Second-Person Pronoun Resolution

Second-person pronoun resolution in multi-party spoken English dialogue*

Raquel Fernández**
University of Amsterdam

Matthew Frampton†
Stanford University

Stanley Peters‡
Stanford University

Matthew Purver§
Queen Mary University of London

This paper discusses the problem of second-person pronoun resolution in dialogue: determining who (if anyone) the word 'you' refers to. We motivate the task, and break it down into three distinct subtasks – distinguishing generic from deictic uses, distinguishing singular from plural uses, and determining individual reference. We then describe a dataset and series of supervised classification experiments, and show that various linguistic and non-linguistic features can be used to achieve overall accuracies of up to 78%.

1. Introduction

In English, the second-person pronoun *you* can be used in semantically distinct ways. It can be used in a *generic* sense, referring to nobody in particular (1a); or in a *deictic* sense, referring to the current addressee(s) (1b). In multi-party dialogue, this deictic reference may be to any one of the non-speaking participants, to all of them together, or to some subset of them (1c). In addition, *you* can appear as part of a *discourse marker*, most commonly the expression *you know* (1d). Resolving instances of *you* – as we must do if we are to, say, infer action item assignments from business meetings, or interpret commands to members of a team of robots – is therefore a far from trivial process. We must not only determine their generic or deictic nature, but for deictic cases we must also determine the identity of the addressee (s).

(1) a. Often, you need to know specific button sequences to get certain functionalities done.
b. I think it's good. You've done a good review.
c. I don't know if you guys have any questions.
d. It's not just, you know, noises like something hitting.

* The authors are listed in alphabetical order.
** Institute for Logic, Language & Computation, University of Amsterdam, P.O Box 94242, 1090 GE Amsterdam, Netherlands. E-mail: raquel.fernandez@uva.nl.
† Center for the Study of Language and Information, Stanford University, Stanford, CA 94305, USA. E-mail: frampton@stanford.edu.
‡ Center for the Study of Language and Information, Stanford University, Stanford, CA 94305, USA. E-mail: peters@csli.stanford.edu.
§ School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK. E-mail: m.purver@qmul.ac.uk.

In this paper, we propose and evaluate automatic methods for both parts of the task. We make use of many sources of linguistic information, including prosodic, lexical, syntactic and pragmatic features, as well as visual information concerning gaze. Distinguishing generic from deictic uses turns out to be essentially a linguistic problem, requiring information about the utterance itself; resolving reference requires more knowledge of the context, both linguistic (e.g. turn-taking) and non-linguistic (gaze direction). By using Bayesian networks, we can combine these various sources to achieve accurate supervised classification: our best integrated resolution system (a cascade of supervised classifiers) achieves an overall accuracy of 78%. Robustness to the unavailability of individual streams of information is good, with accuracies of 73% achieved without visual information, and 75% using only information which is available in real time. Robustness to noisy input data is also good: a system whose features are derived entirely by automatic means (via ASR and an automatic gaze direction detector) achieves 71% accuracy.

1.1 Motivation

Second-person pronouns are pervasive in dialogue. While in written text the pronoun *you* occupies the 20th position in the British National Corpus (BNC) word frequency lists, in unrestricted spoken English conversation *you* is the second most frequent word (Kilgarriff 1997). Effective dialogue processing systems cannot therefore be limited to pronoun resolution modules that can only deal with third person pronouns and demonstratives (as in most systems designed for text or monologue). Given the rich history of research into dialogue systems, it might seem strange that little attention has been paid to this problem until now – but in fact, it's not that surprising. In the past, dialogue systems research has concentrated almost exclusively on two-party dialogue, in which deictic *you* resolution is trivial; and on human-computer interaction in restricted task-oriented or information-seeking domains, in which generic uses are rare.

Recent advances, though, mean that the problem is becoming more pressing. Firstly, interest is increasing into multi-party dialogue applications, perhaps involving interaction between a human user and multiple virtual characters (Traum et al. 2008) or teams of dialogue-capable robots (Hiatt and Cavedon 2005), and in such cases the resolution of addressee reference is crucial. Secondly, improvements in speech recognition and robust language processing are leading to systems which support human-human dialogue – for example, systems which produce summaries or extract information from business meetings (Tur et al. 2010) or broadcast transcripts (Beeferman, Berger, and Lafferty 1999) – which have to deal with more natural, unconstrained and often open-domain language in which both generic and referential uses can be common.

Considering the case of an automatic summarisation system, we can easily see why we need to resolve second-person pronouns. Imagine a business meeting discussion, from which our task is to produce a to-do list for the appropriate participants:

(2) A: Well you know I think we need to get the application in
B: Uh-huh
A: As soon as possible really
B: Yeah if you get it in by November you get a discount
A: Yeah so do you think you could do that?
C: OK sure

Understanding who the task has been assigned to (in this case, apparently C) is a case of resolving the deictic reference of the uses of *you* in A's final question: we need

to determine that this utterance assigns (or attempts to assign) the task "do that" to the addressee *you*, and detect who that addressee is. However, we must avoid drawing the same kind of inferences from the immediately preceding observation by B: the uses of *you* here are generic, and B is not suggesting that the addressee of this utterance (apparently A) should be the one to "get it in" or "get a discount". It may also be important to know that "you know" in A's first utterance is simply a discourse marker, rather than telling us about any particular participant's knowledge state.

A further motivation comes from the field of machine translation. Many languages distinguish between generic and deictic *you* explicitly (French and German, for example, have the generic pronouns *on* and *man*), and/or reflect the number of people being addressed (French's singular *tu* and plural *vous*, or Māori's *koe* for one person, *kōrua* for two, and *koutou* for more than two). In some, pronoun choice can depend on relative status, requiring some knowledge of the individual being addressed (e.g. German's *Sie* or the more complex system in Japanese). Understanding *you* reference must therefore be crucial for any translation system designed for use in dialogue.

1.2 Overview

The rest of this article is organised as follows. We start by surveying some previous approaches to reference resolution and addressee detection that are related to our work. In Section 3, we introduce the sub-corpus which we use in our investigation – a portion of the AMI Meeting Corpus (McCowan et al. 2005) – and describe how the utterances containing the pronoun *you* in our dataset are coded. Section 4 provides a detailed description of the features which we extract in order to tackle the *you* resolution task and its sub-parts. After this, Section 5 reports our experiments and their results. We assess the contribution of different types of features and perform an error analysis for each subtask. We end with conclusions in Section 6.

2. Related Work

In this section we review some related previous work, firstly on reference resolution for pronouns other than *you*, and secondly on addressee resolution in dialogue.

2.1 Pronominal Reference Resolution

2.1.1 Third-person pronouns. Most work on pronoun resolution deals with anaphoric reference resolution in written discourse (Jurafsky and Martin 2009, for example). This contrasts strongly with the case of *you*, where we are forced to deal with exophoric referents that are part of dialogue situations. In recent years, researchers working on reference resolution have begun to extend their systems to deal not only with discourse but also with the more complex scenario of dialogue, where language is more disfluent and far less structured. Some examples are, for instance, Byron (2004) and Arstein and Poesio (2006), who attempt to resolve references to abstract entities with implicit discourse antecedents, and Strube and Müller (2003) and Müller (2007), who focus on resolving third person pronouns and demonstratives. Performance varies substantially amongst systems. For instance, Byron (2004) achieves an F-score of around 69% on task-oriented two-party dialogue transcripts with the help of costly information such as domain-dependent semantic category restrictions for predicate argument positions, while the F-score obtained by the domain-independent system of Müller (2007) on un-preprocessed multi-party transcripts falls to less than 20%.

2.1.2 Non-referential "it". An interesting sub-problem within third person pronoun resolution is the identification of non-referential instances of *it* – similar to the genericity problem for *you*. The work of Evans (2001) and Boyd, Gegg-Harrison, and Byron (2005) on monological text shows that between 20 and 30% of instances of *it* are non-referential. This is even more pronounced in dialogue, where according to Müller (2006), non-referential instances amount to around 37%, including not only pleonastic uses of *it* but also *discarded* uses where the pronoun appears in abandoned or disfluent utterances, which are rather common in spoken dialogue. These approaches use a variety of features to filter out non-referential *it*, including syntactic patterns, lexical information, and the distance between the pronoun and elements such as complementizers or infinitives. The F-scores obtained, for both text and dialogue, are around 70%.

