Dialogue Structure and Pronoun Resolution

Joel R. Tetreault and James F. Allen

University of Rochester
Department of Computer Science
Rochester, NY, 14620, USA
{tetreaul, james}@cs.rochester.edu

Abstract

This paper presents an empirical evaluation of a pronoun resolution algorithm augmented with discourse segmentation information. Past work has shown that segmenting discourse can aid in pronoun resolution by making potentially erroneous candidates inaccessible to a pronoun's search. However, implementing this in practice has been difficult given the complexities associated with deciding on a useful scheme and then generating the segmentation reliably. In our study, we investigate whether or not shallow schemes that are easy to generate will improve pronoun resolution in a task-oriented corpus of dialogues. Our results show that incorporating this shallow segmentation at best marginally improves pronoun resolution performance.

1. Introduction

This paper presents empirical work in using dialog structure to aid in pronoun resolution. Past work has shown that segmenting text into groups of sentences can improve reference resolution accuracy by reducing the search space of potential antecedents. However, most work in the field of reference has focused on using syntax and other surface features, such as number of mentions or distance, to resolve pronouns correctly. In addition, most of these algorithms, hand-built or statistical, never fare much better than 80% on their respective corpora (see Tetreault (2001) and Mitkov (2000) for leading methods). The incorrect cases are usually out of the reach of the information provided. Thus it is necessary to incorporate more information into these algorithms to improve performance.

Discourse structure is one of the information sources commonly believed to help bridge the ``20%" gap. There has been a lot of theoretical work, such as Grosz and Sidner (1986) and Moser and Moore (1996), that has shown that segmenting text can eliminate potential antecedents and thus aid in coreference resolution. Though there are many different ways to segment discourse, the common themes are that some sequences are more closely related than others (discourse segments) and that a discourse can be organized as a tree, with the leaves being the individual utterances and the interior nodes being discourse segments. The embeddedness of a segment affects which previous segments, and thus their entities, are accessible. As a discourse progresses, segments close and unless they are close to the root of the tree (have a low embedding) may not be accessible.

Though there are many good examples of discourse structure, actually detecting discourse structure is quite difficult. Past empirical work such as Poesio and Eugenio (2001) and Ide and Cristea (2000, 1998) has focused more on determining a discourse structure and investigating how many possible antecedents are ruled inaccessible incorrectly. Tetreault (2003) took a different metric by simply seeing if implementing different structures would lead to improved performance in a pronoun resolution algorithm. The study used a large newspaper corpus annotated for coreference and Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) but their results were inconclusive: the algorithm was able to constrain the search space (thus speeding up the search in some cases) but performed the same as their syntax-based baseline.

In our study, we investigate a different domain: dialogues, and investigate whether different shallow segmentations incorporated into a pronoun resolution algorithm will improve its performance. We use a corpus that has a syntactic and semantic parse for each utterance, as well as speech-act information. This rich information is used to empirically test two flat segmentation methods: Dialogue Act Segmentation (Eckert and Strube, 2000) and Questions Under Discussion (Roberts, 1996) over third person pronouns. In the following sections we describe our algorithms, corpus, and evaluation.

2. Background

2.1. Discourse Structure

Grosz and Sidner (1986) claim that discourse structure is composed of three interrelated units: a linguistic structure, an intentional structure, and an attentional structure. The linguistic structure consists of the structure of the discourse segments and an embedding relationship that holds between them.

The intentional component determines the structure of the discourse. When people communicate, they have certain intentions in mind and thus each utterance has a certain purpose to convey an intention or support an intention. Grosz and Sidner call these purposes ``Discourse Segment Purposes" or DSP's. DSP's are related to each other by either dominance relations, in which one DSP is embedded or dominated by another DSP such that the intention of the embedded DSP contributes to the intention of the subsuming DSP, or satisfaction-precedent relations in which satisfying the intentions of a DSP is necessary to satisfy the intentions of the next DSP. Given the nesting of DSP's, the intentional structure forms a tree, with the top node being the main intention of the discourse. The intentional structure is more difficult to compute since it requires recognizing the discourse purpose and the relation between intentions.

