INTERACTION BETWEEN DIALOG STRUCTURE AND COREFERENCE RESOLUTION

Amanda J. Stent and Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue
Florham Park, NJ 07932, USA.
stent,srini@research.att.com

ABSTRACT

Determining the *coreference* of entity mentions in a discourse is a key part of the interpretation process for advanced spoken dialog applications. In this paper, we present the most comprehensive system for statistical coreference resolution in dialog to date. We also compare the impact of two contrasting theories of dialog structure (the *stack model* and the *cache model*) on the performance of statistical coreference resolution, and show that the stack model outperforms the cache model.

Index Terms— Natural language interfaces, Speech communication

1. INTRODUCTION

An *entity* in a dialog can be referred to using a range of linguistic expressions. For example, *George W. Bush*, *the forty third president*, *dubya*, and depending on the context, *George* and *Bush* all refer to the same person. The basic coreference task is to determine which *mentions* in a discourse (typically noun phrases) refer to the same entities in the underlying discourse model. Performing coreference resolution is a key part of the interpretation process for advanced spoken dialog applications. Furthermore, as suggested by [1, 2, 3, 4], coreference is intimately entwined with the task of tracking global dialog structure. Hence, exploiting dialog structure can potentially provide constraints to coreference resolution that improve its accuracy.

There has been considerable work on pronoun resolution in dialog (e.g. [1, 5]), but comparatively little on the larger coreference task. Poesio et al. [6] performed a corpus analysis examining the impact of two models of dialog structure on accessibility of referents for pronouns and definite NPs in tutorial dialogs. However, their analysis was only of 17 dialogs, and they did not implement a system for coreference. The only work on statistical models of coreference for dialog, that of Luo et al. [7], does not focus on task-oriented dialog or incorporate a model of dialog structure.

In this paper, we: (a) describe the implementation of a system for statistical coreference for dialog; (b) evaluate the performance of the system on a large corpus of task-oriented dialogs; and (c) compare the impact of two models of dialog structure, the *stack model* and the *cache model*, in the context of our system.

2. THEORIES OF DISCOURSE STRUCTURE

We contrast two theories of discourse structure and illustrate their relationship to coreference: the *stack model* [8] and the *cache model* [9]. According to both theories, a dialog is comprised of three separate but related elements: the linguistic structure (the linear sequence of clauses), the intentional structure (which in the stack model is captured as a stack of discourse segments, each containing clauses relating to a single discourse purpose), and the attentional state (the set of entities salient at any point in the discourse). In the *stack model*, the attentional state is tied to the intentional structure: the set of elements in the attentional state tracks the entities accessible through the discourse stack [8]. In the *cache model*, by contrast, the attentional state is tied to the linguistic structure: it acts as a moving window over the discourse history, modeling working memory constraints in the human language production and comprehension systems [9].

To illustrate the differences between the stack and cache models, consider the (simplified) dialog extract from the CHILD corpus in Figure 1. The speakers are discussing payment information. In clause 40, speaker B pushes a discount subtask onto the stack. This subtask

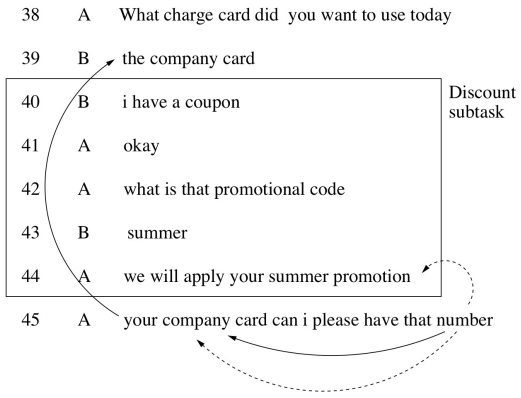

Fig. 1. Example dialog from CHILD corpus

ends after clause 44. In the stack model *summer promotion* is no longer available for reference in clause 45 because the *discount* subtask has been popped from the stack, while in the cache model it is available because it is recent. By contrast, in the cache model *the company card* is not available for reference in clause 45 because it is distant, while in the stack model it is because it is accessible through the stack. Thus, the discourse structure predicts which mentions are available for coreference and this can be exploited to improve the accuracy of coreference resolution.

