Semantics, Dialogue, and Reference Resolution

Joel Tetreault
Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
tetreaul@cs.rochester.edu

James Allen
Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
james@cs.rochester.edu

Abstract

Most pronoun resolution research has focused on written corpora while using syntactical and surface cues. Though big gains have been made in this domain with those methods, it is difficult to do better than the 80% coverage in these domains without some world or semantic knowledge. We investigate this issue by incorporating rich semantic information into a proven reference resolution model over a very difficult domain of human-human task-oriented dialogues. Our results show that semantic information greatly improves performance and can even be viewed as a substitution for the usual syntactic filters.

1 Introduction

In this paper we present an automated corpus-based analysis of pronoun coreference resolution using semantics in a spoken dialogue domain. Most work in pronoun resolution has focused on using syntactic and surface features such as word distance or number of mentions to help improve accuracy rates. While many of these methods perform quite well on large corpora, (for example, (Tetreault, 2001), (Mitkov, 2000)), it seems that these methods can't do much better than 80% accuracy. Error analyses from these studies suggest that other information such as discourse structure, semantic information, reasoning, etc. is required to resolve these hard cases, which typically elude most pronoun resolution algorithms.

In addition, while most empirical work in the field has used large corpora of written text as the basis for evaluation, very little work has been conducted on spoken dialog domain, which are very important for use in natural language understanding systems. These domains are much more difficult than their written counterparts because speech repairs, interruptions, and other disfluencies make it hard to get reliable parses, and also very hard to track the focus (Byron and Stent, 1998). For example, (Byron, 2002) showed that syntax and salience metrics that would perform at 80% on Wall Street Journal articles could only perform at 37% over a large task-oriented spoken-dialog domain. Clearly, something other than syntax and surface methods are necessary for successful reference resolution. Furthermore, what work has been done in reference in spoken dialogs has focused on distinguishing between coreferential and demonstrative pronouns, and then the different types of demonstratives, and then trying to resolve each type (Eckert and Strube, 2000), (Byron, 2002). These metrics typically use semantic information of the verb and tracking of acknowledgments to determine type.

In our study we assume knowledge of the type of each pronoun and focus our work on coreferential pronouns specifically. This research is novel in two ways - first, we use semantic knowledge generated from a deep parser, along with surface constraints to aid in resolution; second, we evaluate our algorithm over a large spoken dialog do-

main. The results show that including semantics improves reference resolution. In the following section we discuss our spoken dialogue corpus. Next we discuss the algorithm and close with results and discussion.

2 Corpus Description

Our corpus consists of transcribed task-oriented dialogs between two humans called the Monroe domain (Stent, 2001). In these domains, one participant was given the task of resolving several medical and weather emergencies in a city by allocating resources to resolve all of them in a timely fashion. The other participant acted as a system to aid the first in planning.

Corpus construction (Swift et al., 2004) and (Tetreault et al., 2004) consisted of three phrases: disfluency annotation, parsing, and reference annotation. We annotated our corpus for disfluencies by marking all repeated phrases, repaired phrases, and also marking incomplete and ungrammatical sentences. Examples of incomplete and ungrammatical utterances are: *Actually it's right a ab* and *So ambulance sends generator*.

After removing the disfluencies, each sentence is parsed by a broad-coverage deep parser. 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 output is a syntactic and semantic representation of a sentence.

The semantic representation is a flat unscoped logical form with events and labeled semantic roles. Each term has associated with it an identifying variable, semantic relationships to other terms, and a semantic vector describing the term. The vector is a typed feature list meaning that there is a main type associated with the term (in our case, one of: physical object, abstract object, situation, and proposition) which licenses certain secondary features. For example, a physical object type would license features such as form, origin, mobility, intentional, etc. Likewise, a situation feature type would license features such as aspect, time-span, cause, etc. Each feature has a list of possible values. Some are binary such as the container feature which means an entity can either hold something, or it can't. And some have a wide range such as mobility: fixed, self-moving, non-self-moving. Examples of a term and the semantic vector (see the :SEM field) for the entity (an ambulance) are shown in Figure 1.

