Data-Driven Interaction Patterns: Authority and Information Sharing in Dialogue

Elijah Mayfield, Michael Garbus, David Adamson, and Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
{*emayfiel, mgarbus, dadamson, cprose*}*@cs.cmu.edu*

Abstract

We explore the utility of a computational framework for social authority in dialogue, codified as utterance-level annotations. We first use these annotations at a macro level, compiling aggregate statistics and showing that the resulting features are predictive of group performance in a task-based dialogue. Then, at a micro level, we introduce the notion of an interaction pattern, a formulation of speaker interactions over multiple turns. We use these patterns to characterize situations where speakers do not share information equally. These patterns are found to be more discriminative at this task than similar patterns using standard dialogue acts.

Introduction

Complex dialogue systems rely on many components of natural language understanding. Beyond the technical and engineering challenges, such as speech recognition, more complex systems must also account for elements of discourse, such as strategies for introducing new information and repairing after a misunderstanding, building models of speakers' shared knowledge and conflicting conceptual pacts, and recognizing the communicative and illocutionary intent of speaker utterances.

Much progress has been made in solving these problems in recent years, and many components have been incorporated into real-world systems. We explore the utility of the Negotiation framework, an annotation scheme inspired by sociolinguistic study of authority and information flow in sequences of utterances in a dialogue (Martin and Rose 2003). Our prior work has shown that these codes can be automatically applied with high accuracy (Mayfield and Rosé 2011).

In previous work, we have studied this framework from a primarily social standpoint. Important constructs from collaborative learning, such as negative affective behavior, task engagement (Howley, Mayfield, and Rosé In Press), and group self-efficacy (Howley, Mayfield, and Rosé 2011), have been shown to correlate with aspects of Negotiation.

Here, we study the utility of Negotiation for a more cognitive process: the completion of a task where information is not perfectly shared between speakers. Half of the task-related objects are visible only to one speaker in a pair, meaning that information sharing is critical to task performance. However, sharing information is still a social process, so we believe that a sociolinguistic framework may be able to identify key information sharing behaviors.

A social framework for sharing information in dialogue makes sense for analysis of pairs in task-based dialogue. Interactional style plays a major role in the way that information is contributed to a discourse, whether it is through cautious introduction of new information (Carletta 1992) or monitoring the understanding of your listener (Brown and Dell 1987). One common aspect of previous work has been to characterize interaction style through dialogue acts, either by examining distributions of dialogue act tags (Boyer et al. 2011) or sequences of tags (Forbes-Riley and Litman 2005). However, these tags are insufficient for capturing the notion of social authority. As we demonstrate in both experiments in this paper, features from the Negotiation framework are more robust in describing the exchange of information between speakers.

To showcase the usefulness of Negotiation for expressing these social relationships, we will attempt two tasks. First, we study the problem of task performance prediction. This requires features representing a whole-dialogue interaction, so we use macro-level statistics about authoritative behavior, using the Negotiation framework as the source of these statistics. Compared to a robust bag-of-words model, we find that social authority features are more predictive of task success. We also find that more traditional representations of information coordination, using task-related object reference, are insufficiently expressive to predict performance.

Second, we dig deeper into *how* information is being shared between speakers. We find all spans of interactions in our corpus where a landmark is being referenced, and divide data into information status - shared, or privileged (visible) to one of the two speakers. We then characterize, through data-driven methods, the behaviors in privileged information settings. We extract patterns based on spans of utterances in dialogue, utilizing stretchy patterns (Gianfortoni, Adamson, and Rosé 2011) and the Negotiation framework annotations. These patterns can be extracted quickly and automatically, and qualitative analysis shows that they align with prior findings about information sharing in dialogue.

For our experiments, we analyze the MapTask corpus (Anderson et al. 1991), which has been studied extensively

in prior work. In each dialogue, a pair of participants are each given a map. The maps share many landmarks but also have several landmarks with differences in placement, name, or even whether they exist. One participant, the instruction giver, has a path on their map, and they must direct the other participant, the instruction follower, to reproduce that path.