3. DATA

The CHILD corpus is a corpus of task-oriented human-human spoken dialog in a catalog ordering domain [10]. The dialogs have been transcribed, split into clauses, and annotated for dialog acts and tasks/subtasks. We used 818 CHILD dialogs that involve only two speakers. These dialogs were manually annotated for coreference information: *mentions* (phrases that could be part of coreference chains) and *coreference links* (indicating that pairs of mentions were coreferent) were labeled. There are 105,859 mentions (excluding first and second person pronouns). There are 19,580 coreference chains of length greater than one which include 60,518 mentions altogether. The average chain size is small (see Table 1), and most chains are in a single subtask; however, the average chain span is more than thirteen clauses. This is mostly due to two entities, *the order* and *the catalog*, which are mentioned throughout the dialogs.

We computed the frequency of conflicts between coreference links and the stack and cache models of dialog. A link conflicts with the stack model if the first mention in the link is in a subtask that has been popped off the stack before the second mention is seen. A link conflicts with the cache model if the mentions are separated by more than four turns (four turns is likely to span three mention-containing clauses; a three-sentence window is used in text-based coreference systems). There are 1.5 times as many conflicts per chain with the cache model as with the stack model (see Table 1). However, the gains in 'perfect recall' with the stack model may be offset in a statistical coreference system by losses in precision, as the stack model makes many more mentions available for coreference than the cache model.

Nonsingleton chains	19580
Mentions per chain	3.09
Clauses per chain	13.32
Tasks per chain	1.38
Conflicts per chain (cache model)	0.41
Conflicts per chain (stack model)	0.28

Table 1. Coreference chains in CHILD

4. DISCOURSE-AWARE COREFERENCE

Since the late 1990s, the predominant approaches to coreference resolution in text have been statistical (e.g. [11, 12]). The stages in a full statistical coreference system typically include: (a) mention identification (extracting text segments corresponding to mentions); (b) feature extraction (extracting lexical, syntactic and other features for each mention); (c) pairwise coreference determination (selecting pairs of mentions that could be coreferent, and using a classifier to determine the likelihood that they are); and (d) mention clustering (combining pairwise coreference decisions to produce mention clusters, each corresponding to one entity). However, not all experiments in statistical coreference involve building a full coreference system; most use data in which mentions, mention features, and coreference links have been annotated by hand.

For this paper, we adopt the standard coreference pipeline. We take an 'overhearer' perspective: each dialog is processed incrementally, clause by clause, as it is 'overheard'. From each clause we extract mentions and mention features as described in Sections 4.1 and 4.2 respectively. We perform pairwise coreference classification as described in Section 4.3. At the end of the dialog, we use the pairwise coreference decisions to produce mention clusters as described in Section 4.4.

Our model for statistical coreference uses the discourse structure in two ways. First, information about the attentional state (recency information, and subtask and stack information) is incorporated into the features

for each mention. Second, the attentional state is used to determine which mentions are *available* to corefer – in the cache model, only recent mentions are available, while in the stack model, only mentions visible in the stack are available. The use of discourse-related information as features is not prescriptive, while the use of attentional state to select available mentions is.

4.1. Mention identification

In this paper, we focus on pairwise coreference determination and mention clustering, so we use the *true mentions*, i.e. the mentions hand-labeled in our data.

4.2. Feature extraction

We use three feature sets: **Dialog**-related features, **Task**-related features, and the **Basic** feature set containing lexical, syntactic and semantic features similar to those used in text-based work on coreference (e.g. [11, 13]). All features are listed in Table 2. The pairwise coreference classifiers are trained using unigrams, bigrams and trigrams over these features.

Basic Features In our dialogs, turns are segmented into clauses which are automatically part-of-speech tagged and supertagged. Our labelers did not identify the heads of mentions, so we use the last word of mentions that are not proper nouns, and the full text of mentions that are proper nouns. We used rules to identify the values of the number (*sg/pl/na*), person (*1st/2nd/3rd/na*) and grammatical form (one of {*indefinite, definite, possessive, demonstrative, quantified, proper, pronoun, deictic, qterm, other*}) features, and a dictionary to identify the values of the gender feature. We excluded mentions labeled with 'qterm' (e.g. *which, what, when*).

In some text-based work on coreference, researchers have used additional features that rely on having rich syntactic parses of the input, such as centering-related features and apposition [5, 13, 14]. Given that we are dealing with spoken dialog, some of these features (highly relevant for newspaper text) are not relevant here, or we cannot obtain them with high accuracy from the clause parses due to interference from disfluencies, interruptions, etc. However, by recording the words between two adjacent mentions in an clause, we approximate certain features (e.g. apposition, existential 'it', presence of a conjunction or of the word 'said').

