Feature Sharing in the Generation and Interpretation of Nominals in Dialogue

Pamela W. Jordan
Intelligent Systems Program & Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA 15260
pjordan@pitt.edu

Abstract

In a fully collaborative, mixed-initiative dialogue it is necessary to both interpret and generate nominal descriptions and so the question of the extent to which these processes can share knowledge is helpful for deciding what to include in the dialogue history and for gaining insight for generation into what enhances interpretation and vice versa. Although theoretical and symbolic models for the interpretation and generation of nominals share much in common, this is not necessarily the case with the statistical models that have been tried. There have been no studies to compare these feature sets and examine whether features that have provided good results for interpretation will do so for generation and vice versa. In this paper, we describe work in progress to do so. So far we have tested generation features for 3 models on the interpretation task and found that they could provide a significant contribution.

Introduction

Both the generation and interpretation of nominals have been widely studied and many theories and symbolic approaches have been proposed ((Appelt 1985; Dale 1992; Heeman & Hirst 1995; Passonneau 1996; Jordan 2000b; van Deemter 2001; Gardent 2002; Grosz 1977; Webber 1978; Sidner 1983; Grosz & Sidner 1986; Vieira & Poesio 2000) *inter alia*). The main factors that are common in this work is the need to consider the discourse and task structure for determining saliency of objects, the recency of last mention, and the frequency of mention. While statistical approaches have been tried in recent times for both interpretation ((Strube, Rapp, & Müller 2002; Ng & Cardie 2002) are a few of the most recent) and generation (Jordan & Walker 2000), when we compare the features used in the statistical models for interpretation and generation, there are few commonalities and there have been no studies that examine whether the features that have provided good results for interpretation would also do so for generation and vice versa.

In a fully collaborative, mixed-initiative dialogue it is necessary to both interpret and generate nominal descriptions and so the question of the extent to which these processes can share knowledge is helpful for deciding what to include in the dialogue history (Clark & Marshall 1981b). For content selection during generation, it is generally worthwhile to give some consideration to the ramifications of particular choices on the successful outcome of interpretation. Conversely, it is also generally worthwhile to consider what motivates generation to give insights on how to successfully decode what is communicated. In this case it may be helpful to have some understanding during generation of features that enhance interpretation and vice versa.

The reason to focus on statistical approaches over symbolic ones (some of which are informed by the same models as the symbolic approaches) is that the interpretation and generation processes can be fine-tuned to a particular domain. Another reason for this focus is because statistical approaches are currently showing better results at correctly resolving some types of anaphoric references and at matching human performance in selecting the attributes to express.

In this paper we will describe work we have in progress for determining which statistical features can be shared by interpretation and generation. So far we have tested the generation features used in (Jordan & Walker 2000) and subsequently updated in (Jordan & Walker 2002) for interpretation and are in the process of evaluating how those same features effect interpretation performance.

Background

Generating Nominals

In previous work, (Jordan & Walker 2000) empirically compared the utility of features important for representing three theoretically complementary models of nominal generation (Jordan & Walker 2002). The feature sets were based on these generation models:

- CONTRAST SET features, inspired by the INCREMENTAL MODEL of (Dale & Reiter 1995);
- CONCEPTUAL PACT features, inspired by the models of Clark and colleagues (Clark & Wilkes-Gibbs 1986; Brennan & Clark 1996);
- INTENTIONAL INFLUENCES features, inspired by the model of (Jordan 2000b).

Dale and Reiter's INCREMENTAL MODEL focuses on the production of near-minimal descriptions that allow the hearer to reliably distinguish the task object from similar task objects. Following (Grosz & Sidner 1986), Dale and Reiter's algorithm utilizes discourse structure as an important factor in determining which objects the current object must be distinguished from. The model of Clark, Brennan and Wilkes-Gibbs is based on the notion of CONCEPTUAL PACTS, i.e. the conversants attempt to coordinate with one another by establishing a conceptual pact for describing an object. Jordan's INTENTIONAL INFLUENCES model is based on the assumption that the underlying task-related inferences required to achieve the task goals are an important factor in content selection for non-minimal descriptions.