The final structure is the attentional state, which is responsible for tracking the participant's mental model of

what entities are salient or not in the discourse. It is modeled by a stack of focus spaces, which is modified by a process called focusing. Each discourse segment has a focus space that keeps track of its salient entities, relations, etc. Focus spaces are removed (popped) and added (pushed) from the stack depending on their respective discourse segment purpose and whether or not their segment is opened or closed. The key points about attentional state for the pronoun resolution task are that it maintains a list of the salient entities, prevents illegal access to blocked entities, is dynamic, and is dependent on intentional state.

Tetreault (2003) used a large subsection of the Penn Treebank annotated for Rhetorical Structure Theory annotations to approximate Grosz and Sidner's pushing and popping model. RST is intended to describe the coherence texts by labeling relations between clauses. The relations are binary so after a text has been completely labeled, it is represented by a binary tree in which the interior nodes are relations. With some sort of segmentation and a notion of clauses one can test pushing and popping, using the depth of the clause in relation to the surrounding clauses. The results were inconclusive as different uses of the RST trees incorporated into their baseline pronoun resolution algorithm failed to perform better than the baseline.

One of the problems of the embedded tree structures for modeling discourse is that it is very difficult to generate automatically, let alone annotate manually reliably. And as the previous pronoun study shows, even with a detailed annotation, it is unclear if it is actually useful for pronoun resolution. In our study, we investigate whether shallow segmentation methods based on easily derived information could improve pronoun resolution.

2.2. Dialogue Act Segmentation

One approach to dialogue segmentation was suggested by Eckert and Strube (2000) to identify and resolve demonstrative and co-indexing anaphora. Their model makes use of the anaphora's surface form, it's predicative context and the discourse structure. Their main argument is that in a dialogue, common ground is very important in determining which entities are available for reference, since it is possible that a participant ignored the other speaker's utterance or misheard it. An acknowledgment, usually no more than a few words, implicitly signals the other participant that his or her words have been heard and understood. If an utterance is not acknowledged then it is implied that the utterance may not have been heard, or an implicit acknowledgment is being done (usually happens when one speaker talks for awhile). Their theory is that utterances that are not acknowledged are not in common ground and therefore should not be in the discourse history. For pronominal reference, this means that any potential candidates in that utterance are not available for reference.

To make use of this assumption about grounding, Eckert and Strube created dialogue acts to describe the communicative content an utterance. These dialogue acts or DA's are a simplified version of speech acts, and there are only three: 1. Acknowledgements (A) – which are words or vocal signals that indicate understanding, 2. Initiations (I): statements or questions and 3. Acknowledgement/Initiations – utterances which both acknowledge what the other speaker just said, as well as add new information to the discourse. These are usually statements following an (I) from the other speaker. Eckert and Strube label these A/I's, but for brevity's sake, we call them C for combination. In addition, we added the code (N) to represent utterances that have no informational content or act as a signal to show that the previous utterance was not understood (such as "I didn't get that"). As a dialogue is processed, the DA of each new utterance is identified and then utterances are added to the discourse history if necessary. For example, if speaker A utters a statement (I), then speaker B acknowledges it, speaker A's utterance is committed to the discourse history. However, if speaker B uttered something unrelated to speaker A's I, then speaker A's utterance does not get committed to the history and thus its entities are not open to reference. A sequence of over 3 or more I's by the same speaker represents an Initiation chain and is thought to be implicitly acknowledged because the other participant is allowing the speaker to continue without interjecting.

