Resolving *It*, *This*, and *That* in Unrestricted Multi-Party Dialog

Christoph Müller
EML Research gGmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
christoph.mueller@eml-research.de

Abstract

We present an implemented system for the resolution of *it*, *this*, and *that* in transcribed multi-party dialog. The system handles NP-anaphoric as well as discourse-deictic anaphors, i.e. pronouns with VP antecedents. Selectional preferences for NP or VP antecedents are determined on the basis of corpus counts. Our results show that the system performs significantly better than a recency-based baseline.

1 Introduction

This paper describes a fully automatic system for resolving the pronouns *it*, *this*, and *that* in unrestricted multi-party dialog. The system processes manual transcriptions from the ICSI Meeting Corpus (Janin et al., 2003). The following is a short fragment from one of these transcripts. The letters FN in the speaker tag mean that the speaker is a female non-native speaker of English. The brackets and subscript numbers are not part of the original transcript.

FN083: Maybe you can also read through the - all the text which is on the web pages cuz I'd like to change the text a bit cuz sometimes $[\text{it}]_1$'s too long, sometimes $[\text{it}]_2$'s too short, *inbreath* maybe the English is not that good, so *inbreath* um, but anyways - So I tried to do $[\text{this}]_3$ today and if you could do $[\text{it}]_4$ afterwards $[\text{it}]_5$ would be really nice cuz I'm quite sure that I can't find every, like, orthographic mistake in $[\text{it}]_6$ or something. (Bns003)

For each of the six 3rd-person pronouns in the example, the task is to automatically identify its referent, i.e. the entity (if any) to which the speaker makes reference. Once a referent has been identified, the pronoun is resolved by linking it to one of its antecedents, i.e. one of the referent's earlier mentions. For humans, identification of a pronoun's referent is often easy: it_1, it_2, and it_6 are probably used to refer to the text on the web pages, while it_4 is probably used to refer to *reading* this text. Humans also have no problem determining that it_5 is not a normal pronoun at all. In other cases, resolving a pronoun is difficult even for humans: this_3 could be used to refer to either *reading* or *changing* the text on the web pages. The pronoun is ambiguous because evidence for more than one interpretation can be found. Ambiguous pronouns are common in spoken dialog (Poesio & Artstein, 2005), a fact that has to be taken into account when building a spoken dialog pronoun resolution system. Our system is intended as a component in an extractive dialog summarization system. There are several ways in which coreference information can be integrated into extractive summarization. Kabadjov et al. (2005) e.g. obtained their best extraction results by specifying for each sentence whether it contained a mention of a particular anaphoric chain. Apart from improving the extraction itself, coreference information can also be used to substitute anaphors with their antecedents, thus improving the readability of a summary by minimizing the number of dangling anaphors, i.e. anaphors whose antecedents occur in utterances that are not part of the summary. The paper is structured as follows: Section 2 outlines the most important challenges and the state of the art in spoken dialog pronoun resolution. Section 3 describes our annotation experiments, and Section 4 describes the automatic

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 816–823,
Prague, Czech Republic, June 2007.

dialog preprocessing. Resolution experiments and results can be found in Section 5.

2 Pronoun Resolution in Spoken Dialog

Spoken language poses some challenges for pronoun resolution. Some of these arise from nonreferential resp. nonresolvable pronouns, which are important to identify because failure to do so can harm pronoun resolution precision. One common type of nonreferential pronoun is pleonastic *it*. Another cause of nonreferentiality that only applies to spoken language is that the pronoun is *discarded*, i.e. it is part of an incomplete or abandoned utterance. Discarded pronouns occur in utterances that are abandoned altogether.

ME010: Yeah. Yeah. No, no. There was a whole co- There was a little contract signed. It was - Yeah. (Bed017)

If the utterance contains a speech repair (Heeman & Allen, 1999), a pronoun in the *reparandum* part is also treated as discarded because it is not part of the final utterance.