The parser was run over the entire corpus of 1756 utterances and its syntactic and semantic output was handchecked 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 would tend to match everything. We decided not to manually parse the utterances that did not parse correctly because we felt a reference resolution model operating in a spoken dialogue domain will have to deal with bad parses and one wants their results to reflect the "real world" situation. Sentences deemed ungrammatical or incomplete were omitted from the parsing and hand-checking phase. We felt that since there were pronouns and possible antecedents in these utterances, it is necessary to maintain some representation of the utterance. So each term in these sentences were generated manually.

The third phase involved annotating the reference relationships between terms. We annotated coreference relationships between noun phrases and also annotated all pronouns. Our annotation scheme is based on the GNOME project scheme (Poesio, 2000) which annotates referential links between entities as well as their respective discourse and salience information. The main difference in our approach is that we do not annotate discourse units and certain semantic features, since most of the basic syntactic and semantic features are produced automatically for us in the parsing phase. We labeled each pronoun with one of the

```
(TERM :VAR V213818
        :LF (A V213818 (:* LF::LAND-VEHICLE W::AMBULANCE)
        :INPUT (AN AMBULANCE))
        :SEM ($ F::PHYS-OBJ
                (SPATIAL-ABSTRACTION SPATIAL-POINT) (GROUP -)
                (MOBILITY LAND-MOVABLE) (FORM ENCLOSURE)
                (ORIGIN ARTIFACT) (OBJECT-FUNCTION VEHICLE)
                (INTENTIONAL -) (INFORMATION -)
                (CONTAINER (OR + -)) (TRAJECTORY -))
                )
```

Figure 1: Excerpt semantic features for "an ambulance"

following relations: coreference (pronoun is in an identity relation with another explicitly mentioned entity), speaker (one of the discourse participants), action (pronoun refers to an event), demonstrative (pronoun refers to an utterance or discourse segment), and functional (pronoun is related to an entity by an indirect relationship). We had a team of annotators work on the files and agree on how to tag each pronoun.

After the annotation phase, a post-processing phase identifies all the noun phrases that refer to the same entity, and generates a unique chain-id for this entity. This is similar to the $ante$ field in the GNOME scheme. The advantage of doing this processing is that it is possible for a referring expression to refer to a past instantiation that was not the last mentioned instantiation, which is usually what is annotated. As a result, it is necessary to mark all coreferential instantiations with the same identification tag.

So the final parsed corpus consists of lists of entities for each sentence. These entities are verbs, noun phrases, etc, and each has a semantic vector associated with it, though at varying degrees of acceptability depending on the parser success. Noun phrases and pronouns entities are annotated for reference.

3 Algorithm

We use a modified version of the Left-Right Centering algorithm (LRC) (Tetreault, 2001) to determine how much of an effect using semantics has in pronoun resolution in a spoken dialogue. We selected this algorithm because it is easy to use and has performed well in other large domains. It works as follows: while processing a sentence, put each noun phrase encountered on a temporary list, and once the sentence has been completely processed, place the temporary list on a history stack. When a pronoun is encountered, we first search the temporary list's elements from left to right taking the first entity (noun phrase) that fits constraints imposed by the pronoun and the context. If a suitable antecedent is not found, we search through the history stack, searching each sentence from left to right.

3.1 Additional Syntactic Constraints

Normally in LRC, the temporary list is sorted by grammatical function (subject, direct object, etc.) before being placed on the history stack. In our domain, syntax is not very helpful in ranking entities within a sentence since the sentences are so short, so we simply rank the list by word order.

We found that gender constraints, though common in written text evaluations, were more of a drawback then an aid. It was not uncommon in our corpus for people to refer to a person with a medical condition with *that*, or to refer to a digging truck with *he*. Number constraints were encoded in the :LF of the term as SET-OF, so it is easy to tell if an entity is a set or not (see Figure 2). Noun phrases such as road crew which have a singular representation but implicitly represent a group of people do not have the SET-OF notation in the :LF but have in their semantics the GROUP feature. When semantics is used, we leverage this information to allow these types of noun phrases to be referred to by plural and singular pronouns.

In addition to the number constraints, we also implemented three other syntax based constraints: binding, predicate-NP linking, and location ranking. Binding is a standard linguistic constraint

```
(TERM :VAR V3337536
      :LF (PRO V3337536 (SET-OF (:* LF::REFERENTIAL-SEM W::THEM))
      :INPUT (THEM)
      :SEM ($ F::PHYS-OBJ
            (F::MOBILITY F::MOVABLE))

            )
```

Figure 2: Excerpt semantic features for "them"

which prevents from noun phrases within the same verb phrase from co-referring. So in the sentence: *They will move that* the two pronouns would be prevented from referring to each other. The obvious exception is reflexives, though none exist in the corpus. This constraint only works well if the utterance parsed properly. There are some instances where a sentence was parsed into fragments so the binding constraint fails.

