Co-reference via Pointing and Haptics in Multi-Modal Dialogues

Lin Chen, Barbara Di Eugenio
Department of Computer Science
University of Illinois at Chicago
851 S Morgan ST, Chicago, IL 60607, USA
{lchen43,bdieugen}@uic.edu

Abstract

This paper describes our ongoing work on resolving third person pronouns and deictic words in a multi-modal corpus. We show that about two thirds of these referring expressions have antecedents that are introduced by pointing gestures or by *haptic-ostensive* actions (actions that involve manipulating an object). After describing our annotation scheme, we discuss the co-reference models we learn from multi-modal features. The usage of haptic-ostensive actions in a co-reference model is a novel contribution of our work.

1 Introduction

Co-reference resolution has received a lot of attention. However, as Eisenstein and Davis (2006) noted, most research on co-reference resolution has focused on written text. This task is much more difficult in dialogue, especially in multi-modal dialogue contexts. First, utterances are informal, ungrammatical and disfluent. Second, people spontaneously use gestures and other body language. As noticed by Kehler (2000), Goldin-Meadow (2003), and Chen et al. (2011), in a multi-modal corpus, the antecedents of referring expressions are often introduced via gestures. Whereas the role played by pointing gestures in referring has been studied, the same is not true for other types of gestures. In this paper, alongside pointing gestures, we will discuss the role played by *Haptic-Ostensive (H-O)* actions, i.e., referring to an object by manipulating it in the world (Landragin et al., 2002; Foster et al., 2008). As far as we know, no computational models of co-reference have been developed that include H-O actions: (Landragin et al., 2002) focused on perceptual salience and (Foster et al., 2008) on generation rather than interpretation. We should point out that at the time of writing we only focus on resolving third person pronouns and deictics.

The rest of this paper is organized as follows. In Section 2 we describe our multi-modal annotation scheme. In Section 3 we present the pronoun/deictic resolution system. In Section 4, we discuss experiments and results.

2 The Data Set

The dataset we use in this paper is a subset of the ELDERLY-AT-HOME corpus (Di Eugenio et al., 2010), a multi-modal corpus in the domain of elderly care. It contains 20 human-human dialogues. In each dialogue, a helper (HEL) and an elderly person (ELD) performed *Activities of Daily Living* (Krapp, 2002), such as getting up from chairs, finding pots, cooking pastas, in a realistic setting, a studio apartment used for teaching and research. The corpus contains videos and voice data in avi format, haptics data collected via instrumented gloves in csv format, and the transcribed utterances in xml format.

We focused on specific subdialogues in this corpus, that we call *Find* tasks: a *Find* task is a continuous time span during which the two subjects were collaborating on finding objects. *Find* tasks arise naturally while helping perform ADLs such as preparing dinner. An excerpt from a *Find* task is shown below, including annotations for pointing gestures and for H-O actions (annotations are per-

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 523–527,
Montréal, Canada, June 3-8, 2012.

formed via the Anvil tool (Kipp, 2001)).

ELD : Can you get me a pot?
HEL: (opens cabinet, takes out pot, without saying a word)
`[Open(HEL,Cabinet1),Take-Out(HEL,Pot1)]`
ELD: Not that one, try over there.
`[Point(ELD,Cabinet5)]`

Because the targets of pointing gestures and H-O actions are real life objects, we designed a referring index system to annotate them. The referring index system consists of compile time indices and run time indices. We give pre-defined indices to targets which cannot be moved, like cabinets, drawers, fridge. We assign run time indices to targets which can be moved, and exist in multiple copies, like cups, glasses. A referring index consists of a type and an index; the index increases according to the order of appearance in the dialogue. For example, "*Pot#1*" means the first pot referred to in the dialogue. If a pointing gesture or H-O action involved multiple objects, we used JSON (JavaScript Object Notation)[1] Array to mark it. For example, [C#1, C#2] means Cabinet#1 and Cabinet#2.

We define a pointing gesture as a hand gesture without physical contact with the target, whereas gestures that involve physical contact with an object are haptic-obstensive (H-O).[2] We use four tracks in Anvil to mark these gestures, two for pointing gestures, and two for H-O actions. In each pair of tracks, one track is used for HEL, one for ELD. For both types of gestures, we mark the start time, end time and the target(s) of the gesture using the referring index system we introduced above. Additionally we mark the type of an H-O action: Touch, Hold, Take-Out (as in taking out an object from a cabinet or the fridge), Close, Open.[3]

Our co-reference annotation follows an approach similar to (Eisenstein and Davis, 2006). We mark the pronouns and deictics which need to be resolved, their antecedents, and the co-reference links between them. To mark pronouns, deictics and textual antecedents, we use the shallow parser from Apache OpenNLP Tools [4] to chunk the utterances in each turn. We use heuristics rules to automatically mark potential textual antecedents and the phrases we need to resolve. Afterwards we use Anvil to edit the results of automatic processing. To annotate co-reference links, we first assign each of the textual antecedents, the pointing gestures and H-O actions a unique markable index. Finally, we link referring expressions to their closest antecedent (if applicable) using the markable indices.