ME10: That's - that's - so that's a - that's a very good question, then - now that it - I understand it. (Bro004)

In the corpus of task-oriented TRAINS dialogs described in Byron (2004), the rate of discarded pronouns is 7 out of 57 (12.3%) for *it* and 7 out of 100 (7.0%) for *that*. Schiffman (1985) reports that in her corpus of career-counseling interviews, 164 out of 838 (19.57%) instances of *it* and 80 out of 582 (13.75%) instances of *that* occur in abandoned utterances.

There is a third class of pronouns which is referential but nonetheless unresolvable: *vague* pronouns (Eckert & Strube, 2000) are characterized by having no clearly defined textual antecedent. Rather, vague pronouns are often used to refer to the topic of the current (sub-)dialog as a whole.

Finally, in spoken language the pronouns *it*, *this*, and *that* are often discourse deictic (Webber, 1991), i.e. they are used to refer to an *abstract object* (Asher, 1993). We treat as abstract objects all referents of VP antecedents, and do not distinguish between VP and S antecedents.

ME013: Well, I mean there's this Cyber Transcriber service, right?

ME025: Yeah, that's true, that's true. (Bmr001)

Discourse deixis is very frequent in spoken dialog: The rate of discourse deictic expressions reported in Eckert & Strube (2000) is 11.8% for pronouns and as much as 70.9% for demonstratives.

2.1 State of the Art

Pronoun resolution in spoken dialog has not received much attention yet, and a major limitation of the few implemented systems is that they are not fully automatic. Instead, they depend on manual removal of unresolvable pronouns like pleonastic *it* and discarded and vague pronouns, which are thus prevented from triggering a resolution attempt. This eliminates a major source of error, but it renders the systems inapplicable in a real-world setting where no such manual preprocessing is feasible.

One of the earliest empirically based works adressing (discourse deictic) pronoun resolution in spoken dialog is Eckert & Strube (2000). The authors outline two algorithms for identifying the antecedents of personal and demonstrative pronouns in two-party telephone conversations from the Switchboard corpus. The algorithms depend on two nontrivial types of information: the incompatibility of a given pronoun with either concrete or abstract antecedents, and the structure of the dialog in terms of dialog acts. The algorithms are not implemented, and Eckert & Strube (2000) report results of the manual application to a set of three dialogs (199 expressions, including other pronouns than *it*, *this*, and *that*). Precision and recall are 66.2 resp. 68.2 for pronouns and 63.6 resp. 70.0 for demonstratives.

An implemented system for resolving personal and demonstrative pronouns in task-oriented TRAINS dialogs is described in Byron (2004). The system uses an explicit representation of domain-dependent semantic category restrictions for predicate argument positions, and achieves a precision of 75.0 and a recall of 65.0 for *it* (50 instances) and a precision of 67.0 and a recall of 62.0 for *that* (93 instances) if all available restrictions are used. Precision drops to 52.0 for *it* and 43.0 for *that* when only domain-independent restrictions are used.

To our knowledge, there is only one implemented system so far that resolves normal and discourse deictic pronouns in unrestricted spoken dialog (Strube & Müller, 2003). The system runs on dialogs from the Switchboard portion of the Penn Treebank. For *it*, *this* and *that*, the authors report 40.41 precision and 12.64 recall. The recall does not reflect the actual pronoun resolution performance as it is calculated against all coreferential links in the corpus, not just those with pronominal anaphors. The system draws some non-trivial information from the Penn Treebank, including correct NP chunks, grammatical function tags (subject, object, etc.) and discarded pronouns (based on the `-UNF`-tag). The treebank information is also used for determining the accessibility of potential candidates for discourse deictic pronouns.

In contrast to these approaches, the work described in the following is fully automatic, using only information from the raw, transcribed corpus. No manual preprocessing is performed, so that during testing, the system is exposed to the full range of discarded, pleonastic, and other unresolvable pronouns.