In several instances in this paper, we will make reference to dialogue acts and references to task-based objects; in these cases, we use the gold standard, human annotations from the corpus. The dialogue act scheme in the MapTask corpus consists of thirteen tags (which we treat as non-hierarchical), denoting general-purpose actions such as "instruct," "clarify," or "query." References are marked as intra-utterance spans of text and are annotated with the map landmark they correspond to.

Outlining the structure of the paper: We will first provide an overview of similar prior work, both in shared knowledge research and on our corpus in particular, followed by an introduction to the Negotiation framework. We show, by predicting group task performance, that certain authoritativeness behaviors predict successful interactions. We then describe our stretchy pattern extraction, using both Negotiation labels and dialogue acts, and comparing the resulting patterns. We conclude with directions for expanding this work in the future, both in improved representation and in dialogue system implementation.

Prior Work

Information flow is a key element of our work, and it is an important concept for cases where information is not perfectly shared between speakers. A key element of the MapTask corpus is that only half of the reference landmarks on each pair of maps are shared, and the remainder are privileged to only one speaker. Prior work has studied these situations through many lenses, such as the Gricean principle of parsimony (Shadbolt 1984) or Clark's principle of Least Collaborative Effort (Clark 1996). For example, it has been shown that people speak egocentrically by default, assuming information is shared until being proven otherwise (Schober 1995). Intelligibility of referring expressions has also been shown to decrease when it is speaker-old information, even if it is addressee-new (Bard et al. 2000).

Our work hinges on automatically identifying dialogue strategies for sharing information. Strategies have been studied before, identified through qualitative study of corpora. Early work on the MapTask corpus studied task planning and recovery strategies between speakers (Carletta 1992), which attempted to model discourse parameters such as expected difference between speaker understanding or the care with which explanations are crafted. An example strategy from this work was the way in which speakers "planned to fail" by giving minimal information and repairing misunderstandings later. Dialogue strategies may also involve monitoring others for misunderstanding in order to maintain common ground (Clark and Krych 2004); on the other hand, when aware of a listener's impediments to understanding, speakers were found to adjust their speech quickly (Lockridge and Brennan 2002).

Many of these strategies have been confirmed empirically, even within the MapTask corpus, either larger theories being examined on small scales (Davies 2010) or for specific problems, such as how the collaborative principle interacts with locative reference and spatial relations (Viethen and Dale 2008). Planning of discourse strategies has been shown to be effective in real-world dialogue-systems, for repairing misunderstandings (Bohus 2007) or inferring intention (Rich, Sidner, and Lesh 2000). Early work also studied specific forms of introduction (Anderson and Boyle 1994) or specific effects of information visibility (Boyle, Anderson, and Newlands 1994).

The Negotiation Framework

For our definition of authority in social interaction, we specifically focus on the use of the Negotiation framework, which attempts to describe how speakers use their role as a source of knowledge or action to position themselves relative to others in a discourse (Martin and Rose 2003).

The Negotiation framework is primarily made up of four main codes, K1, K2, A1, and A2. Numerous rare or highly specific codes from the sociolinguistic literature were discarded to ensure that a machine learning classification task would not be overwhelmed with many infrequent classes.

The four main codes are divided on two axes, illustrated in Figure 1. First, is the utterance related to exchanging information, or to exchanging services and actions? If the former, then it is a K move (knowledge); if the latter, then an A move (action). Second, is the speaker acting as a primary or secondary source of action or knowledge? In the case of knowledge, this often corresponds to the difference between assertions (K1) and queries (K2). For instance, a statement of fact or opinion is a K1:

g	K1	well i've got a great viewpoint here just below the east lake

By contrast, asking for someone else's knowledge or opinion is a K2:

g	K2	what have you got underneath the east lake
f	K1	a tourist attraction

In the case of action, the codes usually correspond to narrating action (A1) and giving instructions (A2), as below:

g	A2	go almost to the edge of the lake
f	A1	yeah okay

All moves that do not fit into one of these categories are classified as other (o). This includes back-channel moves, floor-grabbing moves, false starts, preparatory moves, and any other non-contentful contributions.

With these codes, one application we explore is a quantitative measure of authoritativeness for each speaker. This is the number of authoritative moves divided by the number of authoritative and non-authoritative moves. In the specific, task-based domain of this corpus, we mark A2 (instruction giving) and K1 (knowledge giving) moves as authoritative and A1 (instruction following) and K2 (knowledge requesting) as non-authoritative moves.

