Resolving *"You"* in Multi-Party Dialog*

Surabhi Gupta
Department of Computer Science
Stanford University
Stanford, CA 94305, US
surabhi@cs.stanford.edu

John Niekrasz, Matthew Purver
Center for the Study
of Language and Information
Stanford University
Stanford, CA 94305, US
{niekrasz,mpurver}@stanford.edu

Dan Jurafsky
Department of Linguistics
Stanford University
Stanford, CA 94305, US
jurafsky@stanford.edu

Abstract

This paper presents experiments into the resolution of *"you"* in multi-party dialog, dividing this process into two tasks: distinguishing between generic and referential uses; and then, for referential uses, identifying the referred-to addressee(s). On the first task we achieve an accuracy of 75% on multi-party data. We achieve an accuracy of 47% on determining the actual identity of the referent. All results are achieved without the use of visual information.

1 Introduction

This paper discusses the disambiguation and resolution of second-person pronoun *"you"* in dialog. Our objectives are to automatically disambiguate the *referential* and *generic* senses, and to resolve the referential instances through *addressee identification* using linguistic features.

This work is motivated by the effort to automatically extract useful information from multi-party human-human conversation. For example, Purver et al. (2006) attempt to extract action items: concrete decisions in which one (or more) individuals makes a commitment to perform a given task. One critical roadblock is that of personal deixis: identification of the individuals in question usually involves personal pronouns – either singular or plural, as (1) and (2) – rather than explicit naming.

(1)
A: and um if **you** can get that binding point also maybe with a nice example that would be helpful for **Johno** and **me**.
B: Oh yeah uh O_K.

(2)
A: So y- so **you guys** will send to **the rest of us** um a version of um, this, and - the - uh, description -
B: With sugge- yeah, suggested improvements and -

However, to complicate the matter, many second-person pronoun uses[1] do not refer to the addressee and are generic, as in example (3):

(3)
B: Well, usually what **you** do is just wait until you think it's stopped,
and then **you** patch them up.

The problem of *"you"* resolution can therefore be seen as two sub-problems: (1) determining whether a *"you"* is referential, and if so, (2) determining the number and identity of the addressee(s) referred to.

2 Related Work

Prior linguistic work has recognized that *"you"* is not always addressee-referring – e.g. Jurafsky et al. (2002) distinguish three cases (generic, referential, and the conventional phrase *"you know"*) in an investigation of phonological form – but little work exists on automatic disambiguation.[2] In (Gupta et al., 2007), we described such a study (to our knowledge, the first) on two-person telephone conversations, achieving 84% accuracy on disambiguating generic and referential senses. Here, we extend this to multi-party data, and add addressee resolution.

In multi-party dialog (unlike two-party dialog), addressee identification is an important problem, and there is a growing body of research addressing this (Jovanovic et al., 2006b; Katzenmaier et

*This work was supported by the CALO project (DARPA grant NBCH-D-03-0010) and ONR (MURI award N000140510388).

[1] There are five second-person pronouns in English: *you*, *your*, *yours*, *yourself*, and *yourselves*.

[2] Although there is related work on classifying *"it"*, which also takes various referential and non-referential readings (Müller, 2006).

al., 2004). Jovanovic et al. (2006b) achieve an accuracy of 83.74% when including visual features such as gaze information, along with more complex information such as meeting action types (e.g. discussion, presentation, white-board). Galley et al. (2004) showed some success using only speech-based information for a related problem – identifying the first halves of adjacency pairs (whose speakers will in many cases be the addressees of the second halves) – achieving an accuracy of 90.2%. Our approach in this paper is closer to the latter in using only non-visual information, in order to support a solution in environments lacking video.

3 Data

We used the AMI Meeting Corpus (McCowan et al., 2005), a multi-modal dataset of 4-party meetings. The meetings are scenario-driven – participants are assigned roles in a loosely scripted collaborative design task, averaging about 30 minutes in duration. All meetings are hand-transcribed and annotated for dialog acts; we used a 15-meeting subset which is also annotated for addressee (Jovanovic et al., 2006a), with each utterance labeled with the set of addressees. Jovanovic et al. (2006a) report that 34.2% of utterances were addressed to all participants, 61.7% were addressed to single individuals, with <2% being addressed to 2-person subgroups.