An excerpt from our s2 dialogue shows a sample annotation (Figure 1). The second column is the speaker and the third column is the manually annotated dialogue act.

utt1	S	(I)	so gabriela
utt2	U	(A)	yes
utt3	S	(I)	at the rochester airport there has been a bomb attack
utt4	U	(A)	oh my goodness
utt5	S	(I)	but it's okay
utt6	U	(I)	where is it
utt7	U	(N)	just a second
utt8	U	(I)	i can't find the rochester airport
utt9	S	(N)	it's
utt10	U	(I)	i think i have a disability with maps
utt11	U	(I)	have i ever told you that before
utt12	S	(I)	it's located on brooks avenue
utt13	U	(A)	oh thank you
utt14	S	(I)	do you see it
utt15	U	(A)	yes

Figure 1. Excerpt from s2 with DA annotation

One major advantage of this dialogue act scheme is that it is very simple, and thus is easy to get high reliability between annotators. Their model also has the advantage of working incrementally.

2.3. Questions Under Discussion

In Roberts (1996), discourse segments can be thought of as a series of utterances which address some common theme, called the Question Under Discussion, or QUD. A question (or topic) opens the segment, and the subsequent utterances are instrumental in addressing the question. When the question is fully answered, the segment ends. Our assumption is that once the segment ends, only the most salient entities – namely the ones posed in the question and the ones in the answer are what should remain in the discourse history. Though this metric removes different utterances than what the DA model would predict, the spirit is the same: to remove low-salience (competing) entities from consideration.

We found that the QUD model could be applied to our task-oriented domain since the dialogues each had several questions in them that were easily identifiable by the utterance's speech-acts. Often these questions were clarifications or asides, but sometimes they would initiate a long planning segment. Finally, repeated acknowledgments or repeated key phrases (answers) serve as good markers that the segment is closed. Other discourse cues such as ``so" also serve as good markers for a segment change.

We manually annotated all questions as segment-starts and marked acknowledgment repetitions and other cues as segment-ends. In addition to the start and end points of an utterance, a segment type was also marked – whether the segment was an aside or a non-aside (such as confirmation or clarification). Asides were classified as a sequence of utterances that had little to do with the rest of the discourse, usually jokes, such as the statement-question pair of utterances 10 and 11 in the excerpt in Figure 1. Asides are treated differently from the other questions in that instead of collapsing the entities from the segment and removing non-salient ones, all entities are removed from history. The reasoning is that asides do not contribute to the discourse so should not "clutter" the discourse history.

A sample annotation for the same excerpt of s2 is seen in Figure 2. Since the first segment (utt6-13) is a non-aside, it would be collapsed. The remaining entities would be what the pronoun *it* refers to, and *brooks avenue*.

```
#S(DS
  :START s2-utt6
  :END s2-utt13
  :TYPE clarification
  )

#S(DS
  :START s2-utt10
  :END s2-utt11
  :TYPE aside
  )

#S(DS
  :START s2-utt14
  :END s2-utt15
  :TYPE confirmation
 )
```

Figure 2. QUD annotation for s2

3. Corpus

Our corpus consists of five transcribed task-oriented dialogs (1756 utterances total) between two humans called the Monroe domain (Stent, 2001). The participants were given a set of emergencies and told to collaborate on building a plan to allocate resources to resolve all the emergencies in a timely manner. The corpus construction consisted of four phases (Tetreault, 2004b). First, disfluencies and speech repairs were removed from sentences that were then parsed by a broad-coverage deep parser with a domain-specific ontology. The output of this second stage was both a syntactic and semantic parse of each sentence in the corpus. The parser works by using a bottom-up algorithm and an augmented context-free grammar with hierarchical features. The parser uses a domain independent ontology combined with a domain model for added selectional restrictions and to help prune unlikely parses.

