, and defined on the perceptual space. The learned models include the agents participating in the action, which constitutes the visual arguments of the action. They will next be related to the linguistic input.

Associating with textual phrases: Next, when our computational learner encounters language, it associates perceptual objects under attention to linguistic units in the co-occurring utterances. For this, it first considers those sentences which overlap temporally with the period when the action clusters are active, using an approach similar to (Roy and Reiter(2005)). One can now align sentences with objects in attentive focus to identify the names of objects (nouns) (Yu and Ballard(2004)). At this point, we assume that the learner knows these nouns, which are not considered as labels for verbs. Extremely frequent words (e.g. the, an, etc) are also dropped from consideration for mapping to actions. Using 1-, 2- and 3-word sequences from the text, the strongest associations for the action clusters are shown in Figure 4, and we see that clusters 1 [come-closer] and 2 [move away] have strongest associations with "move toward each" and "move away", but these are not very dominant over other competitors. On the other hand, for clusters 3 and 4 [chase], there is a strong association with the word "chase".

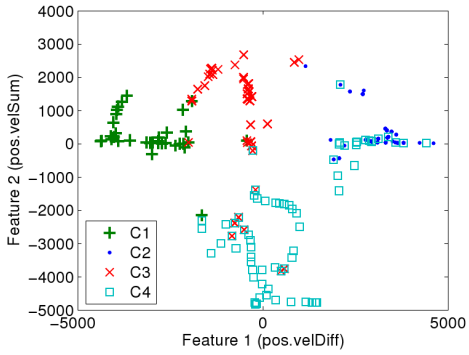


Figure 3: *Feature Vectors of the Four Clusters* : CC: C_1, MA: C_2, Chase(focus is on [chaser]): C_3, Chase(focus is on [chased]): C_4; The clusters reflect the spatio-temporal proximity of the vectors.

Next, it associates sentences uttered during the cognitive focus and correlates them with these actions. The strongest associations are learned as labels for actions (verbs) (Satish and Mukerjee(2008)).

Linguistic Constructions and Argument Structure mapping: At this stage the system knows the most preferred names for the participants (e.g. "big square"), as well as the label for the action (e.g. "chase"). Among the utterances co-occurrent with the action, it now computes the probability of different orderings for the units (e.g. the ordering of "chase"+grammatical-particle, [chased] and [chaser]). Here [chased], [chaser] are used by us for clarity - the system knows these based as a trajector-object distinction, in terms of visual focus. For cluster C3, the pattern *[chaser] chas* [chased]* dominates with frequency 0.90,

CLUSTER 1 (Come-Close)		CLUSTER 2 (Move-Away)		CLUSTER 3 (Chase)		CLUSTER 4 (Chase)	
ONE WORD LONG LINGUISTIC LABELS(MONOGRAMS)							
corner	0.077	away	0.069	chase	0.671	chase	0.429
move	0.055	move	0.055	other	0.185	after	0.112
attack	0.042	chase	0.049	around	0.183	out	0.033
TWO WORD LONG LINGUISTIC LABELS(BIGRAMS)							
each other	0.086	move away	0.111	chase around	0.306	chase after	0.218
move toward	0.065	go into	0.035	each other	0.227	just chase	0.060
toward each	0.065	into with	0.035	chase each	0.198	chase out	0.058
THREE WORD LONG LINGUISTIC LABELS(TRIGRAMS)							
move toward each	0.182	go into with	0.099	chase each other	0.558	just chase out	0.142
toward each other	0.182	run away out	0.051	start run away	0.132	run away out	0.047
move close together	0.114	scare in corner	0.032	begin to move	0.127	to go after	0.031

Figure 4: Figure showing the strongest association of linguistic lebels and action clusters. Dominant association of [chase] with the word "chase" is evident.

and in C4 its frequency is 0.84. This construction matches sentences such as "The square chased the circle" or "The big square was chasing them". In a minority of cases, it also notes the construction *[chased] chase+particle by [chaser]*. Thus, it determines that with high probability, the construction for the action [chase] in English is *[chaser] chase+particle [chased]*. We assume our computational learner has this level of competence (the input to Algorithm 1) before it attempts to detect substituted arguments and missing arguments in linguistic structures. Now we are ready to address the question of coreference.