In our previous work, this framework was formalized and its application was automated (Mayfield and Rosé 2011).

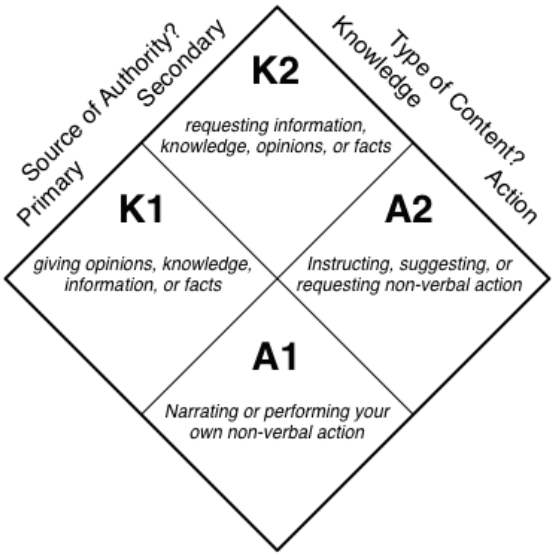

Figure 1: The main codes of the Negotiation framework.

On a per-line basis, agreement with human annotations was found to be $\kappa = 0.58$, compared to $\kappa = 0.71$ inter-annotator agreement. In addition, for a whole-conversation Authoritativeness ratio for a given speaker, automated judgements were found to correlate almost perfectly with human judgements, $r^2 = 0.947$.

Task Success Prediction

We first study authoritativeness to predict task performance in the MapTask corpus. Task success is measured in terms of how far the route that the follower has drawn deviates from the route shown on the giver's map, measured in square centimeters between the drawn and original path.

We predict whether a given interaction will fall above or below the median error of the corpus. We categorize groups into good performance (error $< 56cm^2$), and poor performance (error $\geq 56cm^2$). We are given only the transcription of the conversation, with associated annotations - utterance boundaries, speech transcription, dialogue act annotation, and references to landmarks are all assumed to be given, though each of these is in itself a separate task. We also treat Negotiation labels as given, with 20 manually coded dialogues and 108 dialogues coded automatically with the system from (Mayfield and Rosé 2011).

A key observation from prior work (Carletta 1992) is the concept of cautiousness in dialogue. This suggests that speakers, especially instruction givers, who request more information are qualitatively more likely to perform well. We believe that authority, as framed by the Negotiation framework, can bring this behavior out more clearly than simpler measures, such as counting the number of questions asked.

To illustrate this, we used annotated dialogue acts and counted, in each dialogue, the number of questions (query-yn and query-w tagged lines) asked, by each speaker and the total count. These features represent a shallow measure of "cautiousness." However, we found that they were not significantly discriminative ($p = .23 - .51$) and were not able to predict task success (κ of -0.147). On the other hand, as our data will show, authoritativeness as defined through Negotiation aligns with the expected quantitative results from these previous insights.

Feature Sets

We describe several sets of features. As a baseline, we use a bag-of-words model, which has repeatedly proven to be a robust baseline for classification tasks in language technology. Each word occurring in the corpus vocabulary is treated as one feature, with a boolean value representing whether that word occurs in the dialogue.

Condition Features Our initial set of features are based on the condition. Different dialogues in the corpus were separated by three conditions: relationship to the other participant (either speakers were previously acquainted or were strangers); eye contact (speakers were either blocked or had full view of each others' faces); and map (sixteen different maps were used). The choice of map in particular was highly significant ($p < .01$), with mean error per map ranging from $33cm^2$ all the way to $138cm^2$.

Reference Features Our next set of features is based on reference. We automatically extract several features using the annotations provided with the MapTask corpus. For each speaker, we add features based on their referencing behavior. This includes the total number of utterances they made, the total number of references they made to landmarks, and the average number of references per line. We also observe the number of times they make "cross reference" to landmarks they cannot see on their map - information which is privileged to the other participant. We count the total number of these cross references, the average number of cross references per line, and the percentage of their references that were cross references.