The parser was run over the entire corpus of 1756 utterances and its syntactic and semantic output was hand-checked by trained annotators and marked for acceptability. The parser was able to correctly parse 1334 (85%) of the utterances. Common problems with bad utterances were incorrect word-senses, wrong attachment in the parse tree, or incorrect semantic features. For our purposes, this meant that there were many pronouns that had underconstrained semantics or no semantics at all. Underconstrained pronouns also can be found in utterances that did parse correctly, since sometimes there is simply not enough information from the rest of the sentence to determine a semantics for the pronoun. This becomes problematic in reference resolution because an underconstrained semantics will tend to match everything, and no semantics will match nothing. Sentences deemed ungrammatical or incomplete (5% of the corpus) would not parse so the representations for each term in the sentence were generated manually.

The third phase involved annotating the reference relationships between terms using a variation of the GNOME project scheme (Poesio, 2000). We annotated coreference relationships between noun phrases and also annotated all pronouns. We labeled each pronoun with one of the following relations: coreference, action, demonstrative, and functional.

4. Evaluation

With an annotated corpus complete, the next step is to test the two different metrics to see if gains can be made using a flat segmentation. To do this we first select a baseline algorithm to compare the new metrics against. For a metric to be deemed successful it would have to perform better than the baseline.

We selected Left-Right Centering (Tetreault, 2001) as our baseline algorithm since it fared well against other algorithms in previous studies and is easy to alter with additional constraints. Left-Right Centering (henceforth LRC) is based on Centering Theory (Grosz et al., 1995) in that it uses salience (calculated from grammatical function) to choose which entities should be the antecedents of anaphoric entities. The algorithm works by first searching the current utterance left-to-right for an antecedent that matches number, gender, and syntactic constraints. If one is not found, then it searches past Cf-lists left-to-right (in which the entities are ranked from most salient to least salient) until an antecedent is found. In our evaluation, we assume prior knowledge of the pronoun type, so recall is 100%.

Unlike many corpora used for pronoun resolution evaluation, the Monroe corpus has semantic features associated with each term. By viewing the features as a vector, one can use a matching algorithm to determine if two entities' semantics are compatible. That is, if the semantics of a potential antecedent fits the constraints of the semantics imposed by the pronoun. For example, if the semantics of the pronoun were that it was a physical object and also a movable object, only entities that had those same features (such as an ambulance) would be

viewed as a potential candidate, all others would be filtered out. Incorporating the semantic filter on top of the other constraints improves performance by 6.4%, or 24 pronouns over our 5 dialogue corpus. In addition to the semantics, additional filters were also encoded such as binding constraints and a specialized algorithm for handling the location pronouns "there" and "here." Further information on the use of semantics in this domain can be found in Tetreault (2004a).

We then augmented the LRC algorithm with the ability to handle discourse segmentation information from the two shallow metrics. An automatic version of the DA method was also conducted using the speech acts from the parsed corpus. All acknowledgment type speech-acts were marked as A's, while all others were marked as I's. No C's or N's were marked since they are too difficult to mark automatically without deep interpretation of each utterance and the context. The manual annotation of the DA method was done on 3 of the 5 annotated dialogues (175 pronouns). The annotators reported similar high kappa scores as Eckert and Strube.

For the QUD metric, entities from the start and end segment only remained in discourse history once a non-aside segment was closed. If the segment were an aside, then all entities were removed from consideration as soon as the segment closed.

5. Results

In our first evaluation (Table 1), we used the LRC algorithm performing at its best where it incorporates syntactic, number and gender filters as well as semantic constraints. The algorithm gets 66.9% of the pronouns right in both 3-dialogue and 5-dialogue corpus. The QUD algorithm performs the same over s4 and s12, but gets 3 more pronouns right over s2, so there is a slight advantage. The results show that the DA-metrics do not perform as well as this baseline metric, though in the case of s12, the automatic DA method gets one more pronoun right than the baseline metric because an intervening candidate was removed from consideration. Figure 3 shows this case. The pronoun **that** in utterance 54 refers to **the ambulance** in utterance 50. The baseline LRC selects the heart attack patient incorrectly as its antecedent since the semantics of a patient – that it is movable and a physical object, etc. match the semantics of the pronoun. But because utterance 53 is not acknowledged, the augmented algorithm removes *heart attack patient* from consideration and **the ambulance** is correctly selected by going back through the history list.