Bergsma, Lin, and Goebel (2008) present a novel approach to identifying non-referential *it* in text. Unlike other approaches that use manually defined features, here the authors explore a purely distributional approach that involves first representing the context as distributional patterns with the pronoun's position as a wildcard (e.g. "*you can make * in advance*"), and second enumerating all words that can be found in this position. The distributional patterns are extracted from the Google Web 1T 5-gram Corpus Version 1.1. Referential distributions occur with a variety of noun phrase fillers, while distributions that correlate with non-referential instances tend to have, almost exclusively, the pronoun *it* filling the wildcard position (e.g. "*you can make * in Hollywood*"). The system gives good results, with F-scores between 69 and 83%, showing that lexical and distributional information can be very powerful. As we explain in Section 4, we use lexical features to exploit this kind of information.

2.1.3 Second-person pronouns. Previous linguistic work has recognized that *you* is not always addressee-referring, differentiating between the generic and deictic (or *referential*) uses (Holmes 1998; Meyers 1990). Jurafsky, Bell, and Girand (2002) distinguish between three different cases – the generic and referential cases, and the conventional phrase *"you know"* – in order to empirically investigate the relation between these cases and their realized phonological forms. They found that *"you know"* covered 47% of cases, the referential class 22%, and the generic class 27%, but found no significant differences in surface form (duration or vowel reduction) between the different cases. However, we are not aware of any previous work in resolving second-person pronouns other than our own (see Section 2.3).[1]

2.2 Addressee Identification

2.2.1 Rule-based approaches. Most early work in dialogue processing concentrated on two-person dialogue, in which addressee resolution is essentially trivial. Once researchers started to focus on more complex multi-party data, addressee resolution became an issue. Traum (2004) defines a rule-based algorithm for determining an utterance's addressee which depends mainly on turn-taking behaviour:

(3) a. If utterance specifies a specific addressee (e.g. a vocative or utterance of just a name when not expecting a short answer or clarification of type person) then *Addressee = specified addressee*
b. else if speaker of current utterance is the same as the speaker of the immediately previous utterance then *Addressee = previous addressee*

1 But see (Baldwin, Chai, and Kirchhoff 2010) for a recent approach investigating how information about hand gestures can help distinguish between generic and deictic uses of *you*.

c. else if previous speaker is different from current speaker then *Addressee = previous speaker*
d. else if unique other conversational participant (i.e. a 2-party conversation) then *Addressee = that other participant*
e. else *Addressee = unknown*

Traum, Robinson, and Stephan (2004) show that this gives good performance (F-scores between 0.65 and 1.0) in the domain for which it was designed (a multi-character virtual environment) – but op den Akker and Traum (2009) show that it gives poor performance in the AMI four-person human-human meeting domain (McCowan et al. 2005), with accuracy falling to 36%, primarily because it does not account for the possibility of group addressing. However, they show that incorporating a group-addressing option and information about visual gaze direction can greatly improve results, with accuracy rising to 65% over all utterances (and 68% over utterances which contain *you*). Defining a gazed-at object as one towards which a speaker's gaze is directed for more than 80% of the duration of an utterance:

(4) a. If utterance contains an address term (vocative etc.) then *Addressee = specified addressee*
b. else if (current speaker = previous speaker) and (gazed-at = previous addressee) then *Addressee = previous addressee*
c. else if (current speaker = previous speaker) and (gazed-at $\neq$ previous addressee) then *Addressee = group*
d. else if (current speaker $\neq$ previous speaker) and (current speaker = previous addressee) then *Addressee = previous speaker*
e. else if (current speaker $\neq$ previous speaker) and (gazed-at $\neq$ null) and (utterance contains *you*) then *Addressee = gazed-at*
f. else if (current speaker $\neq$ previous speaker) and (gazed-at = previous speaker) then *Addressee = previous speaker*

2.2.2 Probabilistic approaches. In contrast, Katzenmaier, Stiefelhagen, and Schultz (2004) take a probabilistic approach to addressee identification. Their task is to detect utterances addressed to a conversational robot in human-human-robot situations. They estimate the visual focus of attention of human speakers from head orientation, using a Bayesian approach to maximize the posterior probability that a certain target is the focus of attention given the observed head orientation (with manually set priors). Using only this visual information, the system detects when the robot is being addressed with 93% accuracy (70% F-score). They also exploit the fact that humans tend to speak differently to artificial agents than to other humans to experiment with linguistic cues derived from language models calculated from utterances addressed to the robot. Their results improve (72% F-score) when visual information is combined with linguistic features.

Jovanovic, op den Akker, and Nijholt (2006) and Jovanovic (2007) focus on face-to-face human-human meetings with four participants, using the AMI Meeting Corpus. To predict which participant is the addressee of each dialogue act, they use a Bayesian Network classifier trained on several multimodal features, including visual features such as gaze direction, discourse features such as the speaker and dialogue act of preceding utterances, and utterance features such as lexical clues and utterance duration. They found that using a combination of features from various resources improves performance (the best system achieves an accuracy of 77%).

A slightly different task is tackled by op den Akker and op den Akker (2009), who concentrate on deciding whether a remote participant who attends a meeting via tele-

conferencing is being addressed or not. Again using the AMI corpus, they experiment with several types of classifiers and a variety of feature sets. In addition to the kind of features used by Jovanovic (2007), they also investigate the use of manually annotated topic information, exploiting correlations between particular meeting topics such as "industrial designer presentation" and participant roles such as "industrial designer". Topic and role features improve results significantly for this binary task, with their best system (Logistic Model Trees) achieving an accuracy of 93% over a baseline of 89.87%.

2.3 Our own work

In previous work, we investigated the generic/deictic distinction in two-party dialogue (Gupta, Purver, and Jurafsky 2007), and the subsequent problem of individual reference resolution in multi-party dialogue (Gupta et al. 2007), using only linguistic features. In (Frampton et al. 2009) and (Purver et al. 2009) we investigated the addition of visual information and more detailed lexical features respectively. Here, we summarize our approach, and we improve on the previous results, by using a slightly larger dataset and additional features (prosody), and by averaging features over *yous* in multi-*you* utterances. We also describe a simple but effective method for isolating discourse marker *yous* (instead of just leaving them aside as before), and present the first account of a fully automatic, combined system.

3. Data

Our experiments are performed using the AMI Meeting Corpus (McCowan et al. 2005), a publicly available collection of meetings among four participants. The meetings are scenario-driven: participants are given the task of designing a remote control and are instructed to play different roles within a fictitious design team. Each participant plays one of the following roles: industrial designer, user interface designer, marketing expert, or project manager. The conversations during the meetings, however, are not scripted and the interaction is unconstrained. Each meeting lasts around 30 minutes. The corpus includes audio and video recordings, as well as manual orthographic transcriptions and a wide range of manually annotated information, including dialogue acts, visual focus of attention, and addressees.

For our investigation we use a sub-corpus of 984 *you*-utterances (i.e. utterances containing the word *you*) extracted from 10 different AMI meetings involving both native and non-native English speakers. In the remainder of this section, we explain how these utterances are coded.

3.1 Annotations and data set

We annotated the *you*-utterances in our sub-corpus using a three-way scheme that distinguishes between the following semantic classes: *discourse marker*, *generic* and *deictic*, as exemplified in (1) and (2).[2] Annotators did not have access to the meeting videos; only the transcriptions and/or audio files were provided during annotation. To assess the reliability of the annotations, approximately 10% of utterances were coded by two

2 Gupta, Purver, and Jurafsky (2007), using more casual conversational dialogue data, found a further *reported deictic* class in which a *you* use in reported speech refers not to the current addressee, but the original addressee in the speech being reported; we found no such uses in our more task-focussed data.

Table 1
Distribution of *you* interpretations

Generic	435	44.2%
Deictic	477	48.5%
Total generic/deictic	*912*	*92.7%*
Discourse marker	72	7.3%
Total	*984*	*100.0%*

Table 2
Distribution of deictic *you* reference, where AMI addressee annotations available

Singular	294	65.2%
Plural	157	34.8%
Total	*451*	*100.0%*

different individuals. Inter-annotator agreement was fairly good, with a kappa statistic (Carletta 1996) of 84%.[3] Around 7% of instances were tagged as *discourse marker*, with the remaining *you*-utterances (912) being evenly distributed between *generic* and *deictic* interpretations. For the experiments we report here, we exclude the *discourse marker* class (less interesting for our purposes, and easy to distinguish – see below) and concentrate on the *generic* and *deictic* cases. Table 1 shows an overview of the distribution.