(Jordan & Walker 2000) found that the features from these three models were significantly more accurate predictors of a human's choice for attributes in the context of a dialogue than the majority class baseline in which the most frequent combination of attributes for a corpus of dialogues is always included in a description. In subsequent experiments (Jordan & Walker 2002) have been improving upon the feature-based representations of the three models and have achieved a significant improvement in accuracy at predicting a human's choice of attributes (so far have improved from 50% to 59.9%).

Interpreting Nominals

With the interpretation of nominals, many feature sets also have been tested ((Strube, Rapp, & Müller 2002; Ng & Cardie 2002) *inter alia*). These models have utilized string features, positional features within the discourse, grammatical features, and lexical semantic features. Many of the features relate a new nominal that is to be interpreted to potential antecedents in the discourse. The positional features capture recency information but none of the features attempt to capture other aspects of the discourse structure.

In our initial studies of interpretation we are also interested in seeing the effect on resolution when we combine generation-inspired features with some of the existing interpretation features and in getting a baseline performance measure for resolution for the corpus we are using in this study. Since we do not already have grammatical and lexical semantic features annotated in the corpus we are using for training and testing, we are initially limited to the string and positional features.

One such set of string features described in the literature is a measure of the minimum edit distance between an anaphor and its candidate antecedent. The minimum edit distance is the number of edit operations needed to transform a source string into a target string. We will initially use just the minimun edit distance as defined in (Strube, Rapp, & Müller 2002) in our study because our corpus is annotated for both pronouns and definite noun phrases and their definition was shown to significantly improve overall performance for resolving both pronominal and definite noun phrases when combined with other types of interpretation features.

Methods, Corpus and Data

Our study to date compares the accuracy of a set of nominal expression interpretors that vary according to the feature sets used to train them. The feature sets are grouped by the models that inspired them (as described above). We use machine learning to train and test these nominal-expression interpretors on a set of 332 anaphoric expressions from the corpus of COCONUT dialogues. We evaluate the interpretors by comparing their predictions against a corpus of annotated discourse relations. The interpretors predict which (if any) discourse relation holds between an existing discourse entity in the dialogue history and a description that has just been uttered in the dialogue. By building separate interpretors for each feature set and combinations of feature sets, we can quantify the contributions of each feature set to the task of reference resolution.

G: That leaves us with 250 dollars. I have *a yellow rug for 150 dollars*. Do you have any other furniture left that matches for 100 dollars?"
S: No, I have no furniture left that costs $100. I guess you can buy *the yellow rug for $150*.
G: Okay. I'll buy *the rug for 150 dollars*. I have *a green chair* that I can buy for 100 dollars that should leave us with no money.
S: That sounds good. Go ahead and buy *the yellow rug* and *the green chair*.
G: I'll buy *the green 100 dollar chair*. Design Complete?
S: Sounds good, do you want *the green chair* in the dining room with *the other chairs*? I put *the yellow rug* in the living room. Then the design is complete.
G: Sounds good. Hit the design complete

Figure 1: Excerpt of a COCONUT dialogue illustrating nominal expressions to generate and interpret

We are using the rule learning program RIPPER (Cohen 1996) for training and testing against the annotated corpus. Like other learning programs, RIPPER takes as input the names of a set of *classes* to be learned, the names and ranges of values of a fixed set of *features,* and *training data* specifying the class and feature values for each example in a training set. Its output is a *classification model* for predicting the class of future examples. In RIPPER, the classification model is learned using greedy search guided by an information gain metric, and is expressed as an ordered set of if-then rules.