We add two features comparing relative proportions of references made by each speaker - both the ratio of total number of references, and the ratio of references made to non-visible landmarks. We then add features based on the aggregate number of references made - one feature for the average number of references per line, and one for the longest span of utterances in which no reference is made.

Authority Features Finally, we study the impact of authority-based features. Three features are included in the initial task success experiment based purely on authority: instruction giver authoritativeness ratio, instruction follower authoritativeness ratio, and the difference between speakers' authoritativeness ratios.

We also examine the particular use of A2 moves by the instruction giver, which usually correspond to the lines that are actual instructions to draw. We first calculate the percent of "complex" instructions, which we define as the percent of A2 moves which contain some variant of the terms "slope," "curve," or "round." These are likely to indicate more attention to detail than simpler instructions like "up" or "left."

Condition Features	
friends	Indicates participants were acquainted before dialogue.
eyes	Indicates participants could make eye contact during dialogue.
map	Indicates map (one of sixteen).
mapError	Indicates average error of training dialogues in this map condition.

Table 1: Condition-based features, requiring no analysis of the conversation transcript.

Reference Features	
refs/Line	Average number of landmark references per line of dialogue.
speakerLines*	Lines of dialogue.
speakerRefs*	References to landmarks.
spRefs/Line*	Average landmark references per line.
crossRefs*	References to non-visible landmarks.
crossRefs/Line*	Non-visible landmark references per line.
cross%*	Percent of references made to non-visible landmarks.
lines/Landmark	Average number of lines from first to last mention of each landmark.
refs/Landmark	Average number of references made to each landmark.
refRatio	Ratio of giver landmark references to follower landmark references.
crossRatio	Ratio of giver non-visible landmark references to follower non-visible landmark references.
maxGap	Longest span of utterances in which no landmark reference is made.

Table 2: Features based on reference behavior only. Features marked with * are calculated for each speaker separately.

Experimental Results

We perform a per-conversation classification into good and poor performance. As each conversation represents only one data point, we are faced with a problem of data sparsity. Because of this, we use leave-one-conversation-out cross-validation. We evaluate our models both on accuracy and kappa. All experiments were performed using SIDE (Mayfield and Rosé 2010) using its Weka plugin implementation of support vector machines (Witten and Frank 2002). All results are summarized in Table 4.

To our knowledge, the only prior attempt to predict task success in the MapTask corpus was framed as a regression problem (Reitter and Moore 2007), and as such is not directly comparable in these measures, so we reproduce their experimental setup (10-fold cross-validation using SVM regression) and also present those results, measuring r^2, the amount of variance in performance explained by a model. The findings of that work, which focused on lexical repetition and priming between speakers, are complementary rather than opposed to the findings of this work, as they explore a fundamentally different aspect of speaker interaction.

Authority Features	
giverAuth	Authoritativeness of instruction giver.
followerAuth	Authoritativeness of follower.
authDiff	Difference between speaker authoritativeness ratios.
complex%	Percent of A2 moves with more complex directions.

Table 3: Aggregate Authoritativeness-based features.

Features	Feats.	Acc.	κ	r^2
Unigrams	2218	63.28	.266	.185
Condition Features	4	63.28	.267	.056
Reference Features	18	50.00	-.001	.011
Authority Features	4	64.06	.281	.096
Reference + Authority	22	61.72	.234	.093
Condition + Authority	8	71.09	.422	.161
Reference + Condition	22	64.84	.298	.094
All Non-Unigram	**26**	**72.66**	**.454**	**.216**
Previous Best	-	-	-	.17

Table 4: Results of experiments with various feature sets, with size of feature space also given. The previous best result was set by (Reitter and Moore 2007).

Our best model incorporates all features and utilizing both condition and authority based features, easily outperforms a unigram model, by 9% ($p < .05$). It also outperforms previous work by a smaller margin; without access to their raw data we cannot compute significance.

We find that authority related features are strongly predictive of group task performance. All four features that we extract are given high weight in the resulting SVM model. Qualitatively, we find that the following characteristics are indicative of good performance: low giver authority, high follower authority, and a small gap between speaker authority. This matches the predictions from (Anderson and Boyle 1994) and suggests that those cases where speakers are closer to equal participants in the task led to better performance in the end.

