Dialogue Acts, Synchronising Units and Anaphora Resolution

Miriam Eckert & **Michael Strube**
IRCS – University of Pennsylvania
3401 Walnut St, Suite 400 A
Philadelphia PA 19104, USA
{miriame|strube}@linc.cis.upenn.edu

1 Introduction

We present a method for classifying and resolving anaphora in spoken dialogue which relies crucially on the concept of dialogue acts. Our segmentation is influenced by (Clark & Schaefer, 1989) and (Carletta et al., 1997) and shows a high degree of inter-coder reliability. In our model, which is based on (Strube, 1998), the domain from which potential antecedents for both individual and discourse-deictic anaphors can be elicited is defined in terms of dialogue acts. The recall rate of our algorithm is similar to that of state-of-the-art pronoun resolution algorithms but we achieve a far higher precision than would be achieved by applying these to spoken language because the classification of anaphors prevents the algorithm from co-indexing discourse-deictic anaphora with individual antecedents.

2 Anaphora in Dialogue

Most anaphora resolution algorithms are designed to deal with the co-indexing relation between anaphors and NP-antecedents (IPro, IDem). In the spoken language corpus we examined, this only accounts for 45.1% of all anaphoric references. Essential to the success of our algorithm is its ability to identify and resolve discourse-deictic anaphors, which constitute 22.6% (DDPro, DDDem). The referents of these anaphors are not individual, concrete entities but abstract objects, such as events, facts and propositions, eg

(1) A: ...[we never know what they're thinking]$_i$.
B: **That**$_i$'s right. [I don't trust them]$_j$, maybe I guess **it**$_j$'s because of what happened over there with their own people, how they threw them out of power... (sw3241)

Whilst there have been attempts to classify abstract objects and describe the rules governing anaphoric reference to them (Webber, 1991; Asher, 1993; Dahl & Hellman, 1995), there have been no empirical studies using actual resolution algorithms. In our model, the adjacent dialogue act constitutes the domain where *potential* antecedents for both individual and discourse-deictic anaphors can be found. For discourse-deictic reference this is in accordance with the right-frontier rule (Webber, 1991). Restricting the antecedent domain to a single dialogue act in combination with information supplied by the predicate of the discourse-deictic anaphor is used to determine the *actual* antecedent.

Instead of assuming that all levels of abstract objects are introduced to the discourse model by the sentence that makes them available, it has been suggested that anaphoric discourse-deictic reference involves referent *coercion* (Webber, 1991; Asher, 1993; Dahl & Hellman, 1995). This assumption is further justified by the fact that discourse-deictic reference, as opposed to individual anaphoric reference, is often established by demonstratives rather than pronouns.

A further 13.2% of the anaphors are vague, in the sense that the speaker does not refer to a specific linguistic object but appears to be commenting on the general discourse topic (eg **It***'s awful*.) (VagPro, VagDem). The remaining anaphors constitute a particular type of plural pronoun (19.1%) which indirectly co-specifies with a singular antecedent (eg *In Russia,* **they** *have had problems.*) (Inferrable-Evoked Pronoun - IEPro). These last two types are classfied correctly when the algorithm fails to find an compatible antecedent in the domain specified by the dialogue acts.

3 Building Synchronising Units from Dialogue Acts

Our hypothesis is that the attentional state of the discourse participants has a relation to the dialogue acts, which are used by speakers to indicate that common ground is achieved. These assumptions are based on the (Clark & Schaefer, 1989) theory of contributions (cf also (Traum, 1994)). In their account of the Centering model (Grosz et al., 1995) in dialogues, (Byron & Stent, 1998) provide a summary of issues that need to be addressed when analysing spoken language:

Determination of utterance boundaries. Most anaphor resolution algorithms require the utterance unit as a domain for potential antecedents. In spoken language, annotators must use criteria which do not depend on punctuation.

Utterances with no discourse entities. Eg *Uh-huh; yeah; right.* (Byron & Stent, 1998) and (Walker, 1998) assign no importance to such utterances in their models. In our model these constitute a specific type of dialogue act which is used to acknowledge a preceding utterance.

Center of attention in multi-party discourse. The participants of a dialogue may not be focussing on the same entity at a given point in the discourse. Dialogue acts functioning as acknowledgments can indicate which entities have been entered into the joint discourse model.

We thus assume that the establishment of common ground is indicated by dialogue acts and affects the operations for adding and removing discourse entities from the representation of the attentional state, in our model the list of salient discourse entities (S-list, cf (Strube, 1998)). We divide each dialogue into short, clearly defined dialogue acts – Initiations **I** and Acknowledgments **A** – based on the top of the hierarchy given in (Carletta et al., 1997). Each sentence and each conjoined clause counts as a separate **I**, even if they are part of the same turn. **A**'s do not convey semantic content but have a pragmatic function (eg backchannel). In addition there are utterances which function as an **A** but also have semantic content – these are labelled as **A/I**.

