

Cascaded Lexicalised Classifiers for Second-Person Reference Resolution

Matthew Purver

Department of Computer Science
Queen Mary University of London
London E1 4NS, UK
mpurver@dcs.qmul.ac.uk

Raquel Fernández

ILLC
University of Amsterdam
1098 XH Amsterdam, Netherlands
raquel.fernandez@uva.nl

Matthew Frampton and Stanley Peters

CSLI
Stanford University
Stanford, CA 94305, USA
frampton,peters@csli.stanford.edu

Abstract

This paper examines the resolution of the second person English pronoun *you* in multi-party dialogue. Following previous work, we attempt to classify instances as generic or referential, and in the latter case identify the singular or plural addressee. We show that accuracy and robustness can be improved by use of simple lexical features, capturing the intuition that different uses and addressees are associated with different vocabularies; and we show that there is an advantage to treating referentiality and addressee identification as separate (but connected) problems.

1 Introduction

Resolving second-person references in dialogue is far from trivial. Firstly, there is the *referentiality* problem: while we generally conceive of the word *you*¹ as a deictic addressee-referring pronoun, it is often used in non-referential ways, including as a discourse marker (1) and with a generic sense (2). Secondly, there is the *reference* problem: in addressee-referring cases, we need to know who the addressee is. In two-person dialogue, this is not so difficult; but in multi-party dialogue, the addressee could in principle be any one of the other participants (3), or any group of more than one (4):

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

¹We include *your, yours, yourself, yourselves*.

This paper extends previous work (Gupta et al., 2007; Frampton et al., 2009) in attempting to automatically treat both problems: detecting referential uses, and resolving their (addressee) reference. We find that accuracy can be improved by the use of lexical features; we also give the first results for treating both problems simultaneously, and find that there is an advantage to treating them as separate (but connected) problems via cascaded classifiers, rather than as a single joint problem.

2 Related Work

Gupta et al. (2007) examined the referentiality problem, distinguishing generic from referential uses in multi-party dialogue; they found that 47% of uses were generic and achieved a classification accuracy of 75%, using various discourse features and discriminative classifiers (support vector machines and conditional random fields). They attempted the reference-resolution problem, using only discourse (non-visual) features, but accuracy was low (47%).

Addressee identification in general (i.e. independent of the presence of *you*) has been approached in various ways. Traum (2004) gives a rule-based algorithm based on discourse structure; van Turnhout et al. (2005) used facial orientation as well as utterance features; and more recently Jovanovic (2006; 2007) combined discourse and gaze direction features using Bayesian networks, achieving 77% accuracy on a portion of the AMI Meeting Corpus (McCowan et al., 2005) of 4-person dialogues.

In recent work, therefore, Frampton et al. (2009) extended Gupta et al.'s method to include multi-modal features including gaze direction, again using Bayesian networks on the AMI corpus. This gave a small improvement on the ref-

erentiality problem (achieving 79% accuracy), and a large improvement on the reference-resolution task (77% accuracy distinguishing singular uses from plural, and 80% resolving singular individual addressee reference).

However, they treated the two tasks in isolation, and also broke the addressee-reference problem into two separate sub-tasks (singular vs. plural reference, and singular addressee reference). 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 at one task necessarily affecting performance at another – and we examine this here. In addition, we examine the effect of lexical features, following a similar insight to Katzenmaier et al. (2004); they used language modelling to help distinguish between user- and robot-directed utterances, as people use different language for the two – we expect that the same is true for human participants.

3 Method

We used Frampton et al. (2009)'s AMI corpus data: 948 “*you*”-containing utterances, manually annotated for referentiality and accompanied by the AMI corpus' original addressee annotation. The very small number of two-person addressee cases were joined with the three-person (i.e. all non-speaker) cases to form a single “plural” class. 49% of cases are generic; 32% of referential cases are plural, and the rest are approximately evenly distributed between the singular participants. While Frampton et al. (2009) labelled singular reference by physical location relative to the speaker (giving a 3-way classification problem), our lexical features are more suited to detecting actual participant identity – we therefore recast the singular reference task as a 4-way classification problem and re-calculate their performance figures (giving very similar accuracies).

Discourse Features We use Frampton et al. (2009)'s discourse features. These include simple durational and lexical/phrasal features (including mention of participant names); AMI dialogue act features; and features expressing the similarity between the current utterance and previous/following utterances by other participants. As dialogue act features are notoriously hard to tag automatically, and “forward-looking” information about following utterances may be unavailable in

an on-line system, we examine the effect of leaving these out below.