Dialog Features In the only work on general statistical coreference for dialog that we are aware of, Luo et al. [7] use speaker and turn features. We use speaker, turn and dialog act features.

Task Features We use the subtask label of the clause containing the mention, the whole stack of subtask labels, and the depth of the subtask stack.

4.3. Pairwise classification

Pair construction Most coreference systems construct training and test data using the method outlined in [11]: construct one positive example for a mention m_i and its most recent coreferent mention m_j, and construct a negative example for m_i and each m_k s.t. $j < k < i$. We adopt a variation of this method. For each mention m_i, we construct a positive example using the most recent *possible and available* coreferent mention m_j in the preceding discourse. m_j is *possible* if it does not disagree with m_i in the values of the semantic type, number, gender or person features. *Availability* is determined in one of two ways: (a) cache-based – only mentions in this turn and the previous four turns are available; or (b) stack-based – only mentions in the subtask stack for the dialog so far are available. We construct negative examples for every *possible, available* m_k s.t. $j < k < i$ and m_k is not pronominal.

Classification model Using the LLAMA machine learning toolkit [15], we trained a binary classifier using logistic regression for the following combinations of feature selection and pair construction methods:

- Stack-based pair construction, Task, Dialog and Basic features – This corresponds to a **strong stack** model of dialog structure.
- Cache-based pair construction, Dialog and Basic features – This corresponds to a **strong cache** model of dialog structure.
- Cache-based pair construction, Task, Dialog and Basic features – This corresponds to a **hybrid cache/task** model of dialog structure.
- Cache-based pair construction, Basic features – We use this model as a baseline.
- All pair construction – We include these models for comparison, even though they are computationally complex.

4.4. Mention clustering

Mention clustering is the final step in coreference determination; it partitions the input mentions into mention clusters, each corresponding to one entity. We experimented with *the connected components method* approach to mention clustering, which simply finds

Feature Type	Features	mentions		
		m_1	m_2	m_1 vs. m_2
Lexical	(1-2) unigrams, bigrams and trigrams over mention text	x	x	
	(3-4) text length (words); (5-6) clause id	x	x	
	distances between (in (7) words, (8) mentions, and (9) clauses)			x
	(10) head agrees? (11) contained in? (12) in same clause?			x
	(13) Levenshtein distance (mention text)			x
	(14) unigrams, bigrams and trigrams over words between (same clause, no intervening mentions)			x
Syntactic	(1-2) unigrams, bigrams and trigrams over part of speech (POS) tag sequence	x	x	
	(3-4) unigrams, bigrams and trigrams over supertag sequence	x	x	
	(5-6) gender; (7-8) number; (9-10) person features	x	x	
	(11-12) grammatical form	x	x	
	(13) gender agrees? (14) number agrees? (15) person agrees?			x
	(16) grammatical form agrees?			x
	Levenshtein distances ((17) POS tag sequences, (18) supertag sequences)			x
Semantic	(1-2) semantic type	x	x	
	(3) semantic type agrees?			x
Dialog	(1-2) turn id; (3-4) speaker id; (5-6) dialog act(s) for containing clauses	x	x	
	(7) in same turn? (8) by same speaker?			x
	(9) distance (turns)			x
Task	(1-2) subtask label for containing clause	x	x	
	(3-4) task stack at containing clause	x	x	
	(5-6) task stack depth at containing clause	x	x	
	(7) in same subtask?			x
	(8) distance (task stack actions)			x

Table 2. Features used in finding coreferent pairs of mentions. The Basic feature set includes the lexical, syntactic and semantic features. Numbers indicate feature count within the feature set.

connected components in the graph constructed from the mention pair links with probability greater than 0.5 output by the pairwise coreference classifier. However, pairwise coreference classification does not ensure transitivity; i.e. that if mention pairs (m_i, m_j) and (m_j, m_k) are coreferent, then mention pair (m_i, m_k) is also coreferent. So we also tried the ILP method outlined in [16]. Given a document D, let $M = \{m_i \in D\}$ be the set of mentions in D, and let $P = \{(i,j) | m_i \in M, m_j \in M,$ and $i < j\}$ be the set of possible coreference links over these mentions. For each $(i,j) \in P$, let $p_{(i,j)}$ be the probability assigned to (i,j) by the pairwise coreference classifier, and let $x_{(i,j)}$ be an indicator variable representing (i,j). The objective function we use is: $min \sum_{(i,j) \in P} -log(p_{(i,j)}) * x_{(i,j)} + -log(1 - p_{(i,j)}) * (1 - x_{(i,j)})$ subject to: $(1 - x_{(i,j)}) + (1 - x_{(j,k)}) \geq (1 - x_{(i,k)}) \forall m_i, m_j, m_k \in M$ s.t. $i \neq j \neq k$ and $x_{(i,j)} \in \{0,1\} \forall (i,j) \in P$. We used *lp_solve* as our ILP solver.