We then used the AMI addressee annotations to tag each *deictic* case with further information. These annotations assign to each utterance the addressee(s) – the set of individuals addressed by the utterance. Addressee annotations are not provided for some dialogue act types such as backchannels and fragments.[4] This reduced our useful set of deictic *you*-utterances from 477 to 451 instances (see Tables 1 and 2). According to Jovanovic (2007), the kappa scores for the inter-annotator agreement of the addressee annotation of the whole corpus range from 68% to 81%. Jovanovic (2007) and Reidsma, Heylen, and op den Akker (2008) point out that annotators mainly disagreed on whether an individual or a group had been addressed and had problems distinguishing sub-group addressing from addressing the whole audience. However, when annotators agreed on labelling an utterance as being addressed to an individual, then they also reached high agreement on determining who that single addressee was.

In our set of 451 deictic *you*-utterances, 65% of instances were addressed to a single individual, 33% to the whole audience (three participants in the current scenario), and less than 2% to a sub-group of two participants. This distribution is in line with that reported by Jovanovic, op den Akker, and Nijholt (2006) for the whole AMI corpus. In our experiments, we therefore collapse the two- and three-participant addressee cases into one *plural* class – see Table 2.

Multiple interpretations and/or addressees. Two things are worth mentioning at this point. First, using utterance-level annotations as we do here does not allow us to account for

3 For comparison, Bergsma, Lin, and Goebel (2008) report kappa from 79% to 90% when annotating non-referential *it* in written text, and Müller (2006) up to 65% for the same task in spoken dialogue.
4 See (Jovanovic, op den Akker, and Nijholt 2006) for more details.

utterances containing multiple instances of *you* with different interpretations (e.g. some deictic and some generic, as in the invented example (5a)). While such examples do seem possible, the only cases we encountered in our data were combinations of discourse markers with one other class. No mixtures of generic and deictic, or of singular and plural addressee, were encountered.

Detecting discourse markers, though, seems an almost trivial problem. All discourse markers here have the form *"you know"* or *"you see"*; our dataset contains 131 instances of the bigram *"you know"*, of which only 18 are not discourse markers; and 7 instances of *"you see"*, of which 6 are not discourse markers. Using just two simplistic lexical patterns (*"do/as/if you know/see"* and *"you know how/that"*) we can detect 22 of the 24 non-discourse-marker instances, with no false positives (i.e. 98.6% classification accuracy of *"you know/see"*). It therefore seems reasonable to remove discourse markers from our dataset, and use utterance-level classification.

The second related issue is that the AMI addressee annotations also operate at an utterance level. Using them to give us the identity of the referent of deictic *you* therefore assumes that they apply equally to all parts of the utterance; this would fail to account for cases such as the invented (5b), where the two *you*s refer to different individuals, or (5c) where the utterance as a whole is addressed to the group but *you* apparently addresses or refers to a single individual. Such examples would require either a finer-grained (e.g. phrase-level) set of addressee annotations than AMI provides, or allowing deictic *you* reference to be annotated to a subset of the addressees. However, we did not encounter any such examples in our data, so we leave this aside here. Future work, though, may need to consider both these issues.

(5) a. Do you think that when you drink coffee your blood pressure rises?
b. You and you got great grades in the exam.
c. Tomorrow we could discuss the summary that you sent us.

Of course, many utterances contain multiple *you*s of the *same* class; when calculating features which relate to the individual word (or e.g. its prosody – see below), we average over all *you* instances (after removing discourse markers as described above).

3.2 Labelling scheme for listeners

For those *you*-utterances tagged as singular deictic, we now require a labelling scheme for distinguishing between the three potential addressees (the listeners) who may be the referent of the pronoun. One possibility is to use the scheme employed by Gupta et al. (2007), who label potential addressees in terms of the order in which they speak after the *you*-utterance. Hence, the potential addressee who speaks next is labeled 1, the potential addressee who speaks after that is 2, and the remaining participant is 3. Label 4 is used for group addressing. An advantage of this scheme is that it is domain and setting independent. However, one obvious weakness is that it cannot be used by a real-time *you*-resolution module. Another weakness is that in our dataset (and presumably in general), we end up with a skewed class distribution because the next speaker is often the intended addressee: in our data, the next speaker is the intended addressee 41% of the time, and 38% of instances are plural, leaving only a small percentage for the remaining two classes.

We therefore experimented with two alternative labelling schemes which do not suffer from these weaknesses. The first scheme identifies meeting participants according to their AMI meeting role: industrial designer (ID), marketing expert (ME), project manager (PM), and user-interface designer (UI). As individuals keep their roles throughout

Table 3
Distribution of addressees for singular *you* with two different labelling schemes

	Total	Position rel. to speaker			Individual participant roles			
		L_1	L_2	L_3	ID	ME	PM	UI
percentages	100%	35.0%	30.3%	34.7%	27.6%	24.1%	17.7%	30.6%
raw numbers	294	103	89	102	81	71	52	90

meetings (and in fact, series of meetings) each individual is therefore identified by the same label throughout a meeting. We hypothesized that this would be advantageous when using lexical features due to individual or role-dependent vocabularies. This scheme is similar to that used by Jovanovic (2007) who encodes participants by their absolute seating position rather than their role; this is similarly fixed throughout a meeting, but does not exploit the common roles played (and topics discussed) across meetings.

Our second labelling scheme identifies the potential addressees by their seating position *relative to the current speaker*. The AMI meeting setting includes a rectangular table with two participants seated at each of its opposite longer sides, and absolute seating positions stay constant during a meeting. Thus for a *you*-utterance, we label listeners as either L_1, L_2 or L_3 depending on whether they are sitting diagonally, laterally or opposite from the current speaker. Under this labelling scheme, features which depend on an individual participant are now generated for only 3 instead of 4 individuals. Hence, the feature space is smaller, which is a potential advantage for automatic classification. In contrast to the previous scheme, individuals will not keep the same labels throughout a meeting, so we also include meeting role features which encode the role of the participant currently sitting in each relative position, in order to continue to allow the classifier to exploit the role-dependent lexical features. Table 3 shows the resulting class distributions under both schemes.

4. Feature Extraction

The features which we extracted can be divided into three broad categories: transcript features, prosodic features and visual features. Transcript features, as the name suggests, are extracted from the meeting's transcript and are the most numerous; prosodic features are extracted directly from the audio signal; and visual features are based on the gaze direction of meeting participants.

In the experiments described in Section 5, we test both *manual* and *automatic* systems. For an automatic system, all of the features are derived through entirely automatic means, whereas for a manual system, they are extracted from gold-standard transcriptions and annotations. We describe the transcript features first.

4.1 Transcript features

Table 4 summarizes the transcript features. These are divided into *you*-utterance features and Backward Looking (BL)/Forward Looking (FL) features. The former are features of the *you*-utterance itself (the utterance containing the *you* which we wish to resolve), while the latter are features which depend on context, looking backwards/forwards

from the *you*-utterance. Transcript features are extracted from manual transcripts for a manual system, and from ASR transcripts for an automatic system.

4.1.1 Features of the *you*-utterance. The *you*-utterance features are further sub-divided into three sub-categories: *sentential*, *lexical* and *dialogue act (DA)* features. The sentential features encode structural, durational, lexical and shallow syntactic patterns, while the lexical features include one feature for each distinct word or n-gram seen more than once in the corpus; both are extracted from the transcript (manual or automatic). The dialogue act (DA) features use the manual AMI DA annotations to represent the conversational function of the *you*-utterance. As no high-accuracy DA tagger is currently available for this data,there are no DA features for the automatic systems.

The sentential features are mainly intended to distinguish deictic from generic *you* (e.g. "do you" questions seem more likely to be deictic, while phrases such as "in general" indicate generic statements), with some distinguishing deictic plural from deictic singular (e.g. "you guys" indicates plural reference); the patterns were chosen on the basis of our linguistic intuition, and may therefore lack some coverage, but should be relatively domain-independent. In contrast, the lexical features are extracted entirely automatically, and will cover all distinctive n-gram patterns, but may be dependent on the AMI domain. These lexical features are intended also to aid individual addressee resolution: as different individuals in the AMI meetings have different roles/areas of expertise, utterances addressed to different individuals may contain distinct vocabularies. The DA feature could be useful for the generic/deictic distinction (questions are more likely to contain deictic *you* than statements); and the AMI person Named Entity tag should help to identify individual addressees.