We randomly selected a subset of utterances containing *"you"* to annotate. Only text and/or audio were made available to annotators – no videos were provided during annotation. The result was a 4-way classification on a per-utterance basis using the following classes: *generic*, *referential*, *reported referential*, and *discourse marker*. Examples of the first three of these classes are given above. The *reported referential* class was used to mark when speakers are quoting other speakers' referential uses, as in example (4). Finally, the *discourse marker* class was used to mark instances of the commonly-occurring, semantically bleached version of *"you know"*.

(4)
B: Well, uh, I guess probably the last one I went to I met so many people that I had not seen in probably ten, over ten years.
It was like, don't **you** remember me.
And I am like no.
A: Am I related to **you**?

The reliability of our annotations was acceptable, with kappa of .84 and raw inter-tagger accuracy of .92 (assessed over a subset of 108 instances tagged by two authors). The resulting dataset for generic versus referential consisted of 952 utterances for training and 374 for test; overall, 47.4% of cases were generic. Since the addressee annotations do not cover all utterances in the meetings, the dataset for addressee detection had only 291 utterances for training and 176 utterances for testing (this set of experiments were performed for the utterances marked as referential); 59.7% of the utterances were addressed to one person.

For the experiments below, we excluded the *reported referential* and the *discourse marker* class since they both occurred in less than 2% of the dataset. Note also that the author performing classification experiments annotated the training set, reserving the test set for annotation by another author.

4 Referentiality Detection

We first investigate the disambiguation of generic versus referential uses. In our earlier work on the two-party Switchboard corpus, we achieved an accuracy of 84.4%, significantly above the baseline performance of 54.6% (always predicting the dominant class). The best classifier made use of a diverse set of features including lexical, part-of-speech, and dialog act features, together with a set of oracle context features (which assumed perfect knowledge of the classes of the preceding utterances).

Here, as well as applying the approach to more complex multi-party data, we wanted to remove the requirement for these unrealistic oracle context features. We therefore used a sequence classifier — conditional random field (CRF), first introduced by Lafferty et al. (2001) — allowing us access to the same contextual information, but via the output of the classifier. The full set of features is shown in Table 1.

Note that in the absence of an available DA tagger for this data, we use manually produced DA tags. This is also unrealistic; we therefore investigated the substitution of the full DA tagset features with a single Q_DA feature which indicates the presence of a questioning dialog act (the AMI `elicit` acts).

N	Features
	Sentential Features (Sent)
2	you, you know, you guys
N	number of you, your, yourself
2	you (say|said|tell|told|mention(ed)|mean(t)|sound(ed))
2	you (hear|heard)
2	(do|does|did|have|has|had|are|could|should|n't) you
2	"if you"
2	I|we
2	(which|what|where|when|how) you
	Part of Speech Features (POS)
2	Comparative JJR tag
2	you (VB*)
2	(I|we) (VB*)
2	(PRP*) you
	Dialog Act Features (DA)
16	DA tag of current utterance i
16	DA tag of previous utterance $i-1$
16	DA tag of utterance $i-2$
	Other Features (QM)
2	Question mark

Table 1: Features investigated (adapted from (Gupta et al., 2007)). N indicates the number of possible values (there are 16 DA tags).

Features	Accuracy
Baseline	57.9%
Sent + POS + QM	63.0%
DA	71.9%
Sent + POS + QM + Q_DA	70.6%
Sent + POS + QM + DA	75.1%

Table 2: CRF results: generic versus referential

4.1 Results & Discussion

A dominant class baseline on this data gives an accuracy of 57.9% (see Table 2). Our best set of features achieve an accuracy of 75.1% (see Table 2).

Our automatically extracted features (sentential, part of speech and question mark) achieve an accuracy of 63% which is above the baseline. Adding oracle dialog act information increases accuracy to 75.1%; substituting the more realistic Q_DA feature gives a smaller improvement, resulting in 70.6%. Note that accuracy is lower than the 84.4% achieved for two-person data, suggesting that referentiality in multi-party meetings is a harder task.