Corpus

The COCONUT corpus is a set of 24 computer-mediated dialogues consisting of a total of 1102 utterances. The dialogues were collected in an experiment where two human subjects collaborated on a simple design task, that of buying furniture for two rooms of a house (Di Eugenio *et al.* 1998). An excerpt of a COCONUT dialogue is shown in Figure 1. The participants' main goal is to negotiate the purchases; the items of highest priority are a sofa for the living room and a table and four chairs for the dining room. The participants also have specific secondary goals which further constrain the problem solving task. Participants are instructed to try to meet as many of these goals as possible, and are motivated to do so by rewards associated with satisfied goals. The secondary goals are: 1) match colors within a room, 2) buy as much furniture as you can, 3) spend all your money. The participants are told what rewards are associated with achieving each goal.

Each participant is given a separate budget and inventory of furniture. Neither participant knows what is in the other's inventory or how much money the other has. By sharing information during the conversation, they can combine their budgets and select furniture from each other's inventories. The participants are equals and purchasing decisions are joint. In the experiment, each set of participants solved one to three scenarios with varying inventories and budgets. The problem scenarios varied task complexity by ranging from tasks where items are inexpensive and the budget is relatively large to tasks where the items are expensive and the budget relatively small.

To illustrate the problem of generating and interpreting nominals in dialogue, consider the dialogue in Figure 1. In the process of negotiating the solution, the dialogue participants interpret and generate the nominal expressions (shown in italics) describing the items of furniture.

Each furniture type in the COCONUT task domain has four associated attributes: color, price, owner and quantity. A nominal expression generator must decide which of these four attributes to include in the generated expression. For example, the task domain objects under discussion in the dialogue in Figure 1 are a $150 yellow rug owned by G and a $100 dollar green chair owned by S. The yellow rug is described first as *a yellow rug for 150 dollars* and then subsequently as *the yellow rug for 150 dollars*, *the rug for 150 dollars*, *the yellow rug*. It could also have been described and understood given any of the following non-pronominal expressions: *the rug, my rug, my yellow rug, my $150 yellow rug, the $150 rug*. The content of these descriptions varies depending on which attributes are included in the description. How does the speaker decide which attributes to include and does the hearer need to be aware of this decision-making to better understand which object is referenced? And does awareness of the task negotiations enter into the interpretation and generation of the nominals? For example, consider that *the other chairs* (near the end of the excerpt) could be challenging to interpret without considering the current state of the task negotiations.

Corpus Annotation

After the corpus was collected it was annotated by human coders for two types of features. The DISCOURSE ENTITY LEVEL annotations provide discourse reference information from which initial representations of discourse entities and updates to them can be derived, and explicit attribute usage information that reflects how each discourse entity was evoked. For example, the initial representation for "I have a yellow rug. It costs $150." would include type, quantity, color and owner following the first utterance. Only the quantity attribute is inferred. After the second utterance the entity would be updated to include price. The UTTERANCE LEVEL ANNOTATIONS capture the problem solving state in terms of goals, constraint changes and the size of the solution set for the current constraint equations as well as current variable assignments. The utterance level discourse features encode when an offer is made and the level of a speaker's commitment to a proposal under consideration, i.e. conditional or unconditional.

In order to derive some of the discourse information the task structure must be identified. The COCONUT corpus was encoded via a set of instructions to coders to record all domain goals. Changes to a different domain goal or action were used as a cue to derive the non-linguistic task structure (Terken 1985; Grosz & Sidner 1986). Each domain action provides a discourse segment purpose so that each utterance that relates to a different domain action or set of domain actions defines a new segment. The encoded features all have good intercoder reliability (Di Eugenio *et al.* 1998; Jordan 2000b).

Class Assignment

We are trying to learn which of 6 possible discourse relationships exists between the pairing of a nominal that has just been introduced into the discourse and a potential antecedent. The discourse relationships supported by the annotation are coreference, set, class, predicative, common noun anaphora or none.