5. EXPERIMENTS

We used ten-fold cross-validation on our data. For each test dialog, we performed pairwise coreference classification using each of the five models, and mention clustering using both the connected components method and the ILP method.

We report results using the MUC-6 metric [17], the B^3 metric [18] and the CEAF metric [19]. The MUC-6 metric operates by determining the number of *links* that are common between the set of chains proposed by a model with the set of true chains in the reference corpus. Recall, precision and f-scores are computed by comparing these links. The B^3 metric measures the recall and precision for each *mention* m by comparing the set of elements in the chain containing m between the model's output and the true chain containing m. Overall recall, precision and f-score is obtained as an average of individual mention scores. The CEAF metric first computes the best one-to-one mapping between all the chains proposed by the model and all the true chains. Then the recall, precision and f-scores are computed based on the *mentions in the aligned chains*.

Because these metrics evaluate different aspects of the coreference task, a method may lead to improvements according to one metric but not according to another metric: for example, a method that generates fewer links but with high accuracy may lead to high MUC scores but low CEAF scores.

6. RESULTS

Our experimental results are shown in Table 3. No method achieves high recall in finding coreference links:

Method	Scoring metric								
	MUC-6			B^3			CEAF		
	R	P	F	R	P	F	R	P	F
Strong stack-based (variable history)									
Stack, Task+Dialog+Basic, CC	42.0	69.7	52.4	74.2	91.2	81.9	86.3	69.0	76.7
Stack, Task+Dialog+Basic, ILP	41.6	69.8	52.2	74.1	91.4	81.8	86.4	68.9	76.6
Hybrid cache/task-based (4 turns history)									
Cache, Task+Dialog+Basic, CC	38.9	69.7	49.9	73.2	92.1	81.6	86.5	67.6	75.9
Cache, Task+Dialog+Basic, ILP	38.6	69.8	49.7	73.1	92.2	81.6	86.6	67.5	75.9
Strong cache-based (4 turns history)									
Cache, Dialog+Basic, CC	40.0	72.8	51.6	73.6	92.9	82.2	87.1	67.8	76.2
Cache, Dialog+Basic, ILP	39.7	72.9	51.4	73.5	93.1	82.1	87.1	67.7	76.2
Cache-based baseline (4 turns history, no dialog features)									
Cache, Basic, CC	37.0	71.6	48.7	72.6	93.1	81.6	86.8	66.5	75.3
Cache, Basic, ILP	36.6	71.6	48.5	72.5	93.2	81.6	86.9	66.4	75.3
All (ILP not shown to save space)									
All, Basic, CC	38.1	69.3	49.2	72.7	91.9	81.2	86.3	67.2	75.5
All, Task+Dialog+Basic, CC	42.8	68.0	52.6	74.2	90.2	81.4	85.7	69.5	76.8

Table 3. Coreference resolution results

R for MUC-6 is low across the board. However, as most clusters have only one element the impact on overall performance in finding coreference clusters for mentions is small: P for B^3 is uniformly high. Also, comparing the first and second row for each method, we see that the ILP approach does not give significant performance gains for any metric (see Section 7).

The All method (unlimited history) gives the best F-scores on the MUC-6 and CEAF metrics; however, these models take exponentially longer (days) to train and test. In addition, the best results using this method (last row) are only 0.1% (CEAF) to 0.2% (MUC-6) better than the F-scores for the strong stack model (first row).

We find interesting results for our comparison of models of dialog structure. First, the inclusion of dialog-related features alone gives small but consistent improvements in F-scores on every metric: 2.9% in MUC-6, 0.6% in B^3, and 0.9% in CEAF (compare the strong cache-based model with the cache-based baseline). These results, which agree with those of [7], are mostly due to increased recall.