We also evaluated our algorithms using LRC without the semantic filter. Our justification for holding this out was that in many natural language systems it is not common to get a large list of semantic features automatically or manually. Most systems use the usual morpho-syntactic constraints such as surface form, number and gender. So, we wished to investigate how well discourse segmentation improves performance without this information. We believed that the original baseline was too hard to improve because the semantic filter might get many of the cases correct that discourse segmentation could be used for.

utt50 S (I) so I guess we should send **one ambulance** straight off to marketplace right now, right.
utt51 U (N) a
utt52 U (A) right, yeah
utt53 S (I) that's the *heart attack patient* I guess
utt54 U (I) we should send **that** off

Figure 3: Excerpt from s12

The results in Table 2 indicate otherwise though. The baseline metric still performs better than the DA metrics, but QUD wins out by two pronouns still. Though QUD does still show an improvement over the baseline, the improvement is not really significant, and in this case, not the large boost we were hoping to observe when the semantic filter was removed.

Table 3 shows that over the entire corpus of 5 dialogues (the original 3 plus two more: s16 and s17), with semantics, the baseline algorithm is better than the other two, but without semantics, QUD gets a slight edge once again. The cases that QUD get right are actually all from dialogue s2 because it has several asides which have intervening competitive antecedents for pronouns following the aside.

6. Discussion

Our study accords with the earlier study of Tetreault (2003) that augmenting an existing pronoun resolution algorithm with segmentation information does not improve performance. Unlike the previous study, this one uses a less complex scheme that is easier to annotate reliably and quickly. However, this simpler and flatter scheme is only successful for the QUD metric, and even then, only marginally so.

The QUD metric also has the drawback of currently being generated manually. Detecting asides is usually very difficult since real-world knowledge is hard to generate and use automatically. Detecting the end of a question segment can be nettlesome because some questions lead to other questions and it can be difficult to tell if one or more segments are being closed. Also, people do not always use double acknowledgments at the end of a question segment. So while the manual version does slightly better, an automated implementation would probably not beat the baseline.

Comparison with other work is difficult since very little empirical work has been done on dialogue or discourse structure and reference resolution. Two studies: Byron and Stent (1998) and Byron (2000) showed that task-oriented dialogues usually are much harder to get high accuracy rate for reference resolution since one has to take into grounding and disfluencies, etc. They reported accuracies of 30-40% for third person pronouns, using simply syntactic, number and gender constraints. It could be that our baseline algorithm, with or without semantics performs too well to notice the effects of discourse segmentation.

Dialogue (# of pronouns)	Baseline	DA-Automatic	DA-Manual	QUD
S2 (71)	47	45	41	50
S4 (86)	58	54	55	58
S12 (18)	12	13	10	12
Overall (175)	117 (66.9%)	112 (64.0%)	106 (60.6%)	120 (68.6%)

Table 1: Evaluation with Full Semantics

Dialogue (# of pronouns)	Baseline	DA-Automatic	DA-Manual	QUD
S2 (71)	44	42	37	46
S4 (86)	60	49	47	50
S12 (18)	10	11	9	10
Overall (175)	104 (59.4%)	102 (58.3%)	93 (53.1%)	106 (60.6%)

Table 2: Evaluation without Semantics

Metric	Baseline	DA-Automatic	QUD
Semantics	66.9%	63.4%	66.5%
No Semantics	61.5%	59.7%	61.9%

Table 3: Evaluation over 5 dialogues

Eckert and Strube performed a manual evaluation of their DA method on a corpus of Switchboard dialogues. Their task was slightly different than the study here since they attempted to classify and resolve co-indexing pronouns as well as demonstratives. For the third person pronouns, their precision was 66.2% and recall was 68.2%. However, they do not mention how well their baseline algorithm performs without discourse segmentation, so a comparison is hard to make. It seems that the I and A distinction helps more with demonstratives than co-indexical pronouns.