To illustrate the discourse relationships between nominal expressions, first consider (1). It is an example of a set/subset discourse relationship between *the green set* and the three distinct discourse entities *2 $25 green chairs*, *2 $100 green chairs* and *$200 green table*.

(1) a. : I have [2 $25 green chairs] and [a $200 green table].

b. : I have [2 $100 green chairs]. Let's get [the green set].

A class discourse relationship is illustrated in (2) where the type of the discourse entity for *your green one* is inherited from the discourse entity for *the table*.

(2) Let's decide on [the table] for the dining room. How about [your green one]?

The common noun anaphora discourse relationship labels cases of one anaphora and null anaphora. For example, in (3), each of the marked NPs in the last part of the utterance has a null anaphora relationship to the marked NP in the first part.

(3) I have [a variety of high tables] ,[green], [red] and [yellow] for 400, 300, and 200.

Discourse entities can also be related by predicative relationships such as *is*. For example, in (4), the entities defined by *my cheapest table* and *a blue one for $200* are not the same discourse entities but the information about one provides more information about the other and the discourse entities point to the same physical object.

(4) [My cheapest table] is [a blue one for $200].

We explain how we use the annotations to construct the features in more detail below.

Feature Extraction

In RIPPER, feature values are continuous (numeric), set-valued, or symbolic. We encoded each nominal description pairing in terms of a set of 69 features that were either

Feature Sets	Feature Category	Features	Values
given-new	reference relation	ante-reference-relation	initial,coref,set,class...
inherent (INH)	dialogue specific	problem-number, speaker-pair ana-utterance-number	number, symbol, number
	expression specific	ana-attribute-value: color,price...	red,500...
conceptual pact (CP)	similarities	attributes-agree: color,price...	boolean
		attr-similarities	number
	recency	distance-btwn-ante-ana	number
	ante frequency	freq-attr-expressed: color...	number
		number-prev-mentions	number
	stability history for ante	pact-given-last-2-occasions	boolean
		pact-given-last-3-occasions	boolean
	descriptions	attribute-expressed-in: ante,ana	boolean
	describers of pairing	same-speakers	boolean
contrast set (CS)	attribute distractors	count-of-attr-distractors: color,price...	number
	saliency of distractor attributes	majority-value-for-attr: color,price...	red,500,...
intentional influences (IINF)	task situation for ante	goals,constraints	symbols
	agreement state for ante	influence-on-listener, commit-speaker, solution-size	(action-directive,info-request...) (offer,commit), (det,indet)
	solution-interactions for ante	attr-contrast: color,price	boolean
minimum edit distance (MED)	edit distances between antecedent and anaphor	ante-med, ana-med	number

Table 1: Features listed by feature sets. ANTE = candidate antecedent, ANA = target anaphora

directly annotated by humans as described above, derived from annotated features or inherent to the dialogue (Di Eugenio *et al.* 1998; Jordan 2000b). The dialogue context in which each description occurs is represented in the encodings. Table 1 summarizes the features used in training and testing, grouped by model.

The GIVEN-NEW features encode fundamental attributes of the candidate antecedent that is to be tested for a relationship to the target interpretation expression. It encodes whether the candidate was new (`initial`), given (`coref`) or discourse inferred relative to the discourse history (Clark & Marshall 1981b; Prince 1981). The types of inferences supported by the annotation are set, subset, class and common noun anaphora (e.g. one and null anaphora) (Jordan 2000b). While the generation models tested in (Jordan & Walker 2000) encoded what was mutually known about the discourse entity, this was not relevant for interpreting the target since this is what is to be established by interpretation. However, in refinements of these experiments we will add what is mutually known about the candidate antecedent.

The INHERENT FEATURES in Table 1 are an encoding of particulars about the discourse situation, such as the speaker pair, the task (represented by the problem number), the position within the dialogue and the target entity's expressed attribute values for its five possible attributes. While we don't expect these dialogue specific and the attribute value features to generalize to other dialogue situations, it allows us to examine whether there are individual differences in interpretation algorithms for a speaker pair, specific properties of the object, the location within the dialogue, or the problem difficulty.