A single **I** and an **A** jointly form a *Synchronising Unit (SU)*. Single **I**'s in longer turns constitute **SU**'s by themselves and do not require explicit acknowledgment. The assumption is that by letting the speaker continue, the hearer implicitly acknowledges the utterance. It is only in the context of turn-taking that **I**'s and **A**'s are paired up. The **SU**'s have two functions. Firstly, they are used to indicate at which point the S-list is cleaned up: after each **SU**, discourse entities not evoked in this **SU** are removed from the list. The second point is crucial to our hypothesis that common ground has an influence on attentional state. We assume that only entities in a complete **SU** are entered into the common ground and remain in the S-list for the duration of a further **SU**. If at a turn transition, one speaker's **I** is not acknowledged by the other participant it cannot be included in an **SU**. In this case the discourse entities mentioned in the unacknowledged **I** are added to the S-List but are immediately deleted again when the subsequent **I** shows that they are not part of the common ground. In (2) the *it* in B.51 co-specifies with *sports car* from B.49. and not the most recent but unacknowledged *your driving* in A.50. The latter immediately deleted from the S-List.

(2) **SU**$_i$ **I** B.49: It's kind of a sports coupe, guess a little sports car, but –
S-List: [car]
– **A/I** A.50: You do most of your driving,
S-List: [car, your driving]
SU$_j$ **I** B.51: – **it** does pretty good.
S-List: [car]
A/I A.52: You do most of your driving in the city?(sw2326)

4 The Algorithm

The method for resolving anaphors in spoken dialogue is based on the algorithm described by (Strube, 1998). In our method discourse entities are also added to the S-list (*salience list*) immediately after they are encountered. The order of the list is based on the information status of the discourse entities, basically *hearer-old* discourse entities are preferred over *hearer-new* ones (cf (Prince, 1981) for these terms). At the end of each SU all discourse entities which are not realized in this SU are removed from the S-list. This means that the extension and classification of the dialogue acts determine the set of potential antecedents of an anaphor. A major extension to the algorithm is a method for the classification of different types of pronouns and demonstratives.

The algorithm consists of two branches, one for pronouns and the other for demonstratives. Both of them call the functions *resolveInd* and *resolveDD*, which resolve individual and discourse-deictic anaphora, respectively. *resolveInd* consists only of a search through the S-list for a matching antecedent (with respect to gender and number) as described by (Strube, 1998). *resolveDD* consists of a search through a different list – the A-List, which we assume holds the abstract objects which have previously been referred to by discourse-deictic anaphors. Like the S-List, the A-List is cleaned up at the end of each SU – referents which were not referred to again are removed.

The classification depends in part on the predicative context of the anaphor. We define that an anaphor is *I-incompatible* (cannot refer to an individual object) or *A-incompatible* (cannot refer to an abstract object) if it occurs in one of the contexts described in Table 1. If an anaphor is neither *I-* nor *A-incompatible*, the classification depends on the success of the resolution algorithm.

I-Incompatible (*I) Anaphors in the x-position *cannot* refer to individual, concrete entities.

- Equating constructions where a pronominal referent is equated with an abstract object, eg *x is making it easy, x is a suggestion.*
- Copula constructions whose adjectives can only be applied to abstract entities, eg *x is true, x is false, x is correct, x is right*
- Arguments of propositional attitude verbs which *only* take S'-complements, eg *assume x.*
- Object of *do* (*do x.*)
- Predicate or anaphoric referent is a "reason", eg *x is because I like her, x is why he's late.*

A-Incompatible (*A) Anaphors in the x-position *cannot* refer to abstract entities.

- Equating constructions where a pronominal referent is equated with a concrete individual referent, eg *x is a car.*
- Copula constructions whose adjectives can only be applied to concrete entities, eg *x is expensive, x is tasty, x is loud.*
- Arguments of verbs describing physical contact/stimulation, which cannot be used metaphorically, eg *break x, smash x, eat x, drink x, smell x* but NOT **see x*

Table 1: I-Incompatibility and A-Incompatibility

1. **if** (PRO is I-incompatible)
   then if *resolveDD*(PRO)
   then classify as *DDPro*
   else classify as *VagPro*
2. **else if** (PRO is A-incompatible)
   then if *resolveInd*(PRO)
   then classify as *IPro*
   else classify as *VagPro*
3. **else if** *resolveInd*(PRO)
   then classify as *IPro*
4. **else if** *resolveDD*(PRO)
   then classify as *DDPro*
   else classify as *VagPro*

Table 2: Pronoun Resolution Algorithm

1. **if** (DEM is I-incompatible)
   then if *resolveDD*(DEM)
   then classify as *DDDem*
   else classify as *VagDem*
2. **else if** (DEM is A-incompatible)
   then if *resolveInd*(DEM)
   then classify as *IDem*
   else classify as *VagDem*
3. **else if** *resolveDD*(DEM)
   then classify as *DDDem*
4. **else if** *resolveInd*(DEM)
   then classify as *IDem*
   else classify as *VagDem*