In short, our goal was to investigate whether discourse segmentation could improve performance of pronoun resolution algorithms, thus narrowing the gap left by morpho-syntactic metrics. Our results show that while flat discourse segmentation is generally easier to generate automatically or annotate reliably, it does not offer significant improvement over the baselines. The QUD metric could be successful, but is dependent on correctly identifying the ends of segments reliably, and also dependent on the corpus having a lot of questions in the first place (as is the case with s2). If the QUD could be expanded to take into account statements that initiate plans or other discourse segments, the method could be promising.

7. Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments, as well as Michael Strube and Dan Gildea for discussion. We would especially like to thank our annotators: Alison Rosenberg, David Ganzhorn, and Micha Elsner for their work in hand-checking dialogues and manual annotating the DA scheme.

Partial support for this project was provided by ONR grant no. N00014-01-1-1015, "Portable Dialog Interfaces" and NSF grant 0328810 "Continuous Understanding."

8. References

Byron, Donna and Amanda Stent, 1998. A Preliminary Model of Centering in Dialog. In *Proceedings of the 17th International Conference of Computational Linguistics and 36th Meeting of the ACL*. Montreal, Quebec, Canada, August 10-14, 1998, p.1475-1477.

Byron, Donna, 2002. Resolving Pronominal Reference to Abstract Entities. In *ACL '02*. Philadelphia, PA, USA, July 7 – 12, 2002.

Eckert, Miriam and Michael Strube, 2000. Dialogue Acts, Synchronizing Units and Anaphora Resolution. *Journal of Semantics*. 17:1, p. 51-89.

Grosz, Barbara J. and Candice Sidner, 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*. 12:3. p.175-204.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein, 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*. 21:2, p.203-226.

Ide, Nancy and Dan Cristea, 1998. Veins Theory: An Approach to Global Cohesion and Coherence. *Proceedings of COLING*.

Ide, Nancy and Dan Cristea, 2000. A Hierarchical Account of Referential Accessibility. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.*

Mann, William C., and Sandra A. Thomson, 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*. 8:3.243-281.

Mitkov, Ruslan, 2000. Towards a More Consistent and Comprehensive Evaluation of Anaphora Resolution Algorithms and Systems. *Proceedings of the 2nd Discourse Anaphora and Anaphora Resolution Colloquium.* p. 96-107.

Moser, M., and J.D. Moore, 1996. Towards a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics.*

Roberts, Craige, 1996. Information Structure in Discourse. *Papers in Semantics*. 49. p.43-70.

Stent, Amanda J., 2001. Dialogue Systems as Conversational Partners: Applying Conversation Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue. Ph.D. thesis, University of Rochester.

Poesio, Massimo, 2000. Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms: Issues and Preliminary Results. *LREC '00*. Athens, Greece, May, 2000.

Poesio, Massimo and Barbara di Eugenio, 2001. Discourse Structure and Accessibility. *ESSLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics.*

Tetreault, Joel R., 2001. A Corpus-Based Evaluation of Centering and Pronoun Resolution. *Computational Linguistics*. 27:4.

Tetreault, Joel R., 2003. An Empirical Evaluation of Pronoun Resolution and Clausal Structure. *Proceedings of the 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization* Venice, Italy, June 23-24, 2003. p. 1-8.

Tetreault, Joel R., 2004. Semantics, Dialogue, and Reference Resolution. *To appear in CATALOG '04.* Barcelona, Spain. July 19-21, 2004.

Tetreault, Joel R., Mary Swift, Preethum Prithviraj, Myroslava Dzikovska, and James Allen, Discourse Annotation in the Monroe Corpus. *To appear in the ACL Workshop on Discourse Annotation.* Barcelona, Spain. July 25-26, 2004.