Second, the strong stack model of dialog structure gives small but consistent improvements in F-score over other models that include task and dialog features: 2.5% in MUC-6 and 0.8% in CEAF better than the hybrid cache/task model, and 0.8% in MUC-6 and 0.5% in CEAF over the strong cache-based model. The stack model finds more correct links (increased recall for MUC-6), leading to fewer and more accurate mention clusters (increased precision for CEAF). These results agree with our findings in Section 3 regarding coreference conflicts. In the CHILD dialogs, many subtasks are longer than the 4-turn window provided by the cache model. So one way of interpreting these results is that the subtask structure provides a theoretically informed way of having a dynamically sized window in which to look for available mentions. This leads to improved recall in mention-mention links (MUC-6), and improved precision in clusters (CEAF), with a slight drop in precision for mention-cluster links (B^3).

7. DISCUSSION

In our experiments for this paper, we tried two other methods for pair construction in an effort to improve recall: (1) make an example for each possible available mention, not stopping at the closest coreferent mention; and (2) use the method outlined in [11], but only permit the most recent pronoun to be possibly coreferent. Our basic findings remain the same in either case: dialog features help, and the stack model helps more. Method (1) results in a very large imbalance between positive and negative examples, so that for CHILD data (which has many singleton chains) MUC-6 scores decline dramatically, while B^3 and CEAF scores stay high. However, with method (1) the use of the transitivity constraint does lead to significant improvements over simple connected components clustering. Method (2) leads to slightly lower recall scores. We plan to explore other alternative methods for pair construction.

One way of looking at coreference is the method we have used here: the 'overhearer' method. However, during a dialog the participants have an inside perspective on the interaction, which provides additional constraints on coreference. We are currently exploring participant-specific models of coreference for dialog.

8. CONCLUSIONS

In this paper, we presented the first implementation of a statistical coreference system for task-oriented dialog. In the context of this system, we compared the cache and stack models of global dialog structure and found that the stack model gives improved performance compared to the cache model when incorporated into a statistical coreference system.

9. ACKNOWLEDGMENTS

We thank Barbara Hollister and her annotation team for labeling and checking the CHILD data.

10. REFERENCES

[1] D. Byron and J. Allen, "What's a reference resolution module to do? redefining the role of reference in language understanding systems," in *Proceedings of DAARC*, 2002.

[2] G. Ferguson et al., "CARDIAC: An intelligent conversational assistant for chronic heart failure patient heath monitoring," in *Proceedings of the AAAI Fall Symposium on Virtual Healthcare Interaction*, 2009.

[3] D. Schlangen, T. Baumann, and M. Atterer, "Incremental reference resolution: the task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies," in *Proceedings of SIGDIAL*, 2009.

[4] H. Zender, G.-J. Kruijff, and I. Kruijff-Korbayova, "A situated context model for resolution and generation of referring expressions," in *Proceedings of ENLG*, 2009.

[5] M. Strube and C. Muller, "A machine learning approach to pronoun resolution in spoken dialogue," in *Proceedings of ACL*, 2003.

[6] M. Poesio, A. Patel, and B. di Eugenio, "Discourse structure and anaphora in tutorial dialogues: an empirical analysis of two theories of the global focus," *Research in Language and Computation*, vol. 4, pp. 229–257, 2006.

[7] X. Luo, R. Florian, and T. Ward, "Improving coreference resolution by using conversational metadata," in *Proceedings of NAACL-HLT*, 2009.

[8] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[9] M. Walker, "Limited attention and discourse structure," *Computational Linguistics*, vol. 22, no. 2, pp. 255–264, 1996.

[10] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," in *Proceedings of COLING/ACL*, 2006.

[11] W. Soon, H. Ng, and D. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[12] V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff, "Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art," in *Proceedings of ACL-IJCNLP*, 2009.

[13] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of EMNLP*, 2008.

[14] A. Haghighi and D. Klein, "Simple coreference resolution with rich syntactic and semantic features," in *Proceedings of EMNLP*, 2009.

[15] P. Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.

[16] P. Denis and J. Baldridge, "Joint determination of anaphoricity and coreference resolution using integer linear programming," in *Proceedings of NAACL*, 2007.

[17] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme," in *Proceedings of MUC*, 1995.

[18] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *Proceedings of LREC*, 1998.

[19] X. Luo, "On coreference resolution performance metrics," in *Proceedings of HLT-EMNLP*, 2005.