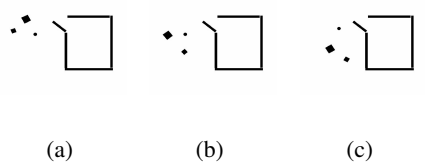
3 Synonyms and Anaphora

We propose a plausible approach towards discovering anaphora-mappings in Algorithm 1. For discovering synonymy, the model needs only to relate participants in known events, such as [chase], with the phrases it observes in the sentence before and after the word "chase" (Steps 1 and 2 of the algorithm). Whie attempting to discover synonyms and named entities of the discourse, the system discovers referentially stable mappings for fixed, single referents. But it also discovers several other units whose referents are dynamically determined by the recent discourse. This may be considered as a semantically-driven approach for discovering grammatical structures like 'the word order of arguments', and 'the phenomenon anaphora'.

Pronominal Anaphora ("it"): In Fig. 5, computing the relative motion features between the two objects in attentive focus (Fig. 2, the big square ([BS]) and the small square ([SS]) the learner finds that the motion sequence matches the visual schema for the action [chase], and given the order of the objects in the feature computation, one can say that the visual schema encodes the semantics of the predicate *chase([BS], [SS])*. Note however, that we do not explicitly use any predicates or logical structures; these are implicit in the visual schema. However, we remove some of the topmost frequent words "the" in this analysis where they appear as part of a phrase. If the entire phrase is a common word (e.g. "it", "they"), it is retained.

We now consider several sentences coteemporaneous with the scene of Fig. 5. For example in *large square chases little square*, when we match the arguments with the linguistic construction, we can associate "large square" with [BS] and "little square" with [SS]. Now, "big square" and "little square" are already known as labels for [BS] and [SS], so "large square" is associated with [BS] as a possible synonym map.

Another sentence aligned with the same action, *it is chasing the small box* results in the associations "it":[BS], and "small box":[SS]. Similarly, in *chases little block*, there is no referent at all for [BS], and "little block" is identified as a possible synonym for [SS].



(a) (b) (c)

Figure 5: Frame sequence in video showing predicate *chase(BS,SS)*.Corresponding narrations include *large square chases little square*, *it is chasing the small box* and *chases little square*.

Estimating Probabilities for Action Maps: In obtaining frequency estimates for synonyms, we require these phrases to co-occcur with instances where a known verb appears. However, even with 36 parallel narratives, the perspectival variation among speakers is such that quite often the same scene will not be focused on, and even where it is, completely unknown phrases may be used (e.g. "tries to get" for "chases"). Thus, one is not able to label these phrases. In order to demonstrate the plausibility of this approach, the results reported below are divided into two parts - one mainly based on "chase", and the other making a further (unimplemented) assumption that other verbs such as "hit" and "push" may also be known using mechanisms similar to those used to discover [chase]. The two differing assumptions are:

a. *Chase-only:* Linguistic forms for [move away] and [come closer] are diffuse, so we consider primarily the learned cluster [chase]. We discover that [chase] maps to "follow", and include sentences with "follow" leading to a corpus of 36+9 sentences which is still small with infrequent specific strings.

b. *+Hit+Push:* In the second model, we assume that in addition to [chase], we have action models and linguistic mappings for the actions [hit] and [push], which occur often in the commentary.

The second (stronger) results should be taken as indicative of the plausibility of the approach, and not as a complete implementation of the algorithm.

Discourses Mapping [chase] Only: Of the three classes of actions for which we have acquired visual schemas from the perceptual data, the narratives for [come-closer] and [move-away] have widely varying constructions. Focusing on the action chase, we discover that it maps to two verbs in the linguistic descriptions: "chase", and "follow". Constructions for both have the structure *[chaser] verb+particle [chased]*.

There are only 36 + 9 sentences with "chase" + "follow", so the data for these arguments is rather sparse. After ruling out phrases that have a sample size of one, cases where the conditional probability of the entity given the phrase is 1 (Steps 3 and 4 of the algorithm), is taken as a synonym (names known earlier in italics) — {[BS]: *big square*, square, big box, large square, big block, bigger square}; {[SS]: *little square*, small square, little box};{[C] : *circle*, little circle, ball, small circle}.

Algorithm 1 A plausible approach towards the discovery of anaphora.