Find Subtasks	142
Length (Seconds)	5009
Speech Turns	1746
Words	8213
Pointing Gestures	362
H-O Actions	629
Pronouns and Deictics	827
Resolved Ref. Expr.	757
Textual Antecedent	218
Pointing Gesture Antecedent	266
H-O Antecedent	273

Table 1: Annotation Statistics

Table 1 shows corpus and annotation statistics. We annotated 142 *Find* subtasks, whose total length is about 1 hour and 24 minutes. This sub-corpus comprises 1746 spoken turns, which include 8213 words. 10% of the 8213 words (827 words) are pronouns or deictics. Note that for only 757/827 (92%) were the annotators able to determine an antecedent. Interestingly, 71% of those 757 pronouns or deictics refer to specific antecedents that are introduced *exclusively* by gestures, either pointing or H-O actions. In the earlier example, only the type for the referent of *that* in *No, not that one* had been introduced textually, but not its specific antecedent `pot1`. Clearly, to be effective on such data any model of co-reference must include the targets of pointing gestures and H-O actions. Our current model does not take into account the type provided by the *de dicto* interpretation of indefinites such as *a pot* above, but we intend to address this issue in future work.

In order to verify the reliability of our annotations, we double coded 15% of the data for pointing gestures and H-O actions, namely the dialogues from 3 pairs of subjects, or 22 *Find* subtasks. We ob-

[1] http://www.json.org/

[2] Whereas not all haptic actions are ostensive, in our dialogues they all potentially perform an ostensive function.

[3] Our subjects occasionally hold objects together, e.g. to fill a pot with water: these actions are not included among the H-O actions, and are annotated separately.

[4] http://incubator.apache.org/opennlp/

tained reasonable κ values: for pointing gestures, κ=0.751, for H-O actions, κ=0.703, and for co-reference, κ=0.70.

3 The Co-reference Model

In this paper we focus on how to use gesture information (pointing or H-O) to solve the referring expressions of interest. Given a pronoun or deictic, we build co-reference pairs by pairing it with the targets of pointing gestures and H-O actions in a given time window. We mark the correct pairs as "True" and then we train a classification model to judge if a co-reference pair is a true pair. The main component of the resolution system is the co-reference classification model. Since our antecedents are not textual, most of the traditional features for co-reference resolution do not apply. Rather, we use the following multi-modal features - *U* is the utterance containing the pronoun / deictic to be solved:

- *Time distance* between the spans of *U* and of the pointing/H-O action. If the two spans overlap, the distance is 0.
- *Speaker agreement:* If the speaker of *U* and the actor of the pointing/H-O action are the same.
- *Markable type agreement:* If the markable type of the pronoun/deictic and of the targets of pointing gesture/H-O action are compatible.
- *Number agreement*: If the number of the pronoun/deictic is the same as that of the targets of the pointing gesture/H-O action.
- *Object agreement*: If the deictic is contained in a phrase, such as "this big blue bowl", we will check if the additional object description "bowl" matches the targets of pointing gesture/H-O action.
- *H-O Action type*: for co-reference pairs with antecedents from H-O actions.

For markable type agreement, we defined two types of markables: PLC (place) and OBJ (object). PLC includes all the targets which cannot easily be moved, OBJ includes all the targets like cups, pots. We use heuristics rules to assign markable types to pronouns/deictics and the targets of pointing gestures/H-O actions. To determine the number of the targets, we extract information from the annotations; if the target is a JSON array, it means it is plural. To extract additional object description for the object agreement feature, we use the Stanford Typed Dependency parser (De Marneffe and Manning., 2008). We check if the pronoun/deictic is involved in "det" and "nsubj" relations, if so, we extract the "gov" element of that relation as the object to compare with the target of gestures/H-O actions.

4 Experiments and Discussions

We have experimented with 3 types of classification models: Maximum Entropy (MaxEnt), Decision Tree and Support Vector Machine (SVM), respectively implemented via the following three packages: MaxEnt, J48 from Weka (Hall et al., 2009), and LibSVM (Chang and Lin, 2011). All of the results reported below are calculated using 10 fold cross validation.

We have run a series of experiments changing the history length from 0 to 10 seconds for generating co-reference pairs (history changes in increments of 1 second, hence, there are 11 sets of experiments). For each history length, we build the 3 models mentioned above. An additional baseline model treats a co-reference pair as "True" if speaker agreement is true for the pair, and the time distance is 0. Beside the specified baseline, J48 can be seen as a more sophisticated baseline as well. When we ran the 10 fold experiment with J48 algorithm, 5 out of 10 generated decision trees only used 3 attributes.