The CONCEPTUAL PACT model suggests that dialogue participants negotiate a description that both find adequate for describing an object (Clark & Wilkes-Gibbs 1986; Brennan & Clark 1996). The speaker generates trial descriptions that the hearer interprets and modifies based on which object he thinks he is suppose to identify. The negotiation continues until the participants are confident that the hearer has correctly identified the intended object. The additional features suggested by this model include the previous descriptions of the candidate antecedent since that is the description that is potentially being negotiated, and how long ago the candidate description was made relative to the target anaphor. If the description has stabilized that would indicate that the negotiation process had been completed. Once a pact is established, the expression may shorten further so we expect frequency and similarity measures to be helpful in predicting when shortening is possible.

The CONCEPTUAL PACT features in Table 1 encode attribute value agreement and attribute similarity, when the entity was last described relative to the target in terms of number of utterances and markables, how frequently the candidate was described and the frequency with which its attributes were expressed, a stability history for the description of the target, which attributes were used to describe the target and candidate and whether the person describing the target and candidate are the same.

The INCREMENTAL MODEL builds a description incrementally by considering the other objects that are currently expected to be in focus for the hearer(Dale & Reiter 1995). These other objects are called *distractors*. The basic idea is to add attributes as necessary until any distractors are ruled out as competing co-specifiers. Based on these ideas, we developed a set of features we call CONTRAST SET features, as shown in Table 1. The goal of our encoding is to represent whether there are distractors present in the focus space which might motivate the inclusion of a particular attribute. Our representation only approximates the INCREMENTAL MODEL since it utilizes a preferred salience ordering of attributes and eliminates distractors as attributes are added to a description. For example, adding the attribute *type* when the object is a chair, eliminates any distractors that aren't chairs. Our encoding treats attributes instead of objects as distractors but this interpretation of Dale and Reiter's model was shown in (Jordan 2000b) to perform similarly to the strict model.

An open issue with deriving the distractors is how to define a focus space (Walker 1996). We use two focus space definitions, one based on recency, and the other on intentional structure. For intentional structure we utilize the task goal segmentation encoded in the COCONUT corpus as discussed above (CS-SEG). For recency, we simply consider the entities from the previous utterance as possible distractors (CS-1UTT). For each focus space definition, the encoding includes a count of the number of attribute values that are different from the candidate for each attribute, the most salient value among the distractors for each attribute and the frequency of the salient values for the distractors.

Jordan (Jordan 2000a) proposed a model to select attributes for nominals called the INTENTIONAL INFLUENCES model. This model posits that the task-related inferences and the agreement process for task negotiation are important factors in selecting attributes. The features used to approximate Jordan's model are in Table 1 and are all relative to the target in the pairing. The task situation features encode inferrable changes in the task situation that are related to target attributes. The agreement state features encode critical points of agreement during the problem solving involving the target. These are features that (Di Eugenio *et al.* 2000) found to be indicative of agreement states and include DAMSL features (*influence-on-listener, commit-speaker*) (Allen & Core 1997), and progress towards a solution (*solution-size*). The solution interactions features represent situations where multiple proposals were under consideration which may have contrasted with one another in terms of solving color-matching goals (*color-contrast*) or price related goals (*price-contrast*). We don't expect these features to significantly add to the resolution process since it is primarily about identification. We expect instead that these features may help in only a small number of cases (as in the case of *the other chairs* in the dialogue excerpt in Figure 1). In future experiments we will also consider these features at the point just before the target expression is issued.

The final feature set is the minimum edit distance features. These features represent the number of edit operations (insert, delete, substitute) needed to transform the target expression into the candidate expression and vice versa.

Learning Trials

The final input for learning is training data, i.e., a representation of discourse relationships between two nominal expressions in terms of feature and class values.

