Improving Pronominal and Deictic Co-Reference Resolution with Multi-Modal Features

Lin Chen, Anruo Wang, Barbara Di Eugenio
Department of Computer Science
University of Illinois at Chicago
851 S Morgan ST, Chicago, IL 60607, USA
{lchen43,awang28,bdieugen}@uic.edu

Abstract

Within our ongoing effort to develop a computational model to understand multi-modal human dialogue in the field of elderly care, this paper focuses on pronominal and deictic co-reference resolution. After describing our data collection effort, we discuss our annotation scheme. We developed a co-reference model that employs both a simple notion of markable type, and multiple statistical models. Our results show that knowing the type of the markable, and the presence of simultaneous pointing gestures improve co-reference resolution for personal and deictic pronouns.

1 Introduction

Our ongoing research project, called RoboHelper, focuses on developing an interface for older people to effectively communicate with a robotic assistant that can help them perform *Activities of Daily Living (ADLs)* (Krapp, 2002), so that they can safely remain living in their home (Di Eugenio et al., 2010). We are devising a multi-modal interface since people communicate with one another using a variety of verbal and non-verbal signals, including haptics, i.e., force exchange (as when one person hands a bowl to another person, and lets go only when s/he senses that the other is holding it). We have collected a mid size multi-modal human-human dialogue corpus, that we are currently processing and analyzing. Meanwhile, we have started developing one core component of our multi-modal interface, a co-reference resolution system. In this paper, we will present the component of the system that resolves pronouns, both personal (*I, you, it, they*), and deictic (*this, that, these, those, here, there*). Hence, this paper presents our first steps toward a full co-reference resolution module, and ultimately, the multi-modal interface.

Co-reference resolution is likely the discourse and dialogue processing task that has received the most attention. However, as Eisenstein and Davis (2006) notes, research on co-reference resolution has mostly been applied to written text; this task is more difficult in dialogue. First, utterances may be informal, ungrammatical or disfluent; second, people spontaneously use hand gestures, body gestures and gaze. Pointing gestures are the easiest gestures to identify, and vision researchers in our project are working on recognizing pointing and other hand gestures (Di Eugenio et al., 2010). In this paper, we replicate the results from (Eisenstein and Davis, 2006), that pointing gestures help improve co-reference, in a very different domain. Other work has shown that gestures can help detect sentence boundaries (Chen and Harper, 2010) or user intentions (Qu and Chai, 2008).

The rest of the paper is organized as follows. In Section 2 we describe the data collection and the ongoing annotation. In Section 3 we discuss our co-reference resolution system, and we present experiments and results in Section 4.

2 The ELDERLY-AT-HOME corpus

Due to the absence of multi-modal collaborative human-human dialogue corpora that include haptic data beyond what can be acquired via point-and-touch interfaces, and in the population of interest,

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 307–311,
Portland, Oregon, June 17-18, 2011.

Figure 1: Experiment Excerpts

we undertook a new data collection effort. Our experiments were conducted in a fully functional studio apartment at Rush University in Chicago – Figure 1 shows two screen-shots from our recorded experiments. We equipped the room with 7 web cameras to ensure multiple points of view. Each of the two participants in the experiments wears a microphone, and a data glove on their dominant hand to collect haptics data. The ADLs we focused on include ambulating, getting up from a bed or a chair, finding pots, opening cans and containers, putting pots on a stove, setting the table etc. Two students in gerontological nursing play the role of the helper (HEL), both in pilot studies and with real subjects. In 5 pilot dialogues, two faculty members played the role of the elderly person (ELD). In the 15 real experiments, ELD resides in an assisted living facility and was transported to the apartment mentioned above. All elderly subjects are highly functioning at a cognitive level and do not have any major physical impairment.

The size of our collected video data is shown in Table 1. The number of subjects refers to the number of different ELD's and does not include the helpers; we do include our 5 pilot dialogues though, since those pilot interactions do not measurably differ from those with the real subjects. Usually one experiment lasts about 50' (recording starts after informed consent and after the microphones and data gloves have been put on). Further, we eliminated irrelevant content such as interruptions, e.g. by the person who accompanied the elderly subjects, and further explanations of the tasks. This resulted in about 15 minutes of what we call *effective* data for each subject; the effective data comprises 4782 turns (see Table 1).

Subjects	Raw(Mins)	Effective(Mins)	Turns
20	482	301	4782

Table 1: ELDERLY-AT-HOME Corpus Size

The effective portion of the data was transcribed by the first two authors using the Anvil video annotation tool (Kipp, 2001). A subset of the transcribed data was annotated for co-reference, yielding 114 sub-dialogues corresponding to the tasks subjects perform, such as finding bowls, filling a pot with water, etc. (see Table 2).