Input :

1. Set of timestamped action predicates *Verb*(*arg1*, *arg2*)
2. Set of timestamped narrative sentences

Alignment :

1. Align co-occurrent predicates with sentences containing the corresponding verb.
2. Increment the object associations against each language phrases L_i:
   - For linguistic constructs of the form [$\langle L_1 \rangle$ verb $\langle L_2 \rangle$], map L_1 to arg1 and L_2 to arg2
   - For constructs of the form [$\langle L_1 \rangle$ verb by $\langle L_2 \rangle$], map L_1 to arg2 and L_2 to arg1
3. For set of three agents (big and small square, circle), plus pairs (total 6 object-groups), estimate the conditional probability P(object/language phrase).
4. If the probability is close to 1, the language phrase is likely to be a proper synonym of the corresponding object.
5. If some linguistic units are acting as a synonym for multiple objects, their referent may not be fixed, but may depend on some other aspect.

Now, after ruling out synonyms and infrequent phrases (those occurring only once), we are left with three units - "it", "them" and "each other" (Table 1). We were surprised ourselves that all three instances found are anaphora. Noticing that these units don't have a fixed referent, other regularities are searched by which their referents can be identified. This may be the start of a process which leads to the idea of anaphora.

With [hit] + [push] : While we have no computational models for actions such as [hit] and [push], there is considerable evidence that these concepts are typically acquired fairly early, and also reflected in early vocabularies (Clark(2003)). In the analysis next (Table 2), we assume the availability of [hit] and [push] models in addition to [chase], and consider the same analysis as above, but now on the larger set of sentences encoding

Phrase (Ph)	# Ph	BS /Ph	SS /Ph	C /Ph	BSSS /Ph	SSC /Ph
it	10	0.5	0.4	0.1	0	0
them	5	0	0	0	0.2	0.8
each other	3	0	0	0	0.66	0.33
[missing]	15	0.46	0.2	0.33	0	0

Table 1: Conditional probability computation (with values in the column headers) for the non-synonymical arguments in sentences mapping [chase] action.

these actions. A few additional synonyms are learned ("he" for [BS], "small box", "little block" for [SS]). Also the labels "square and circle", and "little circle and square" are associated with the combination [SS&C], sentences mapping multiple predicates where both were involved in a patient role. These results may also be interpreted as a slightly advanced stage for the learner, when it has acquired these additional structures.

Step 5 of Algorithm 1 gives the first indication of phenomena such as anaphora. After synonym matching, words remain that are not assigned to any single entity but as in the [chase]-only case, they can be applied to multiple referents. To the learner, this implies that this aspect, that these phrases can be applied to multiple referents, is stable, and not an artifact related to a single action or context. The learner may now attempt to discover other regularities in how the referents for each of these words is assigned. This requires even greater vocabulary, since the prior referent must also be known.

Phrase (Ph)	# Ph	BS /Ph	SS /Ph	C /Ph	BSSS /Ph	SSC /Ph
it	19	0.63	0.26	0.11		0
each other	10	0	0	0	0.9	0.1
they	6	0	0	0	0.66	0.33
them	5	0	0	0	0.2	0.8
[missing]	29	0.59	0.24	0.17	0	0

Table 2: Conditional probability computation (with values in column headers) for the arguments of [chase], [hit] and [push].

Focusing on the word "it", and assuming a greater inventory of verbs, we can consider sequences of sentences such as *The bigger square just went inside the box / Looks like it is chasing the small square*. The "it" in the second sentence is known to our learner as [BS] based on the video parse, and one notes how the agent in the previous sentence is also [BS]. In another situation we have *The large square was chasing the other square / And it got away*. Here the "it" refers to the most recent antecedent, [SS] (though in other examples, it refers to the parallel antecedent). In the chase-only case, we note that "it" refers to the immediately previous referent in 6/10 situations. Two cases involve plural vs single disambiguation: e.g. *Big square is chasing them / They outrun it*, and one case involves parallel reference, e.g. *Now the big square is hitting the small square / It has hit it again* (in fact, unlike our learner, the reader may have difficulty disambiguate the "it"s here). While the referent identification pattern isn't very clear, the learner realizes that "it" at least refers to some earlier referent in the discourse.

Further, even reciprocal anaphors such as "each other" can be recognized since sentences such as *they hit each other* overlap with multiple predicates with switched arguments (*hit([BS],[SS])* and *hit([SS],[BS])*). Beyond this little domain, as our learner is exposed to thousands of linguistic fragments every day, these regularities are likely to get reinforced.