As mentioned above, condition alone is a major indicator of performance. The most influential are choice of map and acquaintance of the speakers. Though prior work has found that eye contact increases communication efficiency (Boyle, Anderson, and Newlands 1994), there is no significant impact on error in task performance.

Unigrams were also found to be surprisingly weak at this task. While they match the performance of our condition and authority based features, they require far more features to do so (over 2,000 for the entire vocabulary). What is normally a robust and very predictive feature space, in problems such as sentiment analysis, produces fair, at best, predictiveness in this task. On the other hand, unigrams perform well at the regression task, outperforming the repetition-based model of prior work.

Perhaps the weakest result in this set of features are those features from reference. If anything, the results are overstated, as they are derived from gold-standard reference annotations, rather than being automatically labelled. Some features are given substantial weight - in particular, cross reference features are highly weighted - but on the whole, these features did not improve on chance agreement.

An ablation study of our feature space shows that both condition and authority features are contributing a statistically significant improvement to accuracy ($p < .02$ and $p < .05$, respectively). The reference features that we use improve performance by 1.57%, but this difference is not statistically significant. Surprisingly, though reference features did not outperform random chance on their own, they improve the performance of both condition and authority features when added, though not by a significant amount.

One insight we can gain from reference features is that follower cross-% (as defined in Table 2. This is not surprising, as spending time discussing information that the follower cannot then use directly to perform the task is likely to be unproductive. It also hinges on information sharing behavior in the specific case of privileged-information interactions. This leads into our next set of experiments. We would like to be able to characterize how information is shared not just as an aggregate statistic (percent of references to privileged information), but in micro-level detail. For this, we need to be able to characterize shorter exchanges between speakers. To do this, we extract patterns of interaction, as described in the next section.

Interaction Patterns

Our definition of a pattern is an extension of a previous application of "stretchy patterns" (Gianfortoni, Adamson, and Rosé 2011). A pattern is comprised of a series of tokens which can be drawn from a small number of classes. These tokens also encode the speaker of an utterance. A token may also be a gap, which is allowed to consume up to some number of concrete tokens. In our case, we set the range of allowed pattern sizes to be $3-6$ tokens, with gaps (marked by $\square$) allowed to consume from $1-3$ tokens. These patterns thus resemble skip n-grams (Guthrie et al. 2006), however, the location of their gaps are enforced rather than being allowed at any point in the pattern.

This definition of a stretchy pattern was first used at the word level, for gender attribution in blog posts. Here we extend it to utterance-level labels, such that a single token in a pattern corresponds to a single utterance in a dialogue. An advantage of using generic classes as tokens is that while a bag of words model is often limited by topic or domain dependency, the topic of discussion and related keywords are not at all tied to the stretchy pattern representation.

This definition of an interaction pattern has multiple advantages. First, the findings of those papers are further verified if they can be rediscovered automatically. Second, an automatic way of identifying certain strategies can be helpful for dialogue system implementation. Finally, the same automatic process may be used to gain further insights about dialogue strategy beyond what has already been studied.

Pattern	**Predictive of:**	κ
gA2 □ go	Shared Knowledge	.322
gA2 □ gA2	Shared Knowledge	.310
gK2 fK1 □ gA2	Shared Knowledge	.269
→ □ ←	Giver-Privileged	.288
→ gK2 □ ←	Giver-Privileged	.176
→ fK1 go	Follower-Privileged	.204
→ fK1 □ gA2	Follower-Privileged	.129

Table 5: Highlighted discriminative Negotiation patterns.

Pattern	**Predictive of:**	κ
→ □ fReply-Y	Shared Knowledge	.316
gInstruct □ gInstruct	Shared Knowledge	.286
gQueryYN □ gInstruct	Shared Knowledge	.217
→ □ fReply-N	Giver-Privileged	.412
→ □ ←	Giver-Privileged	.288
→ □ gReply-N	Follower-Privileged	.188
→ □ gExplain	Follower-Privileged	.077

Table 6: Highlighted discriminative Dialogue Act patterns.