4.1.2 Backward-Looking/Forward-Looking features. The BL/FL features give information about the dialogue context which surrounds the *you*-utterance; they are divided into three sub-categories: *utterance comparison*, *speaker activity* and *dialogue act* features.

BL and *FL* utterances are utterances in the surrounding context spoken by listeners (potential addressees) of the *you*-utterance; a listener's BL1/FL1 utterance is that individual's first utterance as we look backwards/forwards from the *you*-utterance. The utterance comparison features encode similarities and differences between the *you*-utterance and a BL1/FL1 utterance (e.g. their overlap, separation, and lexical similarity). These features thus resemble those used by Galley et al. (2004) for the related task of identifying the first half of an adjacency pair. Here they are primarily intended to help identify individual addressees: we hypothesize that if a listener is being addressed, then their BL1/FL1 utterance will be similar to, and/or close to, the *you*-utterance.

The speaker activity features tell us about the order and frequency of speaker changes immediately before/after the *you*-utterance. Again, these should help to identify individual addressees—the first new speaker following the *you*-utterance is often the intended addressee (see Section 3.2). Frequent speaker changes may also indicate that meeting participants are directly engaging one another, and hence that uses of *you* are more likely to be deictic.

DA features are AMI DA tags as above, but here refer to the contextual BL/FL utterance rather than the *you*-utterance. We expect these features also to be indicative of individual addressee, especially when considered in combination with the DA of the *you*-utterance (e.g. in question-answer pairs): forward-looking DAs (such as "question") are likely to influence the addressee to speak next, while backward-looking acts (such as "answer") might address a recent speaker.

Table 4
Summary of the transcript features. A listener's BL1/FL1 utterance is that individual's first utterance as we look backwards/forwards from the *you*-utterance.

Category	Sub-category	Description/Examples
You-utterance	Sentential	# of *yous*, # of words, duration, speech rate, 1st person pronoun, participant Named Entity, phrasal patterns s.a.: *you guys*, *as you can see*, `auxiliary` *you*, `wh-word` *you*, *in general|always|often*
	Lexical	One feature per distinct word/n-gram seen more than once in the corpus.
	Dialogue act	DA of the *you*-utterance.
BL/FL	Utterance comparison	overlap?, duration of overlap, time separation, ratio of common words, # of utterances between, # of speakers between.
	Speaker activity	# of speakers during previous 5 utt., # of speakers during next 5 utt., past/future speaker order
	Dialogue act	DA of the BL1/FL1 utt.

Table 5
Summary of the prosodic features extracted for each *you*, normalized by the speaker.

Category	Features
Pitch	Minimum, maximum, median, mean, standard deviation, average local variability, # of voiced frames.
Intensity	Minimum, maximum, median, mean.

4.2 Prosodic features

Following preliminary results by Gupta, Purver, and Jurafsky (2007) indicating that the average pitch of generic uses of *you* tended to be lower than that of deictic uses, we also extract some acoustic and prosodic features. For example, we speculate that deictic *you* is more likely to be stressed, and that this will be reflected in its pitch and intensity. Hence we used Praat (Boersma and Weenink 2010) in order to extract acoustic features for each *you*, normalizing each feature by the speaker. These features are summarized in Table 5 and can be divided into two categories: those which relate to pitch, and those which relate to intensity. For both pitch and intensity, we extract the minimum, maximum, median and mean values over the word *you* itself; for pitch, we also extract the standard deviation, average local variability and the number of voiced frames. The average local variability in pitch is the mean absolute change in pitch between adjacent points on the pitch curve. We add a second version of this feature where local pitch changes can never be larger than half an octave. Note that for multi-*you* utterances, we average the prosodic features over the *yous*.

4.3 Visual features

It would be surprising if information about gaze direction was not helpful in resolving *you* in face-to-face conversation. As pointed out by Jovanovic (2007) and others, gaze direction, particularly that of the speaker, is a rather useful clue for identifying individual addressees. Here we want to test whether visual information is also useful for distinguishing generic from deictic *you*, and deictic plurals from deictic singulars. In this section, we describe what our visual features represent and how they are computed.

4.3.1 Visual information used. For a manual system, we can use the "Focus of Attention" (FOA) annotations provided by the AMI corpus, which track meeting participants' head orientation and eye gaze during a meeting. For an automatic system, though, we must generate automatic equivalents from the corpus video. We use the methodology and automatically derived data presented in Frampton et al. (2009). This automatic visual information was extracted using a 6 degree-freedom head tracker, which produces a gaze probability matrix for each frame. Gaze probability $G(i,j)$ is defined as

$$G(i,j) = G_0 e^{-{\alpha_{i,j}}^2/\gamma^2}$$

where α_i is the angular difference between the gaze of individual i and the direction defined by the location of i and the *target* j. The *target* j can be any of the meeting participants or the whiteboard/projector screen in the meeting room. G_0 is a normalization factor such that $\sum_j G(i,j) = 1$ and γ is a user-defined constant (here 15 degrees). We judge that i is gazing at j if the probability is above a certain threshold.[5]

4.3.2 Highest Gaze Duration Proportion (GDP) features. Table 6 summarizes our visual features which are computed from the visual information described above. These features are based on Gaze Duration Proportion (GDP) values, which are similar to the "Degree of Mean Duration of Gaze" values described by Takemae, Otsuka, and Mukawa (2004). A GDP value indicates the proportion of a particular time period for which an individual's gaze is directed at a particular target. As already stated, possible targets for an individual's gaze here are the other meeting participants and the whiteboard/projector screen, and the time periods which we consider are the duration of the whole utterance, of each third, and ± 2 seconds from the start time of the *you*. We compute GDP values for all of the possible combinations of meeting participants, targets and time periods, and then from these GDP values, we derive highest GDP features for each individual [6]. These features indicate where the individual's gaze was directed the longest over each of the different periods of time. In addition, we compute a highest GDP mutual gaze feature for the speaker. This indicates with which other individual, the speaker spent most time engaged in a mutual gaze over the course of the whole *you*-utterance.

Finally, two further features give some indication of the amount of "looking around" that the speaker does during a *you*-utterance: the ratio of the second-highest GDP to the highest, and the ratio of the third-highest to the highest. We hypothesize

5 We found a threshold of 0.6 to give the best kappa scores when comparing to the AMI FOA annotations.
6 Note that for multi-you utterances, we compute the highest GDP value for the combination of the periods that are ± 2 seconds from the start time of each *you*.

Table 6
Visual Features

Category	Features
Computed for each participant	Target with the highest GDP for the whole utterance, the first third of the utterance, the second third of the utterance, the last third of the utterance, and $\pm$2 seconds from the *you* start time.
Computed for the speaker	Participant in mutual gaze with the speaker. Ratio 2nd hyp. target / 1st hyp. target. Ratio 3rd hyp. target / 1st hyp. target.

that the speaker will look around more in utterances with plural addressees, and so these features will help to distinguish deictic plurals from deictic singulars.

5. Experiments

This section presents our experiments and results; as stated in Section 4, we implement classifiers for both *manual* and *automatic you*-resolution systems. Recall that an automatic system's features are computed from ASR and head-tracker output, while a manual system's features come from manual transcripts and the AMI Focus of Attention (FOA) annotations.

Sections 5.1 to 5.3 describe a first set of experiments in which we divide the problem of resolving the pronoun *you* into three sub-tasks: distinguishing generic from deictic usages, distinguishing deictic singulars from deictic plurals, and identifying the addressee of the deictic singulars. These experiments allow us to assess the difficulty of each subtask and to conduct detailed analysis of feature contribution. Since a full computational *you*-resolution module needs to treat all sub-tasks, we then go on to conduct a second set of combined experiments (Section 5.4), first with a single multi-class classifier for all of the different uses of *you*, and then with a cascaded sequence of three classifiers, one for each sub-task.