3 Data Collection

The ICSI Meeting Corpus (Janin et al., 2003) is a collection of 75 manually transcribed group discussions of about one hour each, involving three to ten speakers. A considerable number of participants are non-native speakers of English, whose proficiency is sometimes poor, resulting in disfluent or incomprehensible speech. The discussions are real, unstaged meetings on various, technical topics. Most of the discussions are regular weekly meetings of a quite informal conversational style, containing many interrupts, asides, and jokes (Janin, 2002). The corpus features a semi-automatically generated segmentation in which each segment is associated with a speaker tag and a start and end time stamp. Time stamps on the word level are not available. The transcription contains capitalization and punctuation, and it also explicitly records *interruption points* and *word fragments* (Heeman & Allen, 1999), but not the extent of the related disfluencies.

3.1 Annotation

The annotation was done by naive project-external annotators, two non-native and two native speakers of English, with the annotation tool MMAX2[1] on five randomly selected dialogs[2]. The annotation instructions were deliberately kept simple, explaining and illustrating the basic notions of anaphora and discourse deixis, and describing how markables were to be created and linked in the annotation tool. This practice of using a higher number of naive – rather than fewer, highly trained – annotators was motivated by our intention to elicit as many plausible interpretations as possible in the presence of ambiguity. It was inspired by the annotation experiments of Poesio & Artstein (2005) and Artstein & Poesio (2006). Their experiments employed up to 20 annotators, and they allowed for the explicit annotation of ambiguity. In contrast, our annotators were instructed to choose the single most plausible interpretation in case of perceived ambiguity. The annotation covered the pronouns *it*, *this*, and *that* only. Markables for these tokens were created automatically. From among the pronominal[3] instances, the annotators then identified normal, vague, and nonreferential pronouns. For normal pronouns, they also marked the most recent antecedent using the annotation tool's coreference annotation function. Markables for antecedents other than *it*, *this*, and *that* had to be created by the annotators by dragging the mouse over the respective words in the tool's GUI. Nominal antecedents could be either noun phrases (NP) or pronouns (PRO). VP antecedents (for discourse deictic pronouns) spanned only the verb phrase *head*, i.e. the verb, not the entire phrase. By this, we tried to reduce the number of disagreements caused by differing markable demarcations. The annotation of discourse deixis was limited to cases where the antecedent was a finite or infinite verb phrase expressing a proposition, event type, etc.[4]

3.2 Reliability

Inter-annotator agreement was checked by computing the variant of Krippendorff's α described in Passonneau (2004). This metric requires all annotations to contain the same set of markables, a condition that is not met in our case. Therefore, we report α values computed on the *intersection* of the com-

[1] http://mmax.eml-research.de

[2] Bed017, Bmr001, Bns003, Bro004, and Bro005.

[3] The automatically created markables included all instances of *this* and *that*, i.e. also relative pronouns, determiners, complementizers, etc.

[4] Arbitrary spans of text could not serve as antecedents for discourse deictic pronouns. The respective pronouns were to be treated as vague, due to lack of a well-defined antecedent.

pared annotations, i.e. on those markables that can be found in all four annotations. Only a subset of the markables in each annotation is relevant for the determination of inter-annotator agreement: all non-pronominal markables, i.e. all antecedent markables manually created by the annotators, and all referential instances of *it*, *this*, and *that*. The second column in Table 1 contains the cardinality of the union of all four annotators' markables, i.e. the number of all distinct relevant markables in all four annotations. The third and fourth column contain the cardinality and the relative size of the intersection of these four markable sets. The fifth column contains α calculated on the markables in the intersection only. The four annotators only agreed in the identification of markables in approx. 28% of cases. α in the five dialogs ranges from .43 to .52.

	$\vert 1 \cup 2 \cup 3 \cup 4 \vert$	$\vert 1 \cap 2 \cap 3 \cap 4 \vert$		α
Bed017	397	109	27.46 %	.47
Bmr001	619	195	31.50 %	.43
Bns003	529	131	24.76 %	.45
Bro004	703	142	20.20 %	.45
Bro005	530	132	24.91 %	.52

Table 1: Krippendorff's α for four annotators.