We use two metrics to measure the performance of the models. One are the standard precision, recall and F-Score with respect to the generated co-reference pairs; the other is the number of pronouns and deictics that are correctly resolved. Given a pronoun/deictic p_i, if the classifier returns more than one positive co-reference pair for p_i, we use a heuristic resolver to choose the target. We divide those positive pairs into two subsets, those where the speaker of p_i is the same as the performer of the gesture (SAME), and those with the other speaker (OTHER). If SOME is not empty, we will choose SOME, otherwise OTHER. If the chosen set contains more than one pair, we will choose the target

Model	Hist.	Prec.	Rec.	F.	Number Resolved
Baseline	2	.707	.526	.603	359
J48	1	.801	.534	.641	371
SVM	2	.683	.598	.637	369
MaxEnt	0	.738	.756	**.747**	374
MaxEnt	2	.723	.671	.696	**384**

Table 2: Gesture&Haptics Co-reference Model Results

of the gesture/H-O action in the most recent pair.

Given the space limit, Table 2 only shows the results for each model which resolved most pronouns/deictics, and the model which produced the best F-score. In Table 2, with the change of *History* window setting, the gold standard of co-reference pairs change. When the history window is larger, there are more co-reference candidate pairs, which help resolve more pronouns and deictics.

Given we work on a new corpus, it is hard to compare our results to previous work, additionally our models currently do not deal with textual antecedents. For example Strube and Müller (2003) reports their best F-Measure as .4742, while ours is .747. As concerns accuracy, whereas 384/827 (46%) may appear low, note the task we are performing is harder since we are trying to solve all pronouns/deictics via gestures, not only the ones which have an antecedent introduced by a pointing or H-O action (see Table 1). Even if our feature set is limited, all the classification models perform better than baseline in all the experiments; the biggest improvement is 14.4% in F-score, and solving 25 more pronouns and deictics. There are no significant differences in the performances of the 3 different classification models. Table 2 shows that the history length of the best models is less than or equal to 2 seconds, which is within the standard error range of annotations when we marked the time spans for events.

5 Conclusions

This paper introduced our multi-modal co-reference annotation scheme that includes pointing gestures and H-O actions in the corpus ELDERLY-AT-HOME. Our data shows that 2/3 of antecedents of pronouns/deictics are introduced by pointing gestures or H-O actions, and not in speech. A co-reference resolution system has been built to resolve pronouns and deictics to the antecedents introduced by pointing gestures and H-O actions. The classification models show better performance than the baseline model. In the near future, we will integrate a module which can resolve pronouns and deictics to textual antecedents, including type information provided by indefinite descriptions. This will make the system fully multi-modal. Additionally we intend to study issues of timing. Preliminary studies of our corpus show that the average distance between a pronoun/deictic and its antecedent is 8.26" for textual antecedents, but only 0.66" for gesture antecedents, consistent with our results that show the best models include very short histories, at most 2" long.

Acknowledgments

This work is supported by award IIS 0905593 from the National Science Foundation. Thanks to the other members of the RoboHelper project, especially to Anruo Wang, for their many contributions, especially to the data collection effort. Additionally, we thank the anonymous reviewers for their valuable comments.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Lin Chen, Anruo Wang, and Barbara Di Eugenio. 2011. Improving pronominal and deictic co-reference resolution with multi-modal features. In *Proceedings of the SIGDIAL 2011 Conference*, pages 307–311, Portland, Oregon, June. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Barbara Di Eugenio, Miloš Žefran, Jezekiel Ben-Arie, Mark Foreman, Lin Chen, Simone Franzini, Shankaranand Jagadeesan, Maria Javaid, and Kai Ma. 2010. Towards Effective Communication with Robotic Assistants for the Elderly: Integrating Speech, Vision and Haptics. In *Dialog with Robots, AAAI 2010 Fall Symposium*, Arlington, VA, USA, November.

Jacob Eisenstein and Randall Davis. 2006. Gesture Improves Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 37–40.

Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L. Hill, Jon Oberlander, and Alois Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE international conference on Human Robot Interaction*, HRI '08, pages 295–302. ACM.

S. Goldin-Meadow. 2003. *Hearing gesture: How our hands help us think*. Harvard University Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Andrew Kehler. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *AAAI 00, The 15th Annual Conference of the American Association for Artificial Intelligence*, pages 685–689.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 1367–1370.

Kristine M. Krapp. 2002. *The Gale Encyclopedia of Nursing & Allied Health*. Gale Group, Inc. Chapter Activities of Daily Living Evaluation.

F. Landragin, N. Bellalem, and L. Romary. 2002. Referring to objects with spoken and haptic modalities. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)*, pages 99–104, Pittsburgh, PA.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*.