All of our classifiers are Bayesian Networks and we evaluate their performance in 10-fold cross-validations. In all experiments, we compare our results with a majority class (MC) baseline; in the individual addressee sub-task, we also compare with a *next speaker* (NS) baseline which always selects the next new speaker as the intended addressee. We assess the contribution of different feature types by comparing classifier performance in their presence/absence, and in addition, by using *information gain* to measure their level of correlation with the *you* type. We give results for the transcript, prosodic and visual feature sets separately, and for a combined *multi-modal* (MM) system which uses all feature types. Results are also shown without DA features ("- DA"), which are not likely to be reliably available in any automatic system;[7] and without the FL features which might be problematic for real-time systems ("- FL"). Note that while

7 Results for "-DA" are included only for our manual systems. As mentioned in Section 4.1.1, our automatic classifiers do not use DA features at all.

real-time performance might be essential in some systems (e.g. on-line translation, or dialogue systems used in teams), it is not necessarily a requirement in others (e.g. a post-meeting minute summarizer). For each sub-task, we also perform an error analysis on the output of our best classifier, training and testing on the full data set.

5.1 Generic versus deictic uses of *you*

Our hypothesis as regards the generic/deictic distinction was that the most useful features would be the *you*-utterance (sentential, lexical and prosodic) features. Generic uses of *you* occur within sentences with recognisable syntactic and lexical forms; but are unlikely to be associated with particular turn-taking patterns or gaze directions.

Table 7 summarizes the results for the manual system for this task, and Table 8, the results for the automatic system. For the manual system, all feature sets produced a statistically significant improvement over the majority class baseline. The best systems give 87.2% accuracy in the manual case, and 83.6% accuracy in the automatic case. The generic class F-score reaches highs of .87 and .83 for the manual and automatic systems respectively, and the deictic class, .88 and .85.

Table 7
Generic *vs.* deictic uses, manual system.

Features	Acc	F1-Gen	F1-Deic
MC Baseline	52.3	0	.69
Transcript exc. lexical	79.9	.80	.80
Transcript inc. words	84.5	.84	.85
Transcript inc. 3grams	86.5	.86	.87
Transcript inc. 3grams - DA	86.7	.87	.87
Transcript inc. 3grams - FL	87.5	.87	.88
Prosodic	57.7	.45	.66
Visual	59.4	.62	.57
MM exc. lexical	78.8	.78	.80
MM inc. words	84.0	.83	.84
MM inc. 3grams	87.1	.87	.87
MM inc. 3grams - DA	85.5	.85	.86
MM inc. 3grams - FL	87.0	.86	.88

5.1.1 Feature contribution. As expected, linguistic features are the most useful for this task. Sentential features are amongst the best predictors, especially those which refer to surface lexical properties of the *you*-utterance. For instance, a positive value for the feature "generic expressions", which indicates that the utterance contains words such as *often* or *always*, was useful for detecting generic uses of *you*, as was the number of *you* pronouns within the utterance (with higher numbers increasing the likelihood of generic interpretations). The presence of one or more first person pronouns or of a Named Entity tag referring to one of the participants was predictive of deictic interpretations.

DAs also provide valuable information. As pointed out by Gupta, Purver, and Jurafsky (2007) and Gupta et al. (2007), *you* pronouns within question DAs are strongly associated with deictic uses (in our data set, 136 out of 155 *you*-utterances tagged

Table 8
Generic *vs.* deictic uses, automatic system.

Features	Acc	F1-Gen	F1-Deic
MC Baseline	53.6	0	.70
Transcript exc. lexical	75	.76	.74
Transcript inc. words	81.1	.81	.81
Transcript inc. 3grams	83.6	.82	.85
Transcript inc. 3grams - FL	83.0	.81	.85
Prosodic	59.3	.37	.70
Visual	56.6	.51	.61
MM exc. lexical	76.1	.76	.76
MM inc. words	79.0	.79	.79
MM inc. 3grams	82.3	.81	.84
MM inc. 3grams - FL	81.0	.79	.82

with one of the `elicit` AMI DA tags have a deictic interpretation). Other DA tags such as `inform` are more likely to correlate with generic uses of *you* (245 out of 357 *you*-utterances tagged as `inform` contain generic *you* pronouns) but the association is weaker and hence not as useful.

Lexical features were also very useful, with 3-grams producing better results than simply using words (86.5 vs. 84.5 for the manual system and 83.6 vs. 81.1 for the automatic one, with $p < 0.05$). Since, to some extent, n-grams such as *do you* can capture interrogative structures, lexical features provided clues that had similar predictive power to the most critical information contributed by dialogue acts, namely the presence of questions. Indeed, when 3-grams are used, ignoring DAs does not decrease accuracy. Interestingly, we also found that generic uses of *you* are more likely to appear in utterances containing words related to the main meeting topic, such as *button*, *channel* or *volume*, which refer to properties of the to-be-designed remote control. On the other hand, words related to meeting management, such as *presentation*, *email*, *project* and *meeting* itself, were predictive of deictic uses, as was the presence of discourse and politeness markers such as *okay*, *please* and *thank you*.

Prosodic features perform well. A system that uses only prosody is able to beat the MC baseline (57.7 vs. 52.3, $p < 0.05$). According to information gain, the most predictive prosodic features are mean and median pitch, followed by minimum pitch and average local variability. It seems that as these features increase in value, the *you* is more likely to be deictic than generic. Note that Gupta, Purver, and Jurafsky (2007) found that average pitch was higher in deictic uses.

As expected, the use of contextual BL and FL information did not improve accuracy for this task. Somewhat surprisingly, however, visual features relating to the listeners' gaze were predictive, allowing the visual features alone to beat the baseline in both manual and automatic systems (59.4 and 56.4, respectively, with $p < 0.05$ w.r.t. the baselines). If listeners' gaze direction is mostly toward the white-board/projector screen instead of another individual while an utterance containing *you* is uttered, then the pronoun is more likely to be deictic (in particular, as we shall see in Section 5.2, it is more likely to have a plural referent).

However, in both manual and automatic systems neither prosodic nor visual features improved accuracy significantly for this task when combined with linguistic information.

5.1.2 Error analysis. Using our best classifier training and testing on the full data set yields 94.8% accuracy, giving a total of 47 errors. We found that around a third of these errors were highly ambiguous when looking at the utterance alone without information from the surrounding context – see e.g. the fragmentary utterance in (6a) and the ambiguous example in (6b), which had been annotated as generic but was incorrectly labelled as deictic by the classifier. This indicates that a more detailed model of dialogue context than that provided by our BL and FL features is needed to treat these examples.

Another important source of errors are utterances that require extra-linguistic knowledge about the situation and/or the world. Lexical information can help here, but, not surprisingly, several errors remain – see e.g. examples (6c) and (6d), which were wrongly labelled as generic.

Around 20% of errors stem from the fact that the classifier tends to label all questions (and utterances with particular n-grams such as *do you* and others containing *wh*-words) as deictic, which in general helps to improve accuracy (see above). However, we do find generic uses of *you* in questions as well, as shown in examples (6e) and (6f).

(6) a. you you -disfmarker-
b. Or maybe you want to phone him.
c. So if you want to just go straight to the second slide
d. Here on the left-hand side, you can see a remote control that has lots and lots of buttons
e. How do you wear this thing?
f. Um, how many do y do you need, solar cells?

5.2 Singular versus plural deictic uses

In the task of distinguishing singular and plural uses of *you*, we hypothesized that while sentential and visual features might be of some use, the main contributor would be information about dialogue structure: addressing the group might be expected to result in speaker activity distributed amongst the addressees, in which all make relevant contributions (e.g. answering a question posed to the group).

Table 9 summarizes the results for the manual system, and Table 10, the results for the automatic system. The best manual system achieves 86.5% accuracy and the best automatic system nearly 87%.

5.2.1 Feature contribution. As expected, for this task the utterance comparison features which encode information about dialogue structure are amongst the most useful transcript features. Utterances by individual addressees tend to be more lexically cohesive with the *you*-utterance; when features such as the "ratio of common words" feature indicate a low level of lexical similarity, plural addressing is more likely. Speaker activity features are also useful cues: the presence of only one single speaker in the previous and/or following five utterances correlates with plural addressing, while singular addressing seems to lead to more immediate interaction between the speaker of the *you*-utterance and the individual addressed. Dialogue Act information is also useful: individual addressees tend to acknowledge utterances using backchannels, while it is less common to find backchannels adjecent to the *you*-utterance when the pronoun

Table 9
Singular *vs.* plural deictic, manual system.