Our experimental data is 504 nominal descriptions from 13 dialogues of the COCONUT corpus as well as features constructed from the annotations described above. Of the 504 nominal expressions, 332 have discourse relationships to other nominal expressions in the corpus. As in (Strube, Rapp, & Müller 2002), we paired a potential anaphoric description with a candidate antecedent, and classified it positively if there was a discourse relationship between the two and negatively otherwise. But since our data is all of the nominal descriptions in the corpus and not just the pronouns, we further subcategorized the positive classifications into either coreference, set, class, predicative or common noun anaphora relationships.

To reduce the number of negative instances, we filtered out those negative cases in which the type attributes or the color attributes disagree since these are less often likely to disagree in any of the positive reference relationship cases. The final dataset contains 1554 instances of pairings with 192 of these being instances of coreferences, 111 sets, 76 classes, 13 common noun anaphora, 2 predicative and 1160 negative, with a baseline accuracy of 74.65% if we simply assume all relationships are negative.[1] To further emphasize the positive discourse relationships in the data, we weighted each positive case by 2 (i.e. this has the effect of doubling the number of positive cases) yielding a baseline accuracy of 59.5% for guessing that there is no discourse relationship between a nominal expression and any previously introduced discourse entity.

In order to induce rules from a variety of feature representations, our training data is represented differently in different trials. First, examples are represented using only the GIVEN-NEW features in Table 1 to establish a performance baseline for given-new information. In addition the MED features in Table 1 provide an initial baseline for features that have significantly boosted performance in other interpretation trials but that have no related features in the nominal generation models. Then other generation related feature sets are added in to GIVEN-NEW to examine their individual contribution, culminating with combined feature sets.

The output of each machine learning trial is a model for determining discourse relationships between an anaphor and its possible antecedents for this domain and task, learned from the training data. To evaluate these models, the error rates of the learned models are estimated using 25-fold cross-validation, i.e. the total set of examples is randomly divided into 25 disjoint test sets, and 25 runs of the learning program are performed. Thus, each run uses the examples not in the test set for training and the remaining examples for testing.

Results

Table 2 summarizes our results. For each feature set, we report accuracy rates and standard errors resulting from cross-validation. Accuracy rates are statistically significantly different when the accuracies plus or minus twice the standard error do not overlap (Cohen 1995), p. 134. It is clear that

[1] The positive cases do not sum to 332 because an expression can have discourse relationships with more than one antecedent.

performance depends on the features that the learner has available. The 59.5% MAJORITY CLASS BASELINE accuracy rate in the first row is a standard baseline that corresponds to the accuracy one would achieve from simply classifying a pairing as having no discourse relationship.

First we see that the current GIVEN-NEW model does not improve accuracy beyond the baseline. Next, we see that the MINIMUM EDIT DISTANCE model makes a significant improvement beyond the baseline as we expected. What is surprising is that the INTENTIONAL INFLUENCES model makes the same sort of improvement as the MINIMUM EDIT DISTANCE model and the two CONTRAST SET models. Finally, we see that the CONCEPTUAL PACT model performs significantly better than all of the single feature set models and that when we combine other features with the CONCEPTUAL PACT model there is no further significant improvement.

Feature Sets Used	Accuracy (SE)
MAJORITY CLASS BASELINE	59.5 %
GIVEN-NEW	59.9% (1.5)
GIVEN-NEW,IINF	63.6% (1.3)
GIVEN-NEW,CS-1UTT	65.4% (1.5)
GIVEN-NEW,CS-SEG	65.9% (1.7)
MED	66.5 % (1.3)
GIVEN-NEW, INH	67.4% (1.6)
GIVEN-NEW,CP	82% (1.2)
GIVEN-NEW,CP,CS-SEG	81.4% (1.4)
GIVEN-NEW,CP,CS-1UTT	82.7% (1.2)
GIVEN-NEW,CP,MED	82.4% (1.3)
GIVEN-NEW,CP,IINF	83.3% (1.3)
GIVEN-NEW,IINF,CP,MED,CS-SEG	81.4% (1.1)
GIVEN-NEW,IINF,CP,MED,CS-1UTT	82.3% (1.2)