An annotation excerpt is shown in Figure 2. Markable tokens are classified into *PLC(Place), PERS(Person), OBJ(Object)* types, and numbered by type, e.g., PLC#5. Accordingly, we mark pronouns with types as well, *RPLC, RPERS, ROBJ*, e.g. RPLC#5. If a subject produced a pointing gesture, we generate a markable token to mark what is being pointed to at the end of the utterance (see Utt. 4 and 5 in Figure 2). Within the same task, if two markables have the same type and the same markable index, they are taken to co-refer (hence, longer chains of reference across tasks are cut into shorter spans).

Haptics annotation is at the beginning. We have identified *grab*, *hold*, *give* and *receive* as high-level haptics phonemes that may be useful from the language point of view. We have recently started annotating our corpus with those labels.

Subjects	Tasks	Utterances	Gestures	Pronouns
12	114	1920	896	1635

Table 2: Annotated Corpus Size

In order to test the reliability of our annotation, we double coded about 18% of the data, namely 21 sub-dialogues comprising 213 pronouns, on which we computed the Kappa coefficient (Carletta, 1996). Similar to (Rodrıguez et al., 2010), we measured the reliability of **markable** annotations, and of **link to the antecedent** annotations. As concerns the **markable** level, we obtained κ=0.945, which is high but no surprisingly for such a simple task. At the **link to the antecedent** level, we compared the links from pronouns to antecedents in a specified context of 4 utterances, obtaining a reasonable κ=0.723.

3: PERS#1(HEL/NNP) : RPERS#1(I/PRP) do/VBP n't/RB see/VB any/DT OBJ#3(pasta/NN) ./.
4: PERS#2(ELD/NNP) : Try/VB over/IN RPLC#5(there/RB) ./. {PLC#5(cabinet/NN)}
5: PERS#1(HEL/NNP) : This/DT RPLC#5(**one/NN**) ?/. {PLC#5(cabinet/NN)}
6: PERS#2(ELD/NNP) : Oh/UH ,/, yes/RB ./.

Figure 2: Annotation Excerpt

3 Our approach

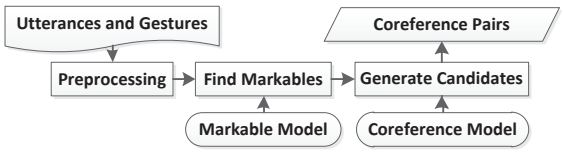

Figure 3: Co-reference System Architecture

The architecture of our co-reference resolution system is shown in Figure 3.

We first pre-process a dialogue by splitting turns into sentences, tokenizing sentences into tokens, POS tagging tokens. The *Markable* model is used to classify whether a token can be referred to and what type of markable it is. The *Markable* model's feature set includes the POS tag of the token, the word, the surrounding tokens' POS tags in a window size of 3. The model outputs markable classes: Place/Object/Person, or None, which means the token is not markable. A pointed-to entity serves as a markable by default.

To perform resolution, each pronoun to be resolved (*I, you, it, they; this, that, these, those, here, there*) is paired with markables in the context of the previous 2 utterances, the current utterance and the utterance that follows, by using {pronoun, markable type} compatibility rules. For example, let's consider the excerpt in Figure 2. To resolve *one* in utterance 5, the system will generate 3 candidate token pairs: <one(5,2), pasta(3,6)>, <one(5,2), cabinet(4,-1)>, <one(5,2), cabinet(5,-1)> (including the pointed-to markable is a way of roughly approximating information that will be returned by the vision component). The elements in those pairs are tokens with their coordinates in the format (*SentenceIndex*, *TokenIndex*); markables pointed to are given negative token indices.

The *Co-reference* model will filter out the pairs <pronoun, markable> that it judges to be incorrect. For the *Co-reference* model, we adopted a subset of features which are commonly used in co-reference resolution in written text. These features apply to each <pronoun, markable> pair and include: *Lexical* features, i.e. words and POS tags for both anaphora and antecedent; *Syntactic* features, i.e. syntactic constraints such as number and person agreement; *Distance* features, i.e. sentence distance, token distance and markable distance. Additionally, the Co-reference model uses pointing gesture information. If the antecedent in the <pronoun, markable> was pointed to, the pair is tagged as *IsPointed*. In our data, people often use pronouns and hand gestures instead of nouns when introducing new entities. It is not possible to map these pronouns to a textual antecedent since none exists. This confirms the findings from (Kehler, 2000): in a multi-modal corpus, he found that no pronoun is used *without* a gesture when it refers to a referent which is not in focus.

4 Experiments and Discussion

The classification models described above were implemented using the Weka package (Hall et al., 2009). Specifically, for each model, we experimented with J48 (a decision tree implementation) and LibSVM (a Support Vector Machine implementation). All the results reported below are calculated using 10 fold cross-validation.

We evaluated the performances of individual models separately (Tables 3 and 4), and of the system as a whole (Table 5).