Table 3: Demonstrative Resolution Algorithm

If a pronoun (3.p.s. neuter) is encountered (Table 2), the functions *resolveDD* or *resolveInd* are evaluated, depending on whether the pronoun is *I-incompatible* (1) or *A-incompatible* (2). In the case of success the pronouns are classified as *DDPro* or *IPro*, respectively. In the case of failure, the pronouns are classified as *VagPro*. If the pronoun is neither *I-* nor *A-incompatible* (ie the predicative context of the pronoun is ambiguous in this respect), the classification is only dependent on the success of the resolution, ie on the availability of referents in the S/A-Lists. The function *resolveInd* is evaluated first (3) because we observed a preference for pronouns to have individual antecedents. If successful, the pronoun is classified as *IPro*, if unsuccessful, the function *resolveDD* attempts to resolve the pronoun (4). If this, in turn, is successful, the pronoun is classified as *DDPro*, if it is unsuccessful it is classified as *VagPro*, indicating that the pronoun cannot be resolved using the linguistic context. The procedure is similar in the case of demonstratives (Table 3). The only difference being that the antecedent of a demonstrative is preferentially an abstract object and (3) and (4) are reversed.

3rd person masculine or feminine pronouns are resolved directly by a look-up in the S-list as these cannot be discourse-deictic. 3rd person plural pronouns which can be resolved this way are classified as *IPro*, if they cannot be resolved, they are marked as *IEPro*.

5 Empirical Evaluation

Our data consisted of five randomly selected dialogues from the Switchboard corpus of spoken telephone conversations (LDC, 1993). Two dialogues were used to train the two annotators (SW2041, SW4877), and three further dialogues for testing (SW2403, SW3117, SW3241).

Dialogue Acts. First, turns were segmented into dialogue act units. For the purpose of applying the κ statistic (Carletta, 1996) we turned the segmentation task into a classification task by using boundaries between dialogue acts as one class and non-boundaries as the other (see (Passonneau & Litman, 1997) for a similar practice). Percent agreement (PA) between the annotators was 98.35%, and $\kappa = 0.92$, indicating high reliability of the annotations. These dialogue act units were then classified into Initiations (I), Acknowledgments (A), Acknowledgment/Initiations (A/I), and no dialogue act (No). For this test we used only those dialogue act units which the annotators agreed about. The PA over labels given to the dialogue act units was 92.6%, $\kappa = 0.87$, again indicating that it is possible to annotate these classes reliably.

Individual and Abstract Object Anaphora. For the classification of pronouns (IPro, DDPro, VagPro, IEPro) a PA of 87.5% was measured, $\kappa = 0.81$. For the classification of demonstratives (IDem, DDDem, VagDem) PA was 90.78%, $\kappa = 0.80$.

Co-Indexation of Anaphora. We used only those anaphors whose classification both annotators agreed upon. The annotators then marked the antecedents and co-indexed them with the anaphors. The results were compared and the annotators agreed upon a reconciled version of the data. Annotator accuracy was then measured against the reconciled version. Accuracy ranged from 98.4% (Annotator A) to 96.1% (Annotator B) for individual anaphors and from 85.7% to 94.3% for abstract anaphors.

We used the reconciled version of the annotation as the key for the individual and abstract anaphora resolution algorithms. For individual anaphors Precision was 66.2% and Recall 68.2%, for discourse-deictic anaphors Precision was 63.6% and Recall 70%. The low value for precision indicates that the classification did not perform very well. Only few of the anaphors resolved incorrectly were classified correctly. One of the most common errors was that a discourse-deictic or vague anaphor was classified as individual because an individual antecedent was avialable. A source of errors with respect to the resolution was that we did not allow the domain of the antecedent to excede one SU. However, exactly this restriction allowed us to resolve many of the discourse-deictic anaphors and also classify a high percentage of *VagPros* and *IEPros* correctly.

References

Asher, N. (1993). *Reference to Abstract Objects.* Dordrecht: Kluwer.

Byron, D. & A. Stent (1998). A preliminary model of centering in dialog. In *Proc. of COLING-ACL-98*, pp. 1475–1477.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon & A. Anderson (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Clark, H. H. & E. F. Schaefer (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

Dahl, Östen. & C. Hellman (1995). What happens when we use an anaphor. In *Presentation at the XVth Scandinavian Conference of Linguistics Oslo, Norway*.

Grosz, B. J., A. K. Joshi & S. Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

LDC (1993). *Switchboard.* Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

Passonneau, R. & D. Litman (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics*, pp. 223–255. New York, N.Y.: Academic Press.

Strube, M. (1998). Never look back: An alternative to centering. In *Proc. of COLING-ACL-98*, pp. 1251–1257.

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation.*, (Ph.D. thesis). Department of Computer Science, University of Rochester.

Walker, M. A. (1998). Centering, anaphora resolution, and discourse structure. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in Discourse*, pp. 401–435. Oxford, U.K.: Oxford Univ. Pr.

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