3.3 Data Subsets

In view of the subjectivity of the annotation task, which is partly reflected in the low agreement even on markable *identification*, the manual creation of a consensus-based gold standard data set did not seem feasible. Instead, we created *core* data sets from all four annotations by means of majority decisions. The core data sets were generated by automatically collecting in each dialog those anaphor-antecedent pairs that at least three annotators identified independently of each other. The rationale for this approach was that an anaphoric link is the more plausible the more annotators identify it. Such a data set certainly contains some spurious or dubious links, while lacking some correct but more difficult ones. However, we argue that it constitutes a plausible subset of anaphoric links that are useful to resolve.

Table 2 shows the number and lengths of anaphoric chains in the core data set, broken down according to the type of the chain-initial antecedent. The rare type OTHER mainly contains adjectival antecedents. More than 75% of all chains consist of two elements only. More than 33% begin with a pronoun. From the perspective of extractive summarization, the resolution of these latter chains is not helpful since there is no non-pronominal antecedent that it can be linked to or substituted with.

	length	2	3	4	5	6	> 6	total
Bed017	NP	17	3	2	-	1	-	23
	PRO	14	-	2	-	-	-	16
	VP	6	1	-	-	-	-	7
	OTHER	-	-	-	-	-	-	-
	all	37 80.44%	4	4	-	1	-	46
Bmr001	NP	14	4	1	1	1	2	23
	PRO	19	9	2	2	1	1	34
	VP	9	5	-	-	-	-	14
	OTHER	-	-	-	-	-	-	-
	all	42 59.16%	18	3	3	2	3	71
Bns003	NP	18	3	3	1	-	-	25
	PRO	18	1	1	-	-	-	20
	VP	14	4	-	-	-	-	18
	OTHER	-	-	-	-	-	-	-
	all	50 79.37%	8	4	1	-	-	63
Bro004	NP	38	5	3	1	-	-	47
	PRO	21	4	-	1	-	-	26
	VP	8	1	1	-	-	-	10
	OTHER	2	1	-	-	-	-	3
	all	69 80.23%	11	4	2	-	-	86
Bro005	NP	37	7	1	-	-	-	45
	PRO	15	3	1	-	-	-	19
	VP	8	1	-	1	-	-	10
	OTHER	3	-	-	-	-	-	3
	all	63 81.82%	11	2	1	-	-	77
Σ	NP	124	22	10	3	2	2	163
	PRO	87	17	6	3	1	1	115
	VP	45	12	1	1	-	-	59
	OTHER	5	1	-	-	-	-	6
	all	261 76.01%	52	17	7	3	3	343

Table 2: Anaphoric chains in core data set.

4 Automatic Preprocessing

Data preprocessing was done fully automatically, using only information from the manual transcription. Punctuation signs and some heuristics were used to split each dialog into a sequence of graphemic sentences. Then, a shallow disfluency detection and removal method was applied, which removed direct repetitions, nonlexicalized filled pauses like *uh, um*, interruption points, and word fragments. Each sentence was then matched against a list of potential discourse markers (*actually*, *like*, *you know*, *I mean*, etc.) If a sentence contained one or more matches, string variants were created in which the respective words were deleted. Each of these variants was then submitted to a parser trained on written text (Charniak, 2000). The variant with the highest probability (as determined by the parser) was chosen. NP chunk markables were created for all non-recursive NP constituents identified by the parser. Then, VP chunk markables were created. Complex verbal constructions like MD + INFINITIVE were modelled by creating markables for the individual expressions, and attaching them to each other with labelled relations like INFINITIVE_COMP. NP chunks were also attached, using relations like SUBJECT, OBJECT, etc.

5 Automatic Pronoun Resolution

We model pronoun resolution as binary classification, i.e. as the mapping of anaphoric mentions to previous mentions of the same referent. This method is not incremental, i.e. it cannot take into account earlier resolution decisions or any other information beyond that which is conveyed by the two mentions. Since more than 75% of the anaphoric chains in our data set would not benefit from incremental processing because they contain one anaphor only, we see this limitation as acceptable. In addition, incremental processing bears the risk of system degradation due to error propagation.