Algorithm	Precision	Recall	F-Measure
J48	0.984	0.984	0.984
LibSVM	0.979	0.936	0.954
Baseline	0.971	0.971	0.971

Table 3: Markable Model Performance

The results in Table 3 are not surprising, since detecting the type of markables is a simple task. Indeed the results of the baseline model are extremely

Method	J48			LibSVM		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Text + Gesture	0.700	0.684	0.686	0.672	0.669	0.670
Text Only	0.655	0.656	0.656	0.624	0.624	0.624

Table 4: Co-reference Model Performance

Words	Method	Features	Precision	Recall	F-Measure
All Pronouns	J48	Text Only	0.544	0.332	0.412
		Text + Gesture	0.482	0.783	0.596
	LibSVM	Text Only	0.56	0.27	0.364
		Text + Gesture	0.522	0.6	0.559
	Baseline	Text Only	0.367	0.254	0.300
		Text + Gesture	0.376	0.392	0.384
3rd Person + Deictic	J48	Text Only	0.264	0.028	0.05
		Text + Gesture	0.438	0.902	0.589
	LibSVM	Text Only	0.6	0.009	0.017
		Text + Gesture	0.525	0.695	0.598
	Baseline	Text Only	0.172	0.114	0.137
		Text + Gesture	0.301	0.431	0.354

Table 5: Co-reference System Performance (Markable + Co-reference Models)

high as well. We compute the baseline by assigning to the potential markable (i.e., each word) its most frequent class in the training set (recall that the four classes include *None* as well).

For the *Co-reference* model, we conducted 2 sets of experiments to ascertain the effect of including *Gesture* in the model. As shown in Table 4, both J48 and LibSVM obtain better results when we include gestures in the model. χ^2 shows that differences in precision and recall [1] are significant at the $p \leq 0.01$ level, though the absolute improvement is not high.

As concerns the evaluation of the whole system, we ran a 4-way experiment, where we examine the performance of the system on all pronouns, and on those pronouns left after eliminating first and second person pronouns, without and with *G*esture information. We also ran two sets of baseline experiments. In the baseline experiments, we link each pronoun we want to resolve, to the most recent utterance-markable token and to a pointed-to markable token (if applicable). Markables are filtered by the same compatibility rules mentioned above.

Regarding the metrics we used for evaluation, we used the same method as Strube and Müller (2003), which is also similar to MUC standard (Hirschman, 1997). As the golden set, we used the human annotated links from the pronouns to markables in the same context of four utterances used by the system. Then, we compared the co-reference links found by the system against the golden set, and we finally calculated precision, recall and F-Measure.

Table 5 shows that the F-measure is higher when including gestures, no matter the type of pronouns. When we include gestures, there is no difference between "All Pronouns" and "$3rd$ Person + Deictic". In the "$3rd$ Person + Deictic" experiments, we observed huge drops in recall, from 0.902 to 0.028 for J48, and from 0.695 to 0.009 for LibSVM algorithm. This confirms the point we made earlier, that 3rd person pronouns/deictic words (Kehler, 2000) often do not have textual antecedents, since when accompanied by simultaneous pointing they introduce new entities in a dialogue.

Comparison to previous work is feasible only at a high level, because of the usage of different corpora and/or measurement metrics. This said, our model with gestures outperforms Strube and Müller (2003), who did not use gesture information to resolve pronouns in spoken dialogue. Strube and Müller (2003) used the 20 Switchboard dialogues as their experiment dataset, and used the MUC metrics. Our re-

[1] χ^2 does not apply to the F-Measure.

sults are similar to Eisenstein and Davis (2006), but there are two main differences. First, the corpus they used is smaller than what we used in this paper. Their corpus was collected by themselves and consisted of 16 videos, each video was 2-3 minutes in length. Second, they used a difference measurement metrics called CEAF (Luo, 2005).

5 Conclusions

In this paper, we presented the new ELDERLY-AT-HOME multi-modal corpus we collected. A co-reference resolution system for personal and deictic pronouns has been developed on the basis of the annotated corpus. Our results confirm that gestures improve co-reference resolution; a simple notion of type also helps. The *Markable* and *Co-reference* modules we presented are a first start in developing a full multi-modal co-reference resolution module. Apart from completing the annotation of our corpus, we will develop an annotation scheme for haptics, and investigate how haptics information affects co-reference and other dialogue phenomena. Ultimately, both pointing gestures and haptic information will automatically be recognized by the collaborators in the project we are members of.

Acknowledgments

This work is supported by award IIS 0905593 from the National Science Foundation. Thanks to the other members of the RoboHelper project, for their many contributions, especially to the data collection effort.

References

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Lei Chen and Mary P. Harper. 2010. Utilizing gestures to improve sentence boundary detection. *Multimedia Tools and Applications*, pages 1–33.

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.

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).

Lynette Hirschman. 1997. Muc-7 coreference task definition.

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.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaolin Qu and Joyce Y. Chai. 2008. Beyond attention: the role of deictic gesture in intention recognition in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 237–246.

Kepa Joseba Rodrıguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of the 7th International Conference on Language Ressources and Evaluation (LREC 2010)*, pages 157–163.

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*.