Predicate-NP linking is the process of replacing an underspecifed pronoun's semantics in a *be* verb phrase with that of its predicate. So in the sentence *it is the digging truck at Avon it* is underspecified but is in a identity relation with the co-theme of the verb phrase so, we replace the pronoun semantics with that of the truck's.

The final constraint, location ranking, is based on research (Tetreault, 2002) on implicit roles which showed that putting a preference order on verb location roles (ie. TO-LOC - where an entity is being taken, FROM-LOC -where an entity is coming from, and AT-LOC, where an entity is situated) improves resolution of implicit roles in a spoken dialogue. Since the dialogues are basically plan-based narratives, where an entity is taken to is more likely to be referred to by a subsequent pronoun than where it was taken from. So when searching for the antecedents for pronouns *there* and *here*, one looks back through each utterance in the discourse history, first re-ranking the possible location candidates, with entities in a TO-LOC role preferred over those in a FROM-LOC role, preferred over those in AT-LOC role or no role at all. For example, in the utterance *Send the digging truck from Elmwood to Mt. Hope* the preferred candidate would be *Mt. Hope* whereas in the original LRC formulation, *Elmwood* would be selected.

3.2 Semantic Filter

A semantic match occurs when the main type between the pronoun and antecedent are the same, and there is no conflict between the features (for example, a match would not occur if the pronoun were mobile but it's candidate was non-moving, but that feature would match if the candidate were self-moving). For pronouns with an underspecified semantics, we simply select the first entity in our search path that meets the remaining constraints. In our study, we only investigate pronouns marked for coreference. Pronouns with other relations, such as functional or demonstrative, were not considered.

4 Evaluation

For our evaluation we selected two baselines (both knowledge-poor versions of LRC): the first uses no semantic knowledge at all and simply selects the first noun phrase in the search regardless of constraints. The second incorporates number and binding constraints. This represents the canonical pronoun resolution constraints used in most systems. The results of both baselines are in Figure 3. The second column indicates what percentage of the 278 pronouns that each algorithm resolved correctly. The fourth column shows how many of the 83 underconstrained pronouns were resolved correctly, and the final column shows how many pronouns with acceptable semantics (out of 195) were resolved) correctly.

The additional rows in the table represent the cumulative effects of adding a constraint onto the constraints in the preceding rows. So adding the location constraint on top of the binding and predicate-NP constraints (and the basic baseline constraints) produces an improvement of 3.2% over not using the constraint. The final row rep-

resents only adding semantics to the baseline constraints.

The main result from this evaluation is that including semantics significantly improves pronoun resolution accuracy. The three syntactic constraints improve performance over the second baseline by 6.5%, or an error reduction of 20.8%. The biggest increase comes from adding semantics (5.4%), or a cumulative error reduction of 31.9%. Another positive outcome from this study is how much only using semantics improves things over the baseline. So from the standpoint of building a natural language system where response time is important, only using the semantic filter is a reasonable alternative to employing a battery of filters on top of semantics.

Another boost can be seen in resolving pronouns with semantics, as it resolves 26 more. This also reflects how useful it is to have a well-parsed corpus to get acceptable semantics for each entity.

We conducted a detailed analysis on the 92 pronouns resolved incorrectly to identify the main categories for error:

Wrong semantics (22) Cases where a bad parse leads to incorrect semantics for either the pronoun or its antecedent so there would be no way for a match to occur. The most common error was plural pronouns having a top-level semantic feature of situation when it should have been physical object. So these pronouns would incorrectly match with events in the discourse as opposed to a set of people, road crews, vehicles, etc.

Underconstrained pronoun - (15) Here there is either not enough information from the rest of the sentence for the parser to give a rich semantics for the pronoun. This means that the pronoun will match more entities than it should.

Difficult (13) There were ten cases in the corpus that required a combination of information and reasoning to resolve the pronoun correctly. Most of the time, the pronoun fit several of the error categories.

