Disambiguating Between Generic and Referential *"You"* in Dialog*

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

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

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

Abstract

We describe an algorithm for a novel task: disambiguating the pronoun *you* in conversation. *You* can be generic or referential; finding referential *you* is important for tasks such as addressee identification or extracting 'owners' of action items. Our classifier achieves 84% accuracy in two-person conversations; an initial study shows promising performance even on more complex multi-party meetings.

1 Introduction and Background

This paper describes an algorithm for disambiguating the generic and referential senses of the pronoun *you*.

Our overall aim is the extraction of *action items* from multi-party human-human conversations, concrete decisions in which one (or more) individuals take on a group commitment to perform a given task (Purver et al., 2006). Besides identifying the task itself, it is crucial to determine the *owner*, or person responsible. Occasionally, the name of the responsible party is mentioned explicitly. More usually, the owner is addressed directly and therefore referred to using a second-person pronoun, as in example (1).[1]

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

It can also be important to distinguish between singular and plural reference, as in example (2) where the task is assigned to more than one person:

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

Use of *"you"* might therefore help us both in detecting the fact that a task is being assigned, and in identifying the owner. While there is an increasing body of work concerning *addressee identification* (Katzenmaier et al., 2004; Jovanovic et al., 2006), there is very little investigating the problem of *second-person pronoun resolution*, and it is this that we address here. Most cases of *"you"* do not in fact refer to the addressee but are generic, as in example (3); automatic referentiality classification is therefore very important.

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

2 Related Work

Previous linguistic work has recognized that *"you"* is not always addressee-referring, differentiating between *generic* and *referential* uses (Holmes, 1998; Meyers, 1990) as well as idiomatic cases of *"you know"*. For example, (Jurafsky et al., 2002) found that *"you know"* covered 47% of cases, the referential class 22%, and the generic class 27%, with no significant differences in surface form (duration or vowel reduction) between the different cases.

While there seems to be no previous work investigating automatic classification, there is related work on classifying *"it"*, which also takes various referential and non-referential readings: (Müller, 2006) use lexical and syntactic features in a rule-based classifier to detect non-referential uses, achieving raw accuracies around 74-80% and F-scores 63-69%.

3 Data

We used the Switchboard corpus of two-party telephone conversations (Godfrey et al., 1992), and annotated the data with four classes: generic, referential singular, referential plural and a *reported referential* class, for mention in reported speech of an

*This work was supported by the CALO project (DARPA grant NBCH-D-03-0010) and ONR (MURI award N000140510388). The authors also thank John Niekrasz for annotating our test data.

[1](1,2) are taken from the ICSI Meeting Corpus (Shriberg et al., 2004); (3,4) from Switchboard (Godfrey et al., 1992).

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 105–108,
Prague, June 2007. ©2007 Association for Computational Linguistics

	Training	Testing
Generic	360	79
Referential singular	287	92
Referential plural	17	3
Reported referential	5	1
Ambiguous	4	1
Total	673	176

Table 1: Number of cases found.

originally referential use (as the original addressee may not be the current addressee – see example (4)). We allowed a separate class for genuinely *ambiguous* cases. Switchboard explicitly tags *"you know"* when used as a discourse marker; as this (generic) case is common and seems trivial we removed it from our data.

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

To test inter-annotator agreement, two people annotated 4 conversations, yielding 85 utterances containing *"you"*; the task was reported to be easy, and the kappa was 100%.

We then annotated a total of 42 conversations for training and 13 for testing. Different labelers annotated the training and test sets; none of the authors were involved in labeling the test set. Table 1 presents information about the number of instances of each of these classes found.

4 Features

All features used for classifier experiments were extracted from the Switchboard LDC Treebank 3 release, which includes transcripts, part of speech information using the Penn tagset (Marcus et al., 1994) and dialog act tags (Jurafsky et al., 1997). Features fell into four main categories:[2] *sentential* features which capture lexical features of the utterance itself; *part-of-speech* features which capture shallow syntactic patterns; *dialog act* features capturing the discourse function of the current utterance and surrounding context; and *context features* which give oracle information (i.e., the correct generic/referential label) about preceding uses of *"you"*. We also investigated using the presence of a question mark in the transcription as a feature, as a possible replacement for some dialog act features. Table 2 presents our features in detail.

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

Table 2: Features investigated. N indicates the number of possible values (there are 46 DA tags; context features can be *generic*, *referential* or *N/A*).

5 Experiments and Results

As Table 1 shows, there are very few occurrences of the referential plural, reported referential and ambiguous classes. We therefore decided to model our problem as a two way classification task, predicting generic versus referential (collapsing referential singular and plural as one category). Note that we expect this to be the major useful distinction for our overall action-item detection task.

Baseline A simple baseline involves predicting the dominant class (in the test set, referential). This gives 54.59% accuracy (see Table 1).[3]

SVM Results We used LIBSVM (Chang and Lin, 2001), a support vector machine classifier trained using an RBF kernel. Table 3 presents results for

[2]Currently, features are all based on perfect transcriptions.

[3]Precision and recall are of course 54.59% and 100%.

Features	Accuracy	F-Score
Ctxt	45.66%	0%
Baseline	54.59%	70.63%
Sent	67.05%	57.14%
Sent + Ctxt + POS	67.05%	57.14%
Sent + Ctxt + POS + QM	76.30%	72.84%
Sent + Ctxt + POS + Q_DA	79.19%	77.50%
DA	80.92%	79.75%
Sent + Ctxt + POS + QM + DA	84.39%	84.21%

Table 3: SVM results: generic versus referential

various selected sets of features. The best set of features gave accuracy of 84.39% and f-score 84.21%.