5 Reference Resolution

For referential cases, we must now identify the reference of *"you"* – in other words, the addressee. As our interest is in resolving *"you"*, we investigate this only for the referential utterances as marked by our annotators (not for all utterances). The AMI corpus has 4 meeting participants for each meeting. As 2-person subgroup addressing is rare (see above), we can model the problem as a four way classification task for each utterance – each of the 3 other participants and the entire group.

Since we have multiple meetings with possibly different participants, it makes little sense to index potential addressees by their real-world identity. Instead, for a given utterance, the potential addressee to speak next gets a label of 1; the other two are given labels of 2 and 3 based on the order in which they next speak. We use a label of 4 to represent addressing to the entire group.

Baseline. We can build two baselines. The *Next Speaker* baseline always predicts the addressee to be the next (different) speaker (i.e. a label of 1). The *Previous Speaker* baseline predicts the addressee to be the most recent previous different speaker.

Features. We expect that the structure of the dialog gives the most indicative cues to addressee: forward-looking dialog acts are likely to influence the addressee to speak next, while backward-looking acts might address a recent speaker. We therefore use similar features to those of Galley et al. (2004) for the related task of identifying the first half of an adjacency pair. However, since their task was retrospective, their features all involve facts about the previous discourse context. We therefore adapt the approach to examine features of subsequent as well as preceding utterances.

For each utterance and potential addressee, we examine the pair made up of the original utterance A and the next (or previous) utterance B spoken by that potential addressee. We then extract features of the pair which might indicate the degree of relatedness of the utterances, including their overlap, separation and lexical similarity, as shown in Table 3.

We also added a feature for the number of speakers that talk during the next 5 utterances to allow for better prediction of group addressing. In addition we included the features from Table 1, to test whether the features found useful for generic vs. referential disambiguation would be useful for the task of addressee detection.

Structural Features
. number of speakers between A and B
. number of utterances between A and B
. number of utterances of speaker B between A and B
. number of speakers that talk during the next 5 utterances
. do A and B overlap?
Durational Features
. duration of A
. if no overlap, time separating A and B
. if overlap, duration of overlap
. time of overlap with previous speaker
. time of overlap with next speaker
. speech rate of A
Lexical Features
. number of words in A
. number of content words in A
. ratio of words in A that are also in B
. ratio of words in B that are also in A
. number of cue words (Hirschberg and Litman, 1993) in A

Table 3: Features for addressee identification adapted from (Galley et al., 2004). We obtain a set of backward looking (BL) and forward looking (FL) features for an utterance.

Features	Accuracy
Baseline: Previous Speaker	23.0%
Baseline: Next Speaker	37.0%
FL + BL + Table 1	47.2%

Table 4: Addressee detection results.

Results & Discussion A CRF trained using all our features achieves an accuracy of 47.2%, which is a significant improvement on the baseline. Table 4 presents all the results.

The biggest confusion was found to be between utterances being classified as 1 or 4 (i.e. the next speaker or the entire group). Future work will therefore involve selecting features which can better discern between these two classes.

6 Conclusion

For generic vs. referential *you* disambiguation, our approach developed on two-party data transfers reasonably well to multi-party data. While accuracy is lower, it is significantly above the baseline. Use of a sequence model classifier has allowed us to operate without oracle context features, and a reduced dialog act tagset (question identification) provides reasonable (though reduced) accuracy. A next step here could be to use automatically classified dialog act tags.

Addressee detection is a hard problem, but we have shown promising results. We expect that investigation of further features, potentially including video information, will improve performance.

References

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

S. Gupta, M. Purver, and D. Jurafsky. 2007. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

J. Hirschberg and D. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

N. Jovanovic, R. op den Akker, and A. Nijholt. 2006a. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23.

N. Jovanovic, R. op den Akker, and A. Nijholt. 2006b. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*.

D. Jurafsky, A. Bell, and C. Girand. 2002. The role of the lemma in form variation. In C. Gussenhoven and N. Warner, editors, *Papers in Laboratory Phonology VII*, pages 1–34. Mouton de Gruyter.

M. Katzenmaier, R. Stiefelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of ICMI*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI Meeting Corpus. In *Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*.

C. Müller. 2006. Automatic detection of nonreferential *It* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

M. Purver, P. Ehlen, and J. Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *MLMI 2006, Revised Selected Papers*.