Three of the errors were related to discourse structure where some notion of common ground or embedded structure could be helpful in eliminating candidates during search. Usually this happens when pronouns have a long distance antecedent but the intervening utterances are an aside and not related to the topic of the pronoun's sentence. For example, utterances 10 and 11 in Figure 4 are an aside and if removed would prevent *it* from resolving to the *disability*.

UTT8	U	i can't fi nd the rochester airport
UTT9	S	it's
UTT10	U	i think i have a disability with maps
UTT11	U	have i ever told you that before
UTT12	S	it's located on brooks avenue

Figure 4: Excerpt from dialog s2

Bad Parse with intervening candidates (9) Unlike the first case, the semantics for the pronoun and entity are acceptable but intervening entities have incorrect semantics that coincidentally match with the pronoun's semantics. Because the algorithm works by selected the first candidate that meets all constraints, this intervening candidate is selected before the real antecedent is considered.

Pred-NP Binding (8) These cases involved pronouns in utterances that did not parse and thus binding constraints were not able to function. So the pronoun would refer to an entity intrasententially when it really should be blocked.

Locatives (8) The locative ranking method does improve performance for *there* and *here* but there are some cases where that ranking fails. For example, *Strong Hospital* in *the ambulance from Strong* should not be highly ranked because it is in an embedded phrase. And in Figure 5, our algorithm selects *east main* as the most salient entity, but the pronoun at the end refers to *rochester general*.

Set (6) We currently don't handle plurals with multiple antecedents, so the 6 cases of set membership are automatically wrong.

Algorithm	% Right	Right	USP Right	ACC Right
baseline 1	44.6%	124	43	81
baseline 2	55.0%	143	51	102
+binding	57.9%	161	54	107
+pred-np	58.3%	162	54	108
+location	61.5%	171	54	117
+semantics	66.9%	186	54	132
b2+semantics	65.5%	182	54	128

Figure 3: Pronoun Resolution Algorithm Performance

UTT198 S so i'm just gonna take the ambulance from rochester general to east main back to rochester general so that we have one ambulance there

Figure 5: Locatives Example

Intervening Candidate (6) In this case, all parses in the local context are good but there is a candidate that matches the pronoun but is not the correct antecedent.

Functional Semantics (2) There were two cases of pronouns in a functional relation being referred to by a co-indexing pronoun. These errors are due to metonymy.

The error analysis shows the effect of erroneous parses on performance. 39 of the errors (wrong semantics, bad parse with intervening candidates, and pred-NP binding) are due to bad parses producing incorrect semantics for the entities. This shows the difficulty to NLP systems that spoken dialoguess impose. Difficult sentences lead to incorrect parses which then can severely effect reference performance. On the other hand, the error distribution shows the great gains that can be made by getting better parses or by compensating with other metrics. Despite the underspecified semantics for some pronouns, or incorrect semantics, using semantics really improves accuracy instead of harming it.

5 Conclusion

In short, we performed an automated empirical evaluation of pronoun coreference resolution in a large spoken dialog domain using rich semantic information from a deep-parser. The results show that semantic information improves performance over recency-based heuristics, and despite the complications imposed by spoken dialogue.

Future work will include researching ways of dealing with underspecified pronouns and also using discourse cues, grounding, and thematic roles of verbs to further aid resolution.

6 Acknowledgments

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

References

Donna K. Byron and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, student session*.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 2002 annual meeting of the Association for Computational Linguistics (ACL '02)*, pages 80–87, Philadelphia, USA, July.

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

R. Mitkov. 2000. Towards a more consistent and comprehensive evaluation of anaphora resolution algo-

rithms and systems. In *2nd Discourse Anaphora and Anaphora Resolution Colloquium*, pages 96–107.

M. Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *LREC '00*, Athens.

Amanda J. Stent. 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.

M. Swift, M. Dzikovska, J. Tetreault, and James F. Allen. 2004. Semi-automatic syntactic and semantic corpus annotation with a deep parser. In *LREC'04*, Lisbon.

Joel Tetreault, Mary Swift, Preethum Prithviraj, Myroslava Dzikovska, and James Allen. 2004. Discourse annotation in the monroe corpus. In *ACL '04 Workshop on Discourse Annotation*, Barcelona, Spain, July 25-26.

Joel R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Joel R. Tetreault. 2002. Implicit role reference. In *2002 International Symposium on Reference Resolution for Natural Language Processing*, pages 109–115.