5.1 Features

In the binary classification model, a pronoun is resolved by creating a set of candidate antecedents and searching this set for a matching one. This search process is mainly influenced by two factors: exclusion of candidates due to *constraints*, and selection of candidates due to *preferences* (Mitkov, 2002). Our features encode information relevant to these two factors, plus more generally descriptive factors like distance etc. Computation of all features was fully automatic.

Shallow constraints for nominal antecedents include number, gender and person incompatibility, embedding of the anaphor into the antecedent, and coargumenthood (i.e. the antecedent and anaphor must not be governed by the same verb). For VP antecedents, a common shallow constraint is that the anaphor must not be governed by the VP antecedent (so-called *argumenthood*). Preferences, on the other hand, define conditions under which a candidate probably is the correct antecedent for a given pronoun. A common shallow preference for nominal antecedents is the parallel function preference, which states that a pronoun with a particular grammatical function (i.e. subject or object) preferably has an antecedent with a similar function. The subject preference, in contrast, states that subject antecedents are generally preferred over those with less salient functions, independent of the grammatical function of the anaphor. Some of our features encode this functional and structural parallelism, including identity of form (for PRO antecedents) and identity of grammatical function or governing verb.

A more sophisticated constraint on NP antecedents is what Eckert & Strube (2000) call *I-Incompatibility*, i.e. the semantic incompatibility of a pronoun with an individual (i.e. NP) antecedent. As Eckert & Strube (2000) note, subject pronouns in copula constructions with adjectives that can only modify abstract entities (like e.g. *true*, *correct*, *right*) are incompatible with concrete antecedents like *car*. We postulate that the preference of an adjective to modify an abstract entity (in the sense of Eckert & Strube (2000)) can be operationalized as the conditional probability of the adjective to appear with a *to*-infinitive resp. a *that*-sentence complement, and introduce two features which calculate the respective preference on the basis of corpus[5] counts. For the first feature, the following query is used:

$$\frac{\#\ \text{it}\ \ (\text{'s}|\text{is}|\text{was}|\text{were})\ \ \text{ADJ}\ \ \text{to}}{\#\ \text{it}\ \ (\text{'s}|\text{is}|\text{was}|\text{were})\ \ \text{ADJ}}$$

According to Eckert & Strube (2000), pronouns that are objects of verbs which mainly take sentence complements (like *assume*, *say*) exhibit a similar incompatibility with NP antecedents, and we capture this with a similar feature. Constraints for VPs include the following: VPs are inaccessible for discourse deictic reference if they fail to meet the *right frontier* condition (Webber, 1991). We use a feature which is similar to that used by Strube & Müller (2003) in that it approximates the *right frontier* on the basis of syntactic (rather than discourse structural) relations. Another constraint is *A-Incompatibility*, i.e. the incompatibility of a pronoun with an abstract (i.e. VP) antecedent. According to Eckert & Strube (2000), subject pronouns in copula constructions with adjectives that can only modify concrete entities (like e.g. *expensive*, *tasty*) are incompatible with abstract antecedents, i.e. they

[5]Based on the approx. 250,000,000 word TIPSTER corpus (Harman & Liberman, 1994).

cannot be discourse deictic. The function of this constraint is already covered by the two corpus-based features described above in the context of *I-Incompatibility*. Another feature, based on Yang et al. (2005), encodes the semantic compatibility of anaphor and NP antecedent. We operationalize the concept of semantic compatibility by substituting the anaphor with the antecedent head and performing corpus queries. E.g., if the anaphor is object, the following query[6] is used:

$$\frac{\text{\# (V|Vs|Ved|Ving) (}\varnothing\text{|a|an|the|this|that) ANTE+} \atop \text{\# (V|Vs|Ved|Ving) (}\varnothing\text{|the|these|those) ANTES}}{\text{\# (ANTE|ANTES)}}$$