Table 2: Accuracy rates for Nominal Interpretation using different feature sets, SE = Standard Error. MED = the MINIMUM EDIT DISTANCE features. CP = the CONCEPTUAL PACT features. IINF = the INTENTIONAL INFLUENCES features. INH = the INHERENT features. CS-SEG = the CONTRAST-SET, SEGMENT features. CS-1UTT = the CONTRAST SET, ONE UTTERANCE features.

In trials in which we isolated the similarity features from the rest of the conceptual pact model features, we find that the similarity features have an accuracy of 77.5% (1.4) while the remaining conceptual pact features have an accuracy of 75.1% (1.3). The similarity features are of high value because they are statistically similar to the combined set while the remainder of the conceptual pact features are significantly worse. The similarity features are conceptually similar to the minimum-edit distance features but does not a priori generalize across all the attributes for an object.

Discussion and Future Work

In our continued study of feature sharing between the interpretation and generation of nominals we will be more closely comparing the other string, grammatical, lexical semantic and positional features used in interpretation with the features we have introduced from the generation models to see what other commonalities exist and will test the interpretation features that are not yet represented in the generation models to see if they will in turn boost generation performance. For example, all of the previously created interpretation models try to model recency by representing the distance between the candidate and target expression in terms of markables and utterances just as the CONCEPTUAL PACT generation features inspired us to do. So this feature is already covered in the generation-inspired feature set.

In this preliminary evaluation, we have not yet fully taken advantage of the CONTRAST SET model in that so far we are training on all the candidates in the utterance containing the nominal to be interpreted and the utterance containing the antecedent and all that are in between the two utterances in order to reduce the number of negative training instances (similarly to what has been done in other interpretation models). While this is a reasonable, cheap approximation of what is in focus, it is possible that the discourse structure will indicate that some of the intervening candidates are no longer in focus because the discourse segment purpose has been achieved and has been removed from the focus space. In our future work we will reduce the candidate pool to just those that are part of the distractor set. Furthermore, in a run-time environment, the model will not have a mechanism for pruning out unlikely candidates without some model of possible distractors.

Finally, at the end of the study, we will need to compute recall and precision measures so that we can compare our results with those of others. This will give us a better sense of how the other interpretation models perform on our data relative to the other corpora on which they have been tested. Finally we will also need to do a MUC-style evaluation in which we assess the accuracy with which the generation-inspired models detect reference chains.

We need to stress that our results are still preliminary and additional refinements are needed to the generation-inspired features. However, relative to the MINIMUM EDIT DISTANCE model, the generation-inspired features are showing promise of being valuable during interpretation. Our future studies will show us if the interpretation features that are not already represented in the generation models will do likewise for generating nominal expressions.

Once our studies are complete, we plan to train interpretation and generation models using the best feature sets for each and integrate these two models into an existing dialogue system (e.g. Why2-Atlas (VanLehn *et al.* 2002)). By including more than coreference relations in the interpretation process, the system can infer additional information about the new objects that are introduced into the dialogue on the basis of the other objects that are already represented in the dialogue history.

All of the features that we have tested so far on the generation side are easily available but since our goal is to use these features during dialogue processing we expect the uncertainties about how to interpret previous descriptions to effect the quality of the attribute choices made. This is because many of the feature values depend on the dialogue history (and not just previous generation choices). On the interpretation side, some of the generation-inspired features are more difficult to obtain because of the need to infer the goals of the dialogue participant.

References

Allen, J., and Core, M. 1997. Draft of DAMSL: Dialog act markup in several layers. Available from http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html, under *Tools and resources*.

Appelt, D. 1985. *Planning English Sentences*. Studies in Natural Language Processing. Cambridge University Press.