Visual Features Again we used Frampton et al. (2009)'s features, extracted from the AMI corpus manual focus-of-attention annotations which track head orientation and eye gaze. Features include the target of gaze (any participant or the meeting whiteboard/projector screen) during each utterance, and information about mutual gaze between participants. These features may also not always be available (meeting rooms may not always have cameras), so we investigate the effect of their absence below.

Lexical Features The AMI Corpus simulates a set of scenario-driven business meetings, with participants performing a design task (the design of a remote control). Participants are given specific roles to play, for example that of project manager, designer or marketing expert. It therefore seems possible that utterances directed towards particular individuals will involve the use of different vocabularies reflecting their expertise. Different words or phrases may also be associated with generic and referential discussion, and extracting these automatically may give benefits over attempting to capture them using manually-defined features. To exploit this, we therefore added the use of lexical features: one feature for each distinct word or n-gram seen more than once in the corpus. Although such features may be corpus- or domain-specific, they are easy to extract given a transcript.

4 Results and Discussion

4.1 Individual Tasks

We first examine the effect of lexical features on the individual tasks, using 10-way cross-validation and comparing performance with Frampton et al. (2009). Table 1 shows the results for the referentiality task in terms of overall accuracy and per-class F1-scores; ‘MC Baseline’ is the majority-class baseline; results labelled ‘EACL’ are Frampton et al. (2009)'s figures, and are presented for all features and for reduced feature sets which might be more realistic in various situations: ‘-V’ removes visual features; ‘-VFD’ removes visual features, forward-looking discourse features and dialogue-act tag features.

As can be seen, adding lexical features (‘+words’ adds single word features, ‘+3grams’ adds n-gram features of lengths 1-3) improves the

Features	Acc	F _{gen}	F _{ref}
MC Baseline	50.9	0	67.4
EACL	79.0	80.2	77.7
EACL -VFD	73.7	74.1	73.2
+words	85.3	85.7	84.9
+3grams	87.5	87.4	87.5
+3grams -VFD	87.2	86.9	87.6
3grams only	85.9	85.2	86.4

Table 1: Generic vs. referential uses

Features	Acc	F _{sing}	F _{plur}
MC Baseline	67.9	80.9	0
EACL	77.1	83.3	63.2
EACL -VFD	71.4	81.5	37.1
+words	83.1	87.8	72.5
+3grams	85.9	90.0	76.6
+3grams -VFD	87.1	91.0	77.6
3grams only	86.9	90.8	77.0

Table 2: Singular vs. plural reference.

performance significantly – accuracy is improved by 8.5% absolute above the best EACL results, which is a 40% reduction in error. Robustness to removal of potentially problematic features is also improved: removing all visual, forward-looking and dialogue act features makes little difference. In fact, using *only* lexical n-gram features, while reducing accuracy by 2.6%, still performs better than the best EACL classifier.

Table 2 shows the equivalent results for the singular-plural reference distinction task; in this experiment, we used a correlation-based feature selection method, following Frampton et al. (2009). Again, performance is improved, this time giving a 8.8% absolute accuracy improvement, or 38% error reduction; robustness to removing visual and dialogue act features is also very good, even improving performance.

For the individual reference task (again using feature selection), we give a further ‘NS baseline’ of taking the next speaker; note that this performs rather well, but requires forward-looking information so should not be compared to ‘-F’ results. Results are again improved (Table 3), but the improvement is smaller: a 1.4% absolute accuracy improvement (7% error reduction); we conclude from this that visual information is most important for this part of the task. Robustness to feature unavailability still shows some improvement: ex-

Features	Acc	F _{P1}	F _{P2}	F _{P3}	F _{P4}
MC baseline	30.7	0	0	0	47.0
NS baseline	70.7	71.6	71.1	72.7	68.2
EACL	80.3	82.8	79.7	75.9	81.4
EACL -V	73.8	79.2	70.7	74.1	71.4
EACL -VFD	56.6	58.9	55.5	64.0	47.3
+words	81.4	83.9	79.7	79.3	81.8
+3grams	81.7	83.9	80.3	79.3	82.5
+3grams -V	74.8	81.3	71.7	75.2	71.4
+3grams -VFD	60.7	66.3	55.9	66.2	53.0
3grams only	60.7	63.1	58.1	52.9	63.4
3grams +NS	74.5	76.7	73.8	75.0	72.7

Table 3: Singular addressee detection.

cluding all visual, forward-looking and dialogue-act features has less effect than on the EACL system (60.7% vs. 56.6% accuracy), and a system using only n-grams and the next speaker identity gives a respectable 74.5%.