If the anaphor is the subject in an adjective copula construction, we use the following corpus count to quantify the compatibility between the predicated adjective and the NP antecedent (Lapata et al., 1999):

$$\frac{\text{\# ADJ (ANTE|ANTES) + \# ANTE (is|was) ADJ+} \atop \text{\# ANTES (are|were) ADJ}}{\text{\# ADJ}}$$

A third class of more general properties of the potential anaphor-antecedent pair includes the type of anaphor (personal vs. demonstrative), type of antecedent (definite vs. indefinite noun phrase, pronoun, finite vs. infinite verb phrase, etc.). Special features for the identification of discarded expressions include the distance (in words) to the closest preceeding resp. following disfluency (indicated in the transcription as an interruption point, word fragment, or *uh* resp. *um*). The relation between potential anaphor and (any type of) antecedent is described in terms of distance in seconds[7] and words. For VP antecedents, the distance is calculated from the *last* word in the entire phrase, not from the phrase *head*. Another feature which is relevant for dialog encodes whether both expressions are uttered by the same speaker.

[6]V is the verb governing the anaphor. Correct inflected forms were also generated for irregular verbs. ANTE resp. ANTES is the singular resp. plural head of the antecedent.

[7]Since the data does not contain word-level time stamps, this distance is determined on the basis of a simple forced alignment. For this, we estimated the number of syllables in each word on the basis of its vowel clusters, and simply distributed the known duration of the segment evenly on all words it contains.

5.2 Data Representation and Generation

Machine learning data for training and testing was created by pairing each anaphor with each of its compatible potential antecedents within a certain temporal distance (9 seconds for NP and 7 seconds for VP antecedents), and labelling the resulting data instance as *positive* resp. *negative*. VP antecedent candidates were created only if the anaphor was either *that*[8] or the object of a form of *do*.

Our core data set does not contain any nonreferential pronouns, though the classifier is exposed to the full range of pronouns, including discarded and otherwise nonreferential ones, during testing. We try to make the classifier robust against nonreferential pronouns in the following way: From the manual annotations, we select instances of *it*, *this*, and *that* that at least three annotators identified as nonreferential. For each of these, we add the full range of all-negative instances to the training data, applying the constraints mentioned above.

5.3 Evaluation Measure

As Bagga & Baldwin (1998) point out, in an application-oriented setting, not all anaphoric links are equally important: If a pronoun is resolved to an anaphoric chain that contains only pronouns, this resolution can be treated as neutral because it has no application-level effect. The common coreference evaluation measure described in Vilain et al. (1995) is inappropriate in this setting. We calculate precision, recall and F-measure on the basis of the following definitions: A pronoun is resolved correctly resp. incorrectly only if it is linked (directly or transitively) to the correct resp. incorrect *non-pronominal* antecedent. Likewise, the number of maximally resolvable pronouns in the core data set (i.e. the evaluation *key*) is determined by considering only pronouns in those chains that do not begin with a pronoun. Note that our definition of precision is stricter (and yields lower figures) than that applied in the ACE context, as the latter ignores incorrect links between two expressions in the *response*

[8]It is a common observation that demonstratives (in particular *that*) are preferred over *it* for discourse deictic reference (Schiffman, 1985; Webber, 1991; Asher, 1993; Eckert & Strube, 2000; Byron, 2004; Poesio & Artstein, 2005). This preference can also be observed in our core data set: 44 out of 59 VP antecedents (69.49%) are anaphorically referred to by *that*.

if these expressions happen to be unannotated in the *key*, while we treat them as precision errors unless the antecedent is a pronoun. The same is true for links in the *response* that were identified by less than three annotators in the *key*. While it is practical to treat those links as wrong, it is also simplistic because it does not do justice to ambiguous pronouns (cf. Section 6).