To illustrate how these interactions are converted into sequences for pattern extraction, consider an interaction sequence `gK1 fo go gA2 fA1`. From this, we can extract n-grams up to 6 tokens long (such as `gK1 fo go` or `fo go gA2 fA1`), as well as patterns with gaps, such as `gK1 □ gA2` or `fo go □ fA1`. In total, the number of stretchy patterns that can be extracted grows rapidly but roughly linearly with the length of the interaction.

Pattern Extraction: Results and Analysis

We divide our data into 1,827 interactions, one for each landmark mentioned in each dialogue. That interaction is then labelled with whether the landmark is shared knowledge, privileged to the instruction giver, or privileged to the instruction follower. Each interaction represents the span from the first utterance containing a mention of a landmark in a conversation to the last utterance, and is marked with beginning-of-utterance ($\rightarrow$) and end-of-utterance ($\leftarrow$) tokens to allow patterns access to this information.

In each interaction, we extract all possible patterns using the parameters described above, and measure each pattern's utility using κ, or the feature's discriminative ability above chance. We measure discriminative ability for two different formulations - giver-privileged landmarks against shared landmarks, and follower-privileged landmarks against shared landmarks.

Some top-ranking features are presented in Tables 5 and 6. Immediately, similarities emerge. In both cases, it was observed that multiple instructions for action in series from the instruction giver are predictive of a shared landmark, where communication is unhindered. However, the predictiveness of this feature is stronger in the case of Negotiation labels ($\kappa = 0.310$ compared to 0.226). This is likely because requests for action may take multiple forms, and while the most common is a direct instruction, the illocutionary force

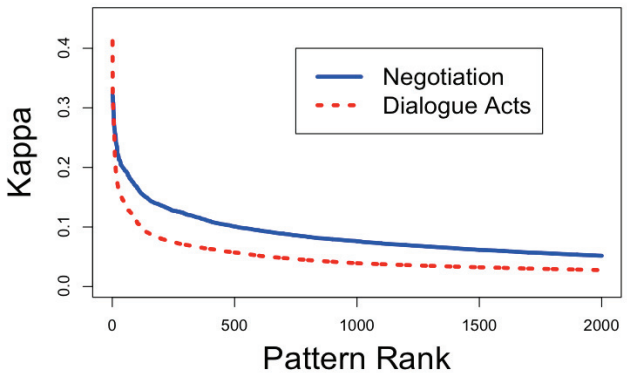

Figure 2: Range of κ values for the top 2,000 features in both Negotiation and Dialogue Act patterns, showing that the dropoff in discriminative power is substantially slower for Negotiation patterns.

is not being accounted for in dialogue acts.

We also see that mere length of time that a landmark is being referenced is highly discriminative. The pattern $\rightarrow \square \leftarrow$ represents any interaction no more than three turns long, and is highly predictive of giver-privileged objects. In examining the data, this often is the result of an instruction being given, the follower responding that they are unable to follow the instruction (they cannot see the landmark), and the dialogue moving on, with the failed landmark reference never being mentioned again. This pattern further confirms the egocentric, planned-failure strategies described in (Carletta 1992; Schober 1995; Clark and Krych 2004), among others.

In general, fewer patterns are highly discriminative in the dialogue act patterns. We also see very few predictive dialogue act patterns beyond three tokens in length, compared to several Negotiation patterns, including the highlighted feature, `gK2 fK1 □ gA2`. This pattern is especially interesting for multiple reasons. First, it complements the observations of forms of introduction discussed in (Anderson and Boyle 1994). Second, it represents what could be considered a "prototypical" interaction - an instruction giver requests information, which is given, and then after some small number of backchannel moves between speakers, an instruction is issued. Such a pattern occurs too infrequently, or in too many different varying combinations, to arise in the dialogue act patterns.

It is tempting to explain this based on merely a larger number of possible dialogue act patterns, due to the increased number of classes (an issue which has been taken up recently in the dialogue act tagging community, for instance in (Bunt et al. 2010)). This cannot be the whole explanation, however; 2,670,866 dialogue act patterns are extracted from our corpus, while only slightly fewer, 2,204,366 patterns are extracted using Negotiation patterns. As Figure 2 shows, the top 2,000 most discriminative Negotiation patterns maintain predictive power far longer than dialogue act patterns. These results suggest that the smaller number of classes in Negotiation labels, then, are not only just aggregating dialogue act categories, but instead are cross-cutting in ways that more clearly describe the actions taken by speakers in relaying information to one another.