Discussion Overall performance is respectable; precision was consistently high (94% for the highest-accuracy result). Perhaps surprisingly, none of the context or part-of-speech features were found to be useful; however, dialog act features proved very useful – using these features alone give us an accuracy of 80.92% – with the referential class strongly associated with question dialog acts.

We used manually produced dialog act tags, and automatic labeling accuracy with this fine-grained tagset will be low; we would therefore prefer to use more robust features if possible. We found that one such heuristic feature, the presence of question mark, cannot entirely substitute: accuracy is reduced to 76.3%. However, using only the binary Q_DA feature (which clusters together all the different kinds of question DAs) does better (79.19%). Although worse than performance with a full tagset, this gives hope that using a coarse-grained set of tags might allow reasonable results. As (Stolcke et al., 2000) report good accuracy (87%) for statement vs. question classification on manual Switchboard transcripts, such coarse-grained information might be reliably available.

Surprisingly, using the oracle context features (the correct classification for the previous *you*) alone performs worse than the baseline; and adding these features to sentential features gives no improvement. This suggests that the generic/referential status of each *you* may be independent of previous *you*s.

Features	Accuracy	F-Score
Prosodic only	46.66%	44.31%
Baseline	54.59%	70.63%
Sent + Ctxt + POS + QM + DA + Prosodic	84.39%	84.21%

Table 4: SVM results: prosodic features

Category	Referential	Generic
Count	294	340
Pitch (Hz)	156.18	143.98
Intensity (dB)	60.06	59.41
Duration (msec)	139.50	136.84

Table 5: Prosodic feature analysis

6 Prosodic Features

We next checked a set of prosodic features, testing the hypothesis that generics are prosodically reduced. Mean pitch, intensity and duration were extracted using Praat, both averaged over the entire utterance and just for the word *"you"*. Classification results are shown in Table 4. Using only prosodic features performs below the baseline; including prosodic features with the best-performing feature set from Table 3 gives identical performance to that with lexical and contextual features alone.

To see why the prosodic features did not help, we examined the difference between the average pitch, intensity and duration for referential versus generic cases (Table 5). A one-sided t-test shows no significant differences between the average intensity and duration (confirming the results of (Jurafsky et al., 2002), who found no significant change in duration). The difference in the average pitch was found to be significant (p=0.2) – but not enough for this feature alone to cause an increase in overall accuracy.

7 Error Analysis

We performed an error analysis on our best classifier output on the training set; accuracy was 94.53%, giving a total of 36 errors.

Half of the errors (18 of 36) were ambiguous even for humans (the authors), if looking at the sentence alone without the neighboring context from the actual conversation – see (5a). Treating these examples thus needs a detailed model of dialog context.

The other major class of errors requires detailed knowledge about sentential semantics and/or the world – see e.g. (5b,c), which we can tell are referential because they predicate inter-personal comparison or communication.

In addition, as questions are such a useful feature (see above), the classifier tends to label all question cases as referential. However, generic uses do occur within questions (5d), especially if rhetorical (5e):

(5) a. so uh and if you don't have the money then use a credit card
b. I'm probably older than you
c. although uh I will personally tell you I used to work at a bank
d. Do they survive longer if you plant them in the winter time?
e. my question I guess are they really your peers?

8 Initial Multi-Party Experiments

The experiments above used two-person dialog data: we expect that multi-party data is more complex. We performed an initial exploratory study, applying the same classes and features to multi-party meetings.

Two annotators labeled one meeting from the AMI corpus (Carletta et al., 2006), giving a total of 52 utterances containing *"you"* on which to assess agreement: kappa was 87.18% for two way classification of generic versus referential. One of the authors then labeled a testing set of 203 utterances; 104 are generic and 99 referential, giving a baseline accuracy of 51.23% (and F-score of 67.65%).

We performed experiments for the same task: detecting generic versus referential uses. Due to the small amount of data, we trained the classifier on the Switchboard training set from section 3 (i.e. on two-party rather than multi-party data). Lacking part-of-speech or dialog act features (since the dialog act tagset differs from the Switchboard tagset), we used only the sentential, context and question mark features described in Table 2.

However, the classifier still achieves an accuracy of 73.89% and F-score of 74.15%, comparable to the results on Switchboard without dialog act features (accuracy 76.30%). Precision is lower, though (both precision and recall are 73-75%).

9 Conclusions

We have presented results on two person and multi-party data for the task of generic versus referential *"you"* detection. We have seen that the problem is a real one: in both datasets the distribution of the classes is approximately 50/50, and baseline accuracy is low. Classifier accuracy on two-party data is reasonable, and we see promising results on multi-party data with a basic set of features. We expect the accuracy to go up once we train and test on same-genre data and also add features that are more specific to multi-party data.

References

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2006. The AMI meeting corpus. In *MLMI 2005, Revised Selected Papers*.

C.-C. Chang and C.-J. Lin, 2001. *LIBSVM: a library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J. J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE ICASSP-92*.

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

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.*

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder.

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

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*.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.

M. W. Meyers. 1990. Current generic pronoun usage. *American Speech*, 65(3):228–237.

C. Müller. 2006. Automatic detection of nonreferential *It* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the EACL.*

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

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop*.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, C. V. Ess-Dykema, R. Martin, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.