Features	Acc	F1-Sing	F1-Plural
MC Baseline	65.2	.79	0
Transcript exc. lexical	76.0	.82	.65
Transcript inc. words	81.6	.86	.72
Transcript inc. 3grams	86.5	.90	.79
Transcript inc. 3grams - DA	84.5	.94	.66
Transcript inc. 3grams - FL	85.4	.89	.78
Visual	68.5	.77	.49
MM exc. lexical	76.0	.82	.65
MM inc. words	82.0	.86	.74
MM inc. 3grams	84.7	.88	.78
MM inc. 3grams - DA	84.7	.89	.76
MM inc. 3grams - FL	84.5	.88	.77

Table 10
Singular *vs.* plural deictic, automatic system.

Features	Acc	F1-Sing	F1-Plural
MC Baseline	63.7	.78	0
Transcript exc. lexical	69.5	.78	.49
Transcript inc. words	85.6	.89	.79
Transcript inc. 3grams	86.5	.90	.78
Transcript inc. 3grams - FL	86.5	.90	.77
Visual	61.9	.69	.50
MM exc. lexical	69.5	.78	.51
MM inc. words	85.2	.89	.77
MM inc. 3grams	85.2	.89	.77
MM inc. 3grams - FL	85.2	.89	.77

refers to more than one individual. Statements show a stronger correlation with plural addressees, while questions—which in Conversation Analytic terms (Sacks, Schegloff, and Jefferson 1974) are DAs often used to "select the next speaker"—tend to be addressed to individual participants and lead to a speaker change.

However, despite these correlations, similarly to the generic *vs.* deictic task taking away DA or FL information has no significant effect on performance (the small differences observed in Table 9 are not statistically significant). Again this seems due to the fact that lexical features are to some extent able to make up for the absence of higher level information, e.g. with n-grams that are indicative of question DAs such as *you do* or of acknowledgement such as *okay*. Indeed lexical features improve performance significantly: from 76 in the manual system and 69.5 in the automatic one up to 86.5 in both systems when 3-grams are used ($p < 0.05$). Other predictive n-grams for this task are *you mean* and *you know*, which are indicative of singular and plural deictic *yous*, respectively. We also find that plural first person pronouns such as *we* correlate with plural deictic *yous*, as does the verb *see* involved in constructions such as *as you can*

see, which in the current setting are commonly used to address the whole audience. Other sentential features contribute too: for instance, the presence of more than two *you* pronouns seems to correlate with plural reference.

Overall the visual features do not improve performance for the task of distinguishing between singular and plural deictic *you*: the results obtained with the visual-only manual system (68.5%) are not significantly higher than the MC baseline (65.2%) and combining visual and linguistic information does not yield better results than using only transcript features. However, as in the generic *vs.* deictic task, the white-board/projector screen value for the listeners' gaze features seems to have discriminative power—when listeners' gaze is directed at this target, it is often indicative of a plural rather than a singular *you*. It seems then, that in our data-set, the speaker often uses the white-board/projector screen when addressing the group, and hence draws the listeners' gaze in this direction. We should also note that the ratio features which we thought might be useful here (see Section 4.3.2) were not.

5.2.2 Error analysis. We run our best classifier on the full data set, obtaining 88.2% accuracy and a total of 53 errors. Of these, 33 were false positives for the majority class, i.e. singular. 15% of these errors were cases such as (7a), where the *you*-utterance is addressed to a (plural) subset of participants but not to the whole group.[8] It is not surprising that these utterances are problematic for the classifier since they were also a source of disagreement between annotators (see Section 3.1).

Questions were another cause of errors. Half of the false positives for the singular class are utterances tagged with question-related DAs (often containing the bi-gram *do you*). If we look at the full data set, around 40% of singular *you*-utterances are tagged with question DAs, while this is true for less than 18% of plural *you*-utterances. Thus the classifier tends to label questions as being addressed to individuals, which in general improves accuracy but also leads to errors for almost half of plural addressee *you*-utterances that are questions, such as (7b).

The *you*-utterances wrongly classified as plural seem to involve features that increase the likelihood of plural addressing. For instance, (7c) contains a plural first person pronoun (*us*) and a reference to the white board during which all participants look at that target—two features that are more common in group addressing, but that also occur in singular *you*-utterances (7c).

(7) a. So what's -disfmarker- what are your ideas about that?
b. Do you think that people like the colour yellow?
c. So I will now ask you [...] for to uh each of us to to draw uh your favourite animal on the white board.

In general, the errors observed seem to require a richer model of dialogue context that presumably can only be achieved with deeper dialogue understanding techniques.

5.3 Detection of individual addressees

We now turn to the task of identifying the referent of singular deictic uses of *you*, which amounts to identifying the addressee of *you*-utterances addressed to a single individual. For this task, a majority class baseline makes little sense (and its performance is poor,

8 Recall that only 2% of *you*-utterances in our data set overall were addressed to a sub-set of dialogue participants—so the 15% found in our error set is disproportionately high.

about 30% accuracy); instead, we evaluate with respect to a *next speaker* (NS) baseline that always selects the next new speaker as the addressee. This simple heuristic is accurate about 70% of the time (see below), but can only be used by offline systems that have access to future context. For online cases, we compare to a *previous speaker* (PS) baseline, selecting the most recent different speaker as the addressee (as in the rule-based approaches described in Section 2.2.1); this gives nearly 60% accuracy.

As explained in Section 3.2, in this task we experiment with two different labelling schemes for potential addressees: a 3-way scheme labelling listeners according to their relative position with respect to the speaker, and a 4-way scheme labelling each individual according to their meeting role: industrial designer (ID), marketing expert (ME), project manager (PM), and user-interface designer (UI).

Here, we hypothesized that explicit lexical and syntactic cues would be weak, but that gaze direction would be a strong predictor, as would subsequent dialogue behaviour (knowing the next speaker is clearly helpful – see above); we also expected general lexical n-gram features to help distinguish between the invididuals on the basis of the different vocabularies associated with their personal topics or areas of expertise. Tables 11 and 12 show the manual and automatic results when using the 4-way scheme (with best systems achieving 83% and 85% accuracy respectively), and Tables 13 and 14 the equivalents for the 3-way scheme (achieving 86% and 78%). In both cases, the performance beats the relevant baselines and exceeds op den Akker and Traum (2009)'s reported 68% accuracy.

Table 11
Addressee detection for singular deictic *yous*, manual system, participants labelled according to meeting role.

Features	Acc	F1-ID	F1-ME	F1-PM	F1-UI
PS Baseline	56.8	.57	.55	.61	.56
NS Baseline	71.4	.73	.72	.73	.69
Transcript exc. lexical	75.2	.79	.74	.76	.73
Transcript inc. words	75.5	.80	.75	.76	.73
Transcript inc. 3grams	75.5	.80	.75	.76	.73
Transcript inc. 3grams - DA	73.8	.80	.70	.72	.73
Transcript inc. 3grams - FL	62.9	.67	.60	.66	.58
Visual	69.0	.72	.69	.60	.74
MM exc. lexical	82.0	.83	.83	.76	.84
MM inc. 3grams	83.0	.85	.84	.76	.84
MM inc. 3grams - DA	83.0	.85	.83	.77	.86
MM inc. 3grams - FL	81.6	.84	.81	.76	.84

5.3.1 Feature contribution. In general, contextual (BL/FL) features are highly predictive, and when FL features are removed, the drop in accuracy is statistically significant ($p < 0.05$). The most predictive BL/FL features are those which encode the order of the previous and next speakers (not surprising given the high accuracy of the NS and PS

Table 12
Addressee detection for singular deictic *yous*, automatic system, participants labelled according to meeting role.

Features	Acc	F1-ID	F1-ME	F1-PM	F1-UI
PS Baseline	59.2	.62	.45	.70	.60
NS Baseline	70.4	.69	.58	.79	.74
Transcript exc. lexical	76.1	.80	.71	.79	.75
Transcript inc. words	78.9	.81	.74	.79	.81
Transcript inc. 3grams	81.0	.82	.74	.84	.84
Transcript inc. 3grams - FL	70.4	.73	.64	.79	.68
Visual	66.2	.68	.68	.70	.61
MM exc. lexical	84.5	.85	.83	.89	.82
MM inc. 3grams	85.2	.85	.82	.91	.85
MM inc. 3grams - FL	78.2	.86	.73	.83	.72

Table 13
Addressee detection for singular deictic *yous*, manual system, non-speakers labelled according to position relative to speaker.