Brennan, S. E., and Clark, H. H. 1996. Lexical choice and conceptual pacts in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Clark, H. H., and Marshall, C. R. 1981a. Definite reference and mutual knowledge. In Joshi, A.; Webber, B.; and Sag, I., eds., *Linguistics Structure and Discourse Setting*. Cambridge, England: Cambridge University Press. 10–63.

Clark, H. H., and Marshall, C. R. 1981b. Definite reference and mutual knowledge. In Joshi; Webber; and Sag., eds., *Elements of Discourse Understanding*. Cambridge: CUP. 10–63.

Clark, H. H., and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition* 22:1–39.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*. Boston: MIT Press.

Cohen, W. 1996. Learning trees and rules with set-valued features. In *Fourteenth Conference of the American Association of Artificial Intelligence*.

Dale, R., and Reiter, E. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Dale, R. 1992. *Generating Referring Expressions*. ACL-MIT Series in Natural Language Processing. The MIT Press.

Di Eugenio, B.; Jordan, P. W.; Moore, J. D.; and Thomason, R. H. 1998. An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proceedings of the Thirty-sixth Conference of the Association for Computational Linguistics*.

Di Eugenio, B.; Jordan, P. W.; Thomason, R. H.; and Moore, J. D. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies* 53(6):1017–1076.

Gardent, C. 2002. Generating minimal definite descriptions. In *Proceedings of Association for Computational Linguistics 2002*.

Grosz, B., and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12:175–204.

Grosz, B. J. 1977. The representation and use of focus in dialogue understanding. Technical Report 151, Artificial Intelligence Center, SRI International, 333 Ravenswood Ave, Menlo Park, Ca. 94025.

Heeman, P. A., and Hirst, G. 1995. Collaborating on referring expressions. *Computational Lingusitics* 21(3).

Jordan, P. W., and Walker, M. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of Association for Computational Linguistics 2000*.

Jordan, P. W., and Walker, M. 2002. Learning to generate nominal expressions: Experiments with the coconut corpus. *Journal of Artificial Intelligence Research* submitted.

Jordan, P. W. 2000a. Can nominal expressions achieve multiple goals?: An empirical study. In *Proceedings of ACL 2000*.

Jordan, P. W. 2000b. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. Dissertation, Intelligent Systems Program, University of Pittsburgh.

Ng, V., and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of Association for Computational Linguistics 2002*.

Passonneau, R. J. 1996. Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech* 39(2-3):229–264.

Prince, E. 1981. Toward a Taxonomy of Given-New Information. In Cole, P., ed., *Radical Pragmatics*. Academic Press. 223–255.

Sidner, C. L. 1983. Focusing in the comprehension of definite anaphora. In Brady, M., and Berwick, R., eds., *Computational Models of Discourse*. MIT Press.

Strube, M.; Rapp, S.; and Müller, C. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of Empirical Methods in Natural Language Processing Conference*.

Terken, J. M. B. 1985. *Use and Function of Accentuation: Some Experiments*. Ph.D. Dissertation, Institute for Perception Research, Eindhoven, The Netherlands.

van Deemter, K. 2001. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics* to appear.

VanLehn, K.; Jordan, P. W.; Rosé, C.; Bhembe, D.; Böttner, M.; Gaydos, A.; Makatchev, M.; Pappuswamy, U.; Ringenberg, M.; Roque, A.; Siler, S.; Srivastava, R.; and Wilson, R. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Systems Conference*.

Vieira, R., and Poesio, M. 2000. An empirically based system for processing definite descriptions. *Conputational Linguistics* 26(4):539–593.

Walker, M. A. 1996. Limited attention and discourse structure. *Computational Linguistics* 22(2):255–264.

Webber, B. L. 1978. *A Formal Approach to Discourse Anaphora*. Ph.D. Dissertation, Harvard University. Garland Press.