One advantage for this task that arises in dialogue act patterns is the distinction made between positive and negative answers to questions. While these patterns are grouped into a K1 (information giving) label with Negotiation labels, the distinction is enough to allow both positive and negative replies to feature heavily in the most discriminative dialogue act patterns. In particular, one pattern, an instruction follower replying negatively early in an interaction, is the most predictive of any single pattern extracted in either case, with $\kappa = 0.412$.

Several high-ranking patterns have been filtered from these lists because of high overlap with each other. This problem, which interferes with the independence assumption of many machine learning algorithms, has been documented before. Before the patterns can be effectively utilized as a class of feature space, more work is needed on feature selection or construction. Previous work has utilized standard information content metrics (Jurafsky et al. 1998), feature construction using high-precision features (Arora et al. 2010), or feature subsumption to remove shorter duplicates of long and predictive sequences (Riloff, Patwardhan, and Wiebe 2006). This is an element of future work with these patterns.

Conclusions and Future Directions

This work presents multiple new uses for the Negotiation framework, which can be coded automatically with high accuracy. First, we attempted to predict task performance, and found that the Negotiation framework was highly predictive, and a very simple model using only condition- and authority-based features significantly outperformed a unigram baseline. In addition, social authority features outperformed aggregate statistics for reference behavior on this prediction task.

To facilitate more detailed analysis, we presented the notion of interaction patterns, which describe interactions between speakers utilizing a stretchy series of utterance-level annotations. These patterns were found to be well correlated with the shared information status of a series of turns discussing a task-related reference. In comparison to the identical method using standard dialogue acts, the Negotiation framework was found to degrade in predictive power much more slowly, suggesting that this representation more clearly represents the transfer of information between speakers.

Overall, our findings have further confirmed the hypotheses and observations that sociolinguists and social scientists have made on smaller scales. However, in this case, the patterns which emerged and the prediction tasks they were used for were on a much larger scale and all feature extraction was done completely automatically. This suggests that the Negotiation framework offers a promising new direction for testing hypotheses about social interactions in dialogue.

The next step in this research is to combine these directions, and make use of micro-level analysis on a whole-conversation task. Our formulation of stretchy patterns may

be useful not only for predicting information status, but also for describing successful interactions in terms of task or dialogue success. It may also be worthwhile to study variation across conditions, for instance by separately extracting patterns for high performance and low performance groups and comparing their differences.

We found that reference-based features have little impact in aggregate, though numerous studies we have mentioned describing the impact of reference style on interactions. This suggests that there is benefit from a more local and contextualized representation of how references are being made, rather than a global count or averaged representation.

This leads into the next future direction. The current formulation of Negotiation represents only a structural interaction sequence - a request for information being followed by giving that information, for example. What is not included is any direct annotation of connectedness between sequences of interaction. For example, there is no direct way to model a landmark being introduced with a K2-initial sequence, followed by a clarifying K2-initial sequence, followed by an A2-initial instruction sequence. For this, sequences would need to be tied together into longer threads of conversation. We will attempt this level of annotation in future work.

Recent work has studied the multiparty domain for dialogue systems, identifying strategies for recovery from misunderstanding and non-understanding when relying on speech recognition (Bohus 2007), strategies for social interaction and idea generation (Kumar, Beuth, and Rosé 2011), or turn-taking in a multi-party interaction (Bohus and Horvitz 2011). As we have shown, our formulation of Negotiation and interaction patterns has been effective in both large and fine-grained scope. It has demonstrated that previously researched patterns of interaction in privileged-knowledge settings can be extracted in a wholly automatic way, based entirely on corpus data. Thus, a logical next step for our work is to incorporate the recognition of these patterns in situated dialogue systems, to test the effectiveness of these patterns, for instance, by recognizing speaker inequality through dialogue behavior.

Acknowledgements

This research was supported by Office of Naval Research grants N000141010277 and N000141110221.

References

Anderson, A., and Boyle, E. 1994. Forms of introduction in dialogues: their discourse contexts and communicative consequences. In *Language and Cognitive Processes*.