Features	Acc	F1-L_1	F1-L_2	F1-L_3
PS Baseline	57.1	.60	.58	.53
NS Baseline	71.4	.69	.74	.72
Transcript exc. lexical	73.5	.72	.75	.74
Transcript inc. words	73.5	.72	.75	.74
Transcript inc. 3grams	73.5	.72	.75	.74
Transcript inc. 3grams - DA	73.5	.72	.74	.74
Transcript inc. 3grams - FL	63.6	.65	.62	.63
Visual	76.2	.82	.72	.75
MM exc. lexical	84.7	.87	.83	.84
MM inc. words	84.7	.86	.83	.85
MM inc. 3grams	86.4	.88	.86	.86
MM inc. 3grams - DA	85.7	.88	.85	.84
MM inc. 3grams - FL	80.6	.86	.76	.80

baselines)[9]. Other useful BL/FL features include the number of utterances between the BL1/FL1 utterance and the *you*-utterance, the time separation, and the ratio of common words, indicating that the utterances spoken by the addressee are often very close in time to the *you*-utterance, and are lexically similar. In contrast, information about DAs is not predictive for this task since any DA type can be used to address any participant.

As expected, and in contrast to the two previous tasks, visual features are very useful cues. With either labelling scheme, a manual system which uses only visual features performs signficantly better ($p < 0.05$) than the PS baseline. All of the visual features

9 Our data contains 1 utterance with no previous speaker (i.e. only 1 person has spoken so far in the meeting), and for this case, the PS baseline hypothesizes a different addressee depending on the labelling scheme, thus producing very slightly different overall accuracy scores.

Table 14
Addressee detection for singular deictic *yous*, automatic system, non-speakers labelled according to position relative to speaker.

Features	**Acc**	**F1-L_1**	**F1-L_2**	**F1-L_3**
PS Baseline	59.9	.61	.54	.64
NS Baseline	70.4	.72	.68	.71
Transcript exc. lexical	72.5	.74	.71	.72
Transcript inc. words	73.9	.76	.74	.72
Transcript inc. 3grams	73.9	.76	.73	.72
Transcript inc. 3grams - FL	64.8	.66	.61	.67
Visual	64.1	.74	.57	.61
MM exc. lexical	78.2	.79	.82	.73
MM inc. words	78.2	.79	.82	.73
MM inc. 3grams	76.8	.78	.78	.73
MM inc. 3grams - FL	66.2	.75	.62	.61

have some degree of predictive power apart from the ratio features. The speaker's gaze direction is the most predictive clue, confirming the results of Jovanovic (2007). In general, whomever the speaker spends most time looking at or engaged in a mutual gaze with is more likely to be the addressee.

The results obtained with the two labelling schemes for listeners differ mostly with respect to the relative impact of the lexical and visual features. For the 4-way role scheme, lexical features are among the best predictors; and as expected, items related to the participant roles help to detect addressees. For instance, the n-grams *sales*, *to sell* and *make money* correlated with utterances addressed to the "marketing expert", while utterances containing *speech recognition* and *technical* are addressed to the "industrial designer". Lexical and transcript features alone do beat the NS baseline (or for the -FL case, the PS baseline), and in both cases the improvements are statistically significant ($p < 0.05$). When visual features are added, performance improves substantially; but visual features on their own fail to beat the PS baseline, achieving only 69.0% accuracy.

As hypothesized, the 3-way position scheme seems less effective in exploiting lexical features; although the results for lexical and transcript features alone do appear to exceed the NS baseline, this is not statistically significant. However, it may be better suited to encoding visual information: the accuracy obtained with visual features alone (76.2% in Table 13) is significantly better than with 4-way encoding (69.0% in Table 11), although the improvement above the NS baseline is still not statistically significant. When lexical and visual features are combined, this scheme obtains our best accuracy of 86.4% – a 15% absolute improvement over the NS baseline – and the accuracy of the manual system is still high even in the absence of FL information (80.6%).

We thus conclude that visual information is critical for this task and that the relative position scheme may encode it best. However, the 4-way scheme seems more robust to noisy data—automatic system scores (Table 12) are actually higher than the manual system (Table 11), and significantly higher than the automatic system which uses the position scheme (Table 14). We hypothesize that this is due to the relative robustness of our ASR data (more effective with the 4-way scheme) and visual data (more effective with the position scheme).

5.3.2 Error analysis. Running our best multimodal classifier with the 3-way position scheme, training and testing on the full data set of deictic singular *you*-utterances yields an accuracy of 93%, giving a total of 20 errors. All errors correspond to utterances where the next speaker is wrongly classified as the addressee. We observe that the most predicitve visual features—the speaker's gaze direction and the participant in mutual gaze with the speaker—had uninformative values for these utterances. Around 50% of errors derive from utterances where the speaker did not clearly gaze at any participant (speaker visual features with value 0) while in all cases the mutual gaze feature had a null value. Thus, since the identity of the next speaker is such an important cue, in the absence of other determining information (such as highly predictive visual features) the classifier tends to select that participant as the most likely referent of *you*.

Some of these errors could be solved by using domain and situation knowledge. For instance, a more knowledgable model would presumably have been able to infer that the utterance in (8) are addressed to the participant currently in control of the slides, possibly relying on the fact that the visual features indicate that the speaker is looking towards the screen. However, since our current classifier is not able to make such inferences, it instead selects the participant who is both the previous and next speaker but who is not addressed in any of these cases.

(8) Can you go to the next slide?

5.4 A single system for classifying all uses of *you*

A full computational *you*-resolution module would need to treat all tasks (either simultaneously as one joint classification problem, or as a cascaded sequence) with inaccuracy in one task necessarily affecting performance in another. We examine this here. Given the superior performance in singular addressee resolution produced by the 3-way position-based encoding for our manual system, we use that approach here.

Table 15
Single multi-class classifier, manual system.

Features	Acc	F1-Gen	F1-Plural	F1-L_1	F1-L_2	F1-L_3
MC Baseline	47.7	.65	0	0	0	0
PS Baseline	18.4	0	0	.28	.29	.26
NS Baseline	23.0	0	0	.34	.37	.34
Transcript exc. lexical	58.6	.74	.32	.50	.52	.48
Transcript inc. 3grams	68.1	.84	.52	.55	.52	.47
Visual	49.4	.64	.25	.48	.48	.15
MM exc. lexical	61.2	.71	.44	.63	.56	.55
MM inc. 3grams	71.1	.83	.54	.69	.62	.56
MM inc. 3grams - DA	67.9	.80	.51	.64	.56	.55
MM inc. 3grams - FL	67.7	.83	.51	.63	.45	.51

Tables 15 (manual) and 16 (automatic) summarize the results for a single multi-class classifier which is trained to classify all of the different usages of *you* simultaneously. Overall accuracy exceeds a majority-class baseline (the generic class) and the next-speaker and previous-speaker baselines in both cases, but is not particularly high—using the full multimodal feature set, the best manual and automatic systems achieve

Table 16
Single multi-class classifier, automatic system.

Features	Acc	F1-Gen	F1-Plural	F1-L_1	F1-L_2	F1-L_3
MC Baseline	46.4	.63	0	0	0	0
PS Baseline	19.3	0	0	.33	.27	.28
NS Baseline	22.7	0	0	.36	.36	.31
Transcript exc. lexical	53.4	.71	.27	.47	.40	.34
Transcript inc. 3grams	66.1	.78	.61	.48	.45	.33
Visual	35.6	.47	.18	.34	.32	.33
MM exc. lexical	51.6	.66	.31	.52	.40	.38
MM inc. 3grams	66.1	.78	.61	.48	.45	.33
MM inc. 3grams - FL	66.1	.78	.61	.48	.45	.33

Table 17
Cascaded classification, manual system.

Features	Acc	F1-Gen	F1-Plural	F1-L_1	F1-L_2	F1-L_3
MC Baseline	47.7	.65	0	0	0	0
PS Baseline	18.4	0	0	.28	.29	.26
NS Baseline	23.0	0	0	.34	.37	.34
Transcript exc. lexical	64.9	.80	.48	.53	.51	.53
Transcript inc. 3grams	73.1	.86	.68	.54	.55	.68
Visual	47.4	.62	.13	.45	.29	.35
MM exc. lexical	66.6	.78	.49	.52	.69	.62
MM inc. 3grams	78.2	.87	.62	.71	.81	.77
MM inc. 3grams - DA	76.3	.85	.67	.70	.71	.66
MM inc. 3grams - FL	75.5	.86	.69	.82	.74	.50

71.1% and 66.1% accuracy respectively. For the generic class, F1 scores are quite good, but for other classes, they are low.