5.4 Experiments and Results

Our best machine learning results were obtained with the Weka[9] Logistic Regression classifier.[10] All experiments were performed with dialog-wise cross-validation. For each run, training data was created from the manually annotated markables in four dialogs from the core data set, while testing was performed on the automatically detected chunks in the remaining fifth dialog. For training and testing, the person, number[11], gender, and (co-)argument constraints were used. If an anaphor gave rise to a positive instance, no negative training instances were created beyond that instance. If a referential anaphor did not give rise to a positive training instance (because its antecedent fell outside the search scope or because it was removed by a constraint), no instances were created for that anaphor. Instances for nonreferential pronouns were added to the training data as described in Section 5.2.

During testing, we select for each potential anaphor the positive antecedent with the highest overall confidence. Testing parameters include `it-filter`, which switches on and off the module for the detection of nonreferential *it* described in Müller (2006). When evaluated alone, this module yields a precision of 80.0 and a recall of 60.9 for the detection of pleonastic and discarded *it* in the five ICSI dialogs. For training, this module was always on. We also vary the parameter `tipster`, which controls whether or not the corpus frequency features are used. If `tipster` is off, we ignore the corpus frequency features both during training and testing.

We first ran a simple baseline system which resolved pronouns to their most recent compatible antecedent, applying the same settings and constraints as for testing (cf. above). The results can be found in the first part of Table 3. Precision, recall and F-measure are provided for ALL and for NP and VP antecedents individually. The parameter `tipster` is not available for the baseline system. The best baseline performance is precision 4.88, recall 20.06 and F-measure 7.85 in the setting with `it-filter` on. As expected, this filter yields an increase in precision and a decrease in recall. The negative effect is outweighed by the positive effect, leading to a small but insignificant[12] increase in F-measure for all types of antecedents.

Setting		Ante	Baseline P	Baseline R	Baseline F	Logistic Regression P	Logistic Regression R	Logistic Regression F
-it-filter	-tipster	NP	4.62	27.12	7.90	18.53	20.34	19.39^{*}
		VP	1.72	2.63	2.08	13.79	10.53	11.94
		ALL	4.40	20.69	7.25	17.67	17.56	17.61^{*}
	+tipster	NP	-	-	-	19.33	22.03	20.59^{***}
		VP	-	-	-	13.43	11.84	12.59
		ALL	-	-	-	18.16	19.12	18.63^{**}
+it-filter	-tipster	NP	5.18	26.27	8.65	17.87	17.80	17.83^{*}
		VP	1.77	2.63	2.12	13.12	10.53	11.68
		ALL	4.88	20.06	7.85	16.89	15.67	16.26^{*}
	+tipster	NP	-	-	-	20.82	21.61	21.21^{**}
		VP	-	-	-	11.27	10.53	10.88
		ALL	-	-	-	18.67	18.50	18.58^{**}

Table 3: Resolution results.

The second part of Table 3 shows the results of the Logistic Regression classifier. When compared to the best baseline, the F-measures are consistently better for NP, VP, and ALL. The improvement is (sometimes highly) significant for NP and ALL, but never for VP. The best F-measure for ALL is 18.63, yielded by the setting with `it-filter` off and `tipster` on. This setting also yields the best F-measure for VP and the second best for NP. The contribution of the it-filter is disappointing: In both `tipster` settings, the it-filter causes F-measure for ALL to go down. The contribution of the corpus features, on the other hand, is somewhat inconclusive: In both `it-filter` settings, they cause an increase in F-measure for ALL. In the first setting, this increase is accompanied by an increase in F-measure for VP, while in the second setting, F-measure for VP goes down. It has to be noted, however, that none of the improvements brought about by the it-filter or the tipster corpus features is statistically significant. This also confirms some of the findings of Kehler et al. (2004), who found features similar to

[9]http://www.cs.waikato.ac.nz/ml/weka/

[10]The full set of experiments is described in Müller (2007).

[11]The *number* constraint applies to *it* only, as *this* and *that* can have both singular and plural antecedents (Byron, 2004).

[12]Significance of improvement in F-measure is tested using a paired one-tailed t-test and $p <= 0.05$ (*), $p <= 0.01$ (**), and $p <= 0.005$ (***).

our tipster corpus features not to be significant for NP-anaphoric pronoun resolution in written text.