Anderson, A.; Bader, M.; Bard, E.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H.; and Weinert, R. 1991. The hcrc map task corpus. In *Language and Speech*.

Arora, S.; Mayfield, E.; Rosé, C. P.; and Nyberg, E. 2010. Sentiment classification using automatically extracted subgraph features. In *NAACL Workshop on Emotion in Text*.

Bard, E.; Anderson, A.; Sotillo, C.; Aylett, M.; Doherty-Sneddon, G.; and Newlands, A. 2000. Controlling the intelligibility of referring expressions in dialogue. In *Memory and Language*.

Bohus, D., and Horvitz, E. 2011. Multiparty turn taking in situated dialog. In *Procedings of SIGDIAL*.

Bohus, D. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. Dissertation.

Boyer, K.; Grafsgaard, J.; Ha, E. Y.; Phillips, R.; and Lester, J. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the Association for Computational Linguistics*.

Boyle, E.; Anderson, A.; and Newlands, A. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. In *Language and Speech*.

Brown, P., and Dell, G. 1987. Adapting production to comprehen- sion: The explicit mention of instruments. In *Cognitive Psychology*.

Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; Soria, C.; and Traum, D. 2010. Towards an iso standard for dialogue act annotation. In *International Conference on Language Resources and Evaluation*.

Carletta, J. 1992. *Risk-Taking and Recovery in Task-Oriented Dialogue*. Ph.D. Dissertation.

Clark, H., and Krych, M. 2004. Speaking while monitoring addressees for understanding. In *Memory and Language*.

Clark, H. 1996. *Using Language*.

Davies, B. 2010. Principles we talk by: Testing dialogue principles in task-oriented dialogues. In *Pragmatics*.

Forbes-Riley, K., and Litman, D. 2005. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In *Proceedings of SIGDIAL*.

Gianfortoni, P.; Adamson, D.; and Rosé, C. P. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *EMNLP Workshop on Modelling of Dialects and Language Varieties*.

Guthrie, D.; Allison, B.; Liu, W.; Guthrie, L.; and Wilks, Y. 2006. A closer look at skip-gram modelling. In *Language Resources and Evaluation Conference*.

Howley, I.; Mayfield, E.; and Rosé, C. P. 2011. Missing something? authority in collaborative learning. In *Proceedings of Computer Supported Collaborative Learning*.

Howley, I.; Mayfield, E.; and Rosé, C. P. In Press. Linguistic analysis methods for studying small groups. In *International Handbook of Collaborative Learning*.

Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; and Ess-Dykema, C. V. 1998. Switchboard discourse language modelling final report. Technical report.

Kumar, R.; Beuth, J.; and Rosé, C. P. 2011. Conversational strategies that support idea generation productivity in groups. In *Proceedings of Computer Supported Collaborative Learning*.

Lockridge, C., and Brennan, S. 2002. Addressees needs

influence speakers early syntactic choices. In *Psychonomic Bulletin and Review*.

Martin, J., and Rose, D. 2003. *Working with Discourse: Meaning Beyond the Clause*.

Mayfield, E., and Rosé, C. P. 2010. An interactive tool for supporting error analysis for text mining. In *NAACL Demonstration Session*.

Mayfield, E., and Rosé, C. P. 2011. Recognizing authority in dialogue with an integer linear programming constrained model. In *Proceedings of Association for Computational Linguistics*.

Reitter, D., and Moore, J. 2007. Predicting success in dialogue. In *Proceedings of ACL*.

Rich, C.; Sidner, C.; and Lesh, N. 2000. Collagen: Applying collaborative discourse theory to human-computer interaction.

Riloff, E.; Patwardhan, S.; and Wiebe, J. 2006. Feature subsumption for opinion analysis. In *Proceedings of EMNLP*.

Schober, M. 1995. Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations? In *Discourse Processes*.

Shadbolt, N. 1984. *Constituting reference in natural language: the problem of referential opacity*. Ph.D. Dissertation.

Viethen, J., and Dale, R. 2008. The use of spatial relations in referring expression generation. In *Conference on Natural Language Generation*.

Witten, I., and Frank, E. 2002. *Data mining: practical machine learning tools and techniques with Java implementations*.