Cascaded (or *pipelined*) classification, on the other hand, has the advantage of allowing us to exploit the fact that different feature sets work best in different subtasks. As our tasks are of a sequential nature, we can use a sequence of three independent classifiers: first separate generic from deictic cases; then, for deictic cases, separate plural from singular; and finally, apply addressee detection for the singular cases. Table 17 summarizes the results for a manual system which uses this cascaded approach, and Table 18, the results for an equivalent automatic system. Accuracy improves greatly in both, rising to 78.2% for the manual system and 71.8% for the automatic. The same is true for the individual class F1-scores; for example, in the manual system, the F1-score for each class is now between 0.62 and 0.87. However, class F1-scores are lower and more variable in the automatic system.

6. Conclusions

As we have explained, the English second-person pronoun *you* is very frequent in spoken dialogue and has distinct usages. The ability to resolve the meaning of *you*

Table 18
Cascaded classification, automatic system.

Features	Acc	F1-Gen	F1-Plural	F1-L_1	F1-L_2	F1-L_3
MC Baseline	46.4	.63	0	0	0	0
PS Baseline	19.3	0	0	.33	.27	.28
NS Baseline	22.7	0	0	.36	.36	.31
Transcript exc. lexical	60.5	.76	.33	.52	.55	.42
Transcript inc. 3grams	70.7	.82	.69	.60	.68	.48
Visual	40.2	.51	.28	.29	.43	.39
MM exc. lexical	60.9	.76	.34	.48	.66	.51
MM inc. 3grams	71.8	.82	.70	.66	.74	.47
MM inc. 3grams - FL	65.9	.79	.70	.68	.57	.31

can be essential for systems which engage in or support spoken dialogue. In two-party dialogue, this is a matter of detecting discourse markers (a relatively simple task) and distinguishing generic from deictic uses (rather less simple). Multi-party dialogue brings the additional challenge of determining the individual (singular) or group (plural) addressee referred to by deictic uses.

We have explained how both manual and automatic classification systems for resolving *you* in multi-party human-human spoken dialogue can be implemented using Bayesian Networks. The manual systems use features derived from gold-standard manual transcripts and annotations, while the automatic systems use equivalent features derived by entirely automatic means. Some systems can perform resolution in real-time, while others cannot because they use features which are derived from context which occurs after the *you*-utterance (so-called forward-looking features). One limitation of our classification systems is that they do not deal with multiple *you*s within a single utterance which have different deictic/generic meanings. However, our data contained no such utterances, suggesting that they are rare.

The features which we have investigated are wide-ranging, and include lexical, prosodic and visual gaze direction features. In our first experiments, we divided the resolution problem into three sub-tasks, namely distinguishing generic versus deictic usages, then deictic singular versus deictic plural, and finally addressee resolution for deictic singulars. Different features were more useful in different tasks: the generic/deictic distinction seems primarily to be expressed by linguistic means, and can be captured using features based primarily on the words in the *you*-utterance itself; while individual addressee resolution requires knowledge either of dialogue context (the surrounding speaker activity) or extra-linguistic information (participant gaze).

The best accuracy results for manual systems in the three sub-tasks are respectively, 87.5%, 84.9% and 86.4%. and for automatic systems, they are 83.6%, 87.9% and 85.2%. These scores were all well above the baselines (majority class and in the case of individual addressee resolution, a next-speaker baseline), and seem high enough to be useful in practical applications.

Since a full computational *you*-resolution module would need to treat all tasks, we then examined combined approaches, and showed that a cascade (pipeline) of independent sub-task classifiers outperforms a single multiclass classifier, as it can exploit different features for the different sub-tasks. The best overall accuracy scores for

manual and automatic systems are 78.2% and 71.8% – again, outperforming the relevant baselines.

Acknowledgements

The authors would like to thank Mario Christoudias, Trevor Darrell, David Demirdjian, Patrick Ehlen, Surabhi Gupta, Dan Jurafsky, and John Niekrasz, who all made significant contributions to this work in its earlier stages. This work was supported by the CALO project (DARPA grant NBCH-D-03-0010), the Dutch NWO VENI project 275-80-002, and the UK ESRC DynDial project (RES-062-23-0962).

References

Arstein, Ron and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06)*, pages 56–63, Potsdam, Germany.

Baldwin, Tyler, Joyce Y. Chai, and Katrin Kirchhoff. 2010. Hand gestures in disambiguating types of *you* expressions in multiparty meetings. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, pages 306–313, University of Tokyo, Japan.

Beeferman, Doug, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of ACL*, pages 10–18.

Boersma, Paul and David Weenink. 2010. Praat: doing phonetics by computer (version 5.1.29). Available from http://www.praat.org/. [Computer program].

Boyd, Adriane, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Byron, Donna. 2004. *Resolving pronominal reference to abstract entities*. Ph.D. thesis, University of Rochester, Department of Computer Science.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–255.

Evans, Richard. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45 – 57.

Frampton, Matthew, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is you? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Lingusitics (EACL)*, pages 273–281, Athens, Greece.

Galley, Michel, Kathleen McKeown, Julia Hirschberg, and Elizabeth 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)*.

Gupta, Surabhi, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007. Resolving "you" in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

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

Hiatt, Laura and Lawrence Cavedon. 2005. Enabling spoken dialogue interaction about team activities. In *1st International Workshop on Multi-Agent Robotic Systems*.

Holmes, Janet. 1998. Generic pronouns in the Wellington corpus of spoken New Zealand English. *Kōtare*, 1(1).

Jovanovic, N., H.J.A. op den Akker, and A. Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23. ISSN=1574-020X.

Jovanovic, Natasa. 2007. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. Ph.D. thesis, University of Twente, Enschede, The Netherlands.

Jovanovic, Natasa, Rieks op den Akker, and Anton Nijholt. 2006. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*, pages 169–176, Trento, Italy.

Jurafsky, Daniel, Alan Bell, and Cynthia Girand. 2002. The role of the lemma in form variation. In C. Gussenhoven and N. Warner, editors, *Papers in Laboratory Phonology VII*. Mouton de Gruyter, Berlin/New York, pages 1–34.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing*. Prentice-Hall, 2nd edition.

Katzenmaier, Michael, Rainer Stiefelhagen, and Tanja Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 144–151, State College, Pennsylvania.

Kilgarriff, Adam. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

McCowan, Iain, Jean 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, the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Meyers, Miriam Watkins. 1990. Current generic pronoun usage: An empirical study. *American Speech*, 65(3):228–237.

Müller, Christoph. 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)*, pages 49–56, Trento, Italy.

Müller, Christoph. 2007. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 816–823, Prague, Czech Republic.

op den Akker, Harm and Rieks op den Akker. 2009. Are you being addressed? - real-time addressee detection to support remote participants in hybrid meetings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 21–28, London, UK, September. Association for Computational Linguistics.

op den Akker, Rieks and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Proceedings of the 13th SemDial Workshop on the Semantics and Pragmatics of Dialogue*, Stockholm, June.

Purver, Matthew, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded lexicalised classifiers for second-person reference resolution. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 306–309, London, UK, September. Association for Computational Linguistics.

Reidsma, D., D. Heylen, and R. op den Akker. 2008. On the contextual analysis of agreement scores. In *Proceedings of the LREC Workshop on Multimodal Corpora*, pages 52–55, Marrakech, Morrocco. ELRA.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Strube, Michael and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan, July. Association for Computational Linguistics.

Takemae, Yoshinao, Kazuhiro Otsuka, and Naoki Mukawa. 2004. An analysis of speakers' gaze behaviour for automatic addressee identification in multiparty conversation and its application to video editing. In *Proceedings of IEEE Workshop on Robot and Human Interactive Communication*, pages 581–586.

Traum, David. 2004. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*. Springer-Verlag, pages 201–211.

Traum, David, Stacy Marsella, Jonathan Gratch, Jina Lee, , and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *8th International Conference on Intelligent Virtual Agents*, September.

Traum, David, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1699–1702.

Tur, Gokhan, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Michael Frandsen, Clint Frederickson, Martin Graciarena, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, Jing Tien, Dimitra Vergyri, and Fan Yang. 2010. The CALO meeting assistant system. *IEEE Transactions on Audio, Speech and Language Processing*, to appear.