6 Conclusions and Future Work

The system described in this paper is – to our knowledge – the first attempt towards fully automatic resolution of NP-anaphoric and discourse deictic pronouns (*it*, *this*, and *that*) in multi-party dialog. Unlike other implemented systems, it is usable in a realistic setting because it does not depend on manual pronoun preselection or non-trivial discourse structure or domain knowledge. The downside is that, at least in our strict evaluation scheme, the performance is rather low, especially when compared to that of state-of-the-art systems for pronoun resolution in written text. In future work, it might be worthwhile to consider less rigorous and thus more appropriate evaluation schemes in which links are weighted according to how many annotators identified them.

In its current state, the system only processes manual dialog transcripts, but it also needs to be evaluated on the output of an automatic speech recognizer. While this will add more noise, it will also give access to useful prosodic features like stress.

Finally, the system also needs to be evaluated extrinsically, i.e. with respect to its contribution to dialog summarization. It might turn out that our system already has a positive effect on extractive summarization, even though its performance is low in absolute terms.

Acknowledgments. This work has been funded by the Deutsche Forschungsgemeinschaft as part of the DIANA-Summ project (STR-545/2-1,2) and by the Klaus Tschira Foundation. We are grateful to the anonymous ACL reviewers for helpful comments and suggestions. We also thank Ron Artstein for help with significance testing.

References

Artstein, R. & M. Poesio (2006). Identifying reference to abstract objects in dialogue. In *Proc. of BranDial-06*, pp. 56–63.

Asher, N. (1993). *Reference to Abstract Objects in Discourse.* Dordrecht, The Netherlands: Kluwer.

Bagga, A. & B. Baldwin (1998). Algorithms for scoring coreference chains. In *Proc. of LREC-98*, pp. 79–85.

Byron, D. K. (2004). *Resolving pronominal reference to abstract entities.*, (Ph.D. thesis). University of Rochester.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proc. of NAACL-00*, pp. 132–139.

Eckert, M. & M. Strube (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Harman, D. & M. Liberman (1994). *TIPSTER Complete LDC93T3A.* 3 CD-ROMS. Linguistic Data Consortium, Philadelphia, Penn., USA.

Heeman, P. & J. Allen (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Janin, A. (2002). Meeting recorder. In *Proceedings of the Applied Voice Input/Output Society Conference (AVIOS),* San Jose, California, USA, May 2002.

Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke & C. Wooters (2003). The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Hong Kong, pp. 364–367.

Kabadjov, M. A., M. Poesio & J. Steinberger (2005). Task-based evaluation of anaphora resolution: The case of summarization. In *Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research,* Borovets, Bulgaria.

Kehler, A., D. Appelt, L. Taylor & A. Simma (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proc. of HLT-NAACL-04*, pp. 289–296.

Lapata, M., S. McDonald & F. Keller (1999). Determinants of adjective-noun plausibility. In *Proc. of EACL-99*, pp. 30–36.

Mitkov, R. (2002). *Anaphora Resolution.* London, UK: Longman.

Müller, C. (2006). Automatic detection of nonreferential it in spoken multi-party dialog. In *Proc. of EACL-06*, pp. 49–56.

Müller, C. (2007). *Fully automatic resolution of* it, this, *and* that *in unrestricted multi-party dialog.*, (Ph.D. thesis). Eberhard Karls Universität Tübingen, Germany. To appear.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proc. of LREC-04*.

Poesio, M. & R. Artstein (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 76–83.

Schiffman, R. J. (1985). *Discourse constraints on 'it' and 'that': A Study of Language Use in Career Counseling Interviews.*, (Ph.D. thesis). University of Chicago.

Strube, M. & C. Müller (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proc. of ACL-03*, pp. 168–175.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly & L. Hirschman (1995). A model-theoretic coreference scoring scheme. In *Proc. of MUC-6*, pp. 45–52.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Yang, X., J. Su & C. L. Tan (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In *Proc. of ACL-05*, pp. 165–172.