Feature Analysis We examined the contribution of particular lexical features using Information Gain methods. For the referentiality task, we found that generic uses of *you* were more likely to appear in utterances containing words related to the main meeting topic, such as *button*, *channel*, or *volume* (properties of the to-be-designed remote control). In contrast, words related to meeting management, such as *presentation*, *email*, *project* and *meeting* itself, were predictive of referential uses. The presence of first person pronouns and discourse and politeness markers such as *okay*, *please* and *thank you* was also indicative of referentiality, as were n-grams capturing interrogative structures (e.g. *do you*).

For the plural/singular distinction, we found that the plural first person pronoun *we* correlated with plural references of *you*. Other predictive n-grams for this task were *you mean* and *you know*, which were indicative of singular and plural references, respectively. Finally, for the individual reference task, useful lexical features included participant names, and items related to their roles. 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* were addressed to the “industrial designer”.

Discussion The best F-score of the three sub-tasks is for the generic/referential distinction; the

Features	Acc	F _{gen}	F _{plur}	F _{P1}	F _{P2}	F _{P3}	F _{P4}
MC baseline	49.1	65.9	0	0	0	0	0
EACL	58.3	73.3	24.3	57.6	57.0	36.0	51.1
+3grams	60.9	74.8	42.0	57.7	52.2	35.6	50.2
3grams only	67.5	84.8	61.6	39.1	39.3	30.6	38.6
Cascade +3grams	78.1	87.4	59.1	64.1	76.4	75.0	82.6

Table 4: Combined task: generic vs. plural vs. singular addressee.

worst is for the detection of plural reference (F_{plur} in Table 2). This is not surprising: humans find the former task easy to annotate – Gupta et al. (2007) report good inter-annotator agreement ($\kappa = 0.84$) – but the latter hard. In their analysis of the AMI addressee annotations, Reidsma et al. (2008) observe that most confusions amongst annotators are between the group-addressing label and the labels for individuals; whereas if annotators agree that an utterance is addressed to an individual, they also reach high agreement on that addressee’s identity.

4.2 Combined Task

We next combined the individual tasks into one combined task; for each *you* instance, a 6-way classification as generic, group-referring or referring to one of the 4 participants. This was attempted both as a single classification exercise using a single Bayesian network; and as a cascaded pipeline of the three individual tasks; see Table 4. Both used correlation-based feature selection.

For the single joint classifier, n-grams again improve performance over the EACL features. Using *only* n-grams gives a significant improvement, perhaps due to the reduction in the size of the feature space on this larger problem. Accuracy is reasonable (67.5%), but while F-scores are good for the generic class (above 80%), others are low.

However, use of three cascaded classifiers improves performance to 78% and gives large per-class F-score improvements, exploiting the higher accuracy of the first two stages (generic/referential, singular/plural), and the fact that different features are good for different tasks.

5 Conclusions

We have shown that the use of simple lexical features can improve performance and robustness for all aspects of second-person pronoun resolution: referentiality detection and reference identification. An overall 6-way classifier is feasible, and cascading individual classifiers can help. Future

plans include testing on ASR transcripts, and investigating different classification techniques for the joint task.

References

- M. Frampton, R. Fernández, P. Ehlen, M. Christoudias, T. Darrell, and S. 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 EACL*.
- S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. 2007. Resolving “you” in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.
- N. Jovanovic, R. op den Akker, and A. Nijholt. 2006. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the EACL*.
- N. Jovanovic. 2007. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. Ph.D. thesis, University of Twente, The Netherlands.
- 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 the 6th International Conference on Multimodal Interfaces*.
- 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 the 5th International Conference on Methods and Techniques in Behavioral Research*.
- D. Reidsma, D. Heylen, and R. op den Akker. 2008. On the contextual analysis of agreement scores. In *Proceedings of the LREC Workshop on Multimodal Corpora*.
- D. Traum. 2004. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag.
- K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. 2005. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of ICMI*.