Finally, considering the cases of missing arguments, there are two cues available to the early learner: a) that the relevant action involves two arguments, but fewer are available in the discourse, and b) that the missing argument refers to an antecedent in the discourse. In English, zero anaphora is a very common phenomenon. Even in our very small corpus, there are 570 agents, of which 99 are zero anaphors. Clearly this is a sufficiently high probability phenomenon which deserves the attention of the early learner. Once the absent argument is observed, it can be associated with the appropriate argument. Note that since this substitution is occurring at the semantic level and not in the syntax, only antecedents matching the activity will be considered. Estimating the probabilities in terms of frequencies even for this very small dataset, reveals that of the 99 zero anaphors, 96 refer to the most recent agent argument, often coming as a series e.g. *big square says "uh uh, don't do that" / pushes little square*

around / pushes little square around again/ chases little square. Thus, the most recent argument may emerge as a dominant reference pattern for zero anaphora. Also, we note how considerable knowledge beyond syntax is involved in the remaining situations e.g. *Door is shut/ Went into the corner*.

4 Conclusion

We have outlined how an unsupervised approach correlating prior sensori-motor knowledge with linguistic structures, might be used to eventually learn complex aspects of grammar such as argument structure, and lead to the discovery of phenomena such as anaphora. Also, we highlight many cases of zero anaphora, and show how these may also be inferred, most commonly as the most recent agent in the perceptual input.

However, this work, even though it is different from traditional discourse-only-input-based attempts at anaphora resolution, is clearly just a beginning. We have demonstrated unsupervised learning for only one verb, "chase", and it is by no means clear that other action models needed for other verbs can be similarly learned. Nonetheless, there is considerable work that hints at the infants being able to use perceptual cues to learn the base model of many motion primitives of this nature (Pasek(2006)). But clearly more work is needed to be able to approach verbs that are not directly based on motion. Also, the mapping to language also may not be as straightforward for many other verbs.

This limited demonstration, nonetheless, highlights several points. First, it underscores the role of concept argument structures in aligning with linguistic expressions. It provides some evidence for the position that some aspects of semantics may be ontologically prior to syntax, at least for human-like learning processes. Secondly, it addresses the very vexed question of learning grammar from domain-general capabilities. While a computational demonstration such as this cannot provide full answers, certainly it raises a very plausible mechanism, and attempts to learn some complex grammatical constructs such as anaphora. Finally, it addresses some of the issues related to learning language from shared perception, such as the radical translation argument highlighted by Quine's *gavagai* example (Quine(1960)), and instantiates a possibility that dynamic attention may prune the visual input and align with linguistic focus.

A key aspect underscored by this work is the necessity of creating multimodal databases with video, audio and textual corpora, so that more such learning can take place. This work may be taken merely as a straw model that raises more questions than it answers. It will take considerably more work, and creation of significantly larger resources.

References

Clark, EV. 2003. *First language acquisition*. Cambridge University Press.

Dominey, PF and JD Boucher. 2005. Learning to talk about events from narrated video in a construction grammar framework. *AI* 167(1-2):31–61.

Fang, R., J.Y. Chai, and F. Ferreira. 2009. Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *Proc. of ICMI*. ACM, pages 143–150.

Heider, F. and M.Simmel. 1944. An experimental study of apparent behavior. *American Journal of Psychology* 57:243–259.

Pasek, R ,M.Golinkoff, K Hirsh, editor. 2006. *Action meets word: how children learn verbs*. Oxford University Press US.

Quine, WVO. 1960. *Word and Object*. MIT Press, Cambridge,MA.

Roy, D and E Reiter. 2005. Connecting language to the world. *AI: Special Issue* 167:112.

Satish, G. and A. Mukerjee. 2008. Acquiring linguistic argument structure from multimodal input using attentive focus. pages 43 –48.

Steels, Luc. 2003. Evolving grounded communication for robots. *Trends in Cognitive Sciences* 7(7):308–312.

Stoyanov, V., N. Gilbert, C. Cardie, and E.Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proc. of 47th AMACL and 4th IJCNLP of AFNLP*. ACL, pages 656–664.

Strickert, M and B Hammer. 2005. Merge SOM for temporal data. *Neurocomputing* 64:39–71.

Yu, C. and D.H. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM TAP(TAP)* 1(1):57–80.