Modeling Socio-Cultural Phenomena in Discourse

Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Samira Shaikh, Sarah Taylor[†] and Nick Webb
University at Albany, SUNY
[†]Lockheed Martin Corporation
tomek@albany.edu

Abstract

In this paper, we describe a novel approach to computational modeling and understanding of social and cultural phenomena in multi-party dialogues. We developed a two-tier approach in which we first detect and classify certain social language uses, including topic control, disagreement, and involvement, that serve as first order models from which presence the higher level social constructs such as leadership, may be inferred.

1. Introduction

We investigate the language dynamics in small group interactions across various settings. Our focus in this paper is on English online chat conversations; however, the models we are developing are more universal and applicable to other conversational situations: informal face-to-face interactions, formal meetings, moderated discussions, as well as interactions conducted in languages other than English, e.g., Urdu and Mandarin.

Multi-party online conversations are particularly interesting because they become a pervasive form of communication within virtual communities, ubiquitous across all age groups. In particular, a great amount of communication online occurs in virtual chat-rooms, typically conducted using a highly informal text dialect. At the same time, the reduced-cue environment of online interaction necessitates more explicit linguistic devices to convey social and cultural nuances than is typical in face-to-face or even voice conversations.

Our objective is to develop computational models of how certain social phenomena such as leadership, power, and conflict are signaled and reflected in language through the choice of lexical, syntactic, semantic and conversational forms by discourse participants. In this paper we report the results of an initial phase of our work during which we constructed a prototype system called DSARMD-1 (Detecting Social Actions and Roles in Multiparty Dialogue). Given a representative segment of multiparty task-oriented dialogue, DSARMD-1 automatically classifies all discourse participants by the degree to which they deploy selected *social language uses*, such as topic control, task control, involvement, and disagreement. These are the mid-level social phenomena, which are deployed by discourse participants in order to achieve or assert higher-level *social constructs*, including leadership. In this work we adopted a two-tier empirical approach where social language uses are modeled through *observable* linguistic features that can be automatically extracted from dialogue. The high-level social constructs are then inferred from a combination of language uses attributed to each discourse participant; for example, a high degree of influence and a high degree of involvement by the same person may indicate a leadership role. In this paper we limit our discussion to the first tier only: how to effectively model and classify social language uses in multi-party dialogue.

2. Related Research

Issues related to linguistic manifestation of social phenomena have not been systematically researched before in computational linguistics; indeed, most of the effort thus far was directed towards the communicative dimension of discourse. While the Speech Acts theory (Austin, 1962; Searle, 1969) provides a generalized framework for multiple levels of discourse analysis (locution, illocution and perlocution), most current approaches to dialogue focus on information content and structural components (Blaylock, 2002; Car-

berry & Lambert, 1999; Stolcke, et al., 2000) in dialogue; few take into account the effects that speech acts may have upon the social roles of discourse participants. Also relevant is research on modeling sequences of dialogue acts – to predict the next one (Samuel et al. 1998; Ji & Bilmes, 2006 inter alia) – or to map them onto subsequences or "dialogue games" (Carlson 1983; Levin et al., 1998), which are attempts to formalize participants' roles in conversation (e.g., Linell, 1990; Poesio & Mikheev, 1998; Field et al., 2008).

There is a body of literature in anthropology, linguistics, sociology, and communication on the relationship between language and power, as well as other social phenomena, e.g., conflict, leadership; however, existing approaches typically look at language use in situations where the social relationships are known, rather than using language predictively. For example, conversational analysis (Sacks et al., 1974) is concerned with the structure of interaction: turn-taking, when interruptions occur, how repairs are signaled, but not what they reveal about the speakers. Research in anthropology and communication has concentrated on how certain social norms and behaviors may be reflected in language (e.g., Scollon and Scollon, 2001; Agar, 1994) with few systematic studies attempting to explore the reverse, i.e., what the linguistic phenomena tell us about social norms and behaviors.

3. Data & Annotation

Our initial focus has been on on-line chat dialogues. While chat data is plentiful on-line, its adaptation for research purposes presents a number of challenges that include users' privacy issues on the one hand, and their complete anonymity on the other. Furthermore, most data that may be obtained from public chat-rooms is of limited value for the type of modeling tasks we are interested in due to its high-level of noise, lack of focus, and rapidly shifting, chaotic nature, which makes any longitudinal studies virtually impossible. To derive complex models of conversational behavior, we need the interaction to be reasonably focused on a task and/or social objectives within a group.

Few data collections exist covering multi-party dialogue, and even fewer with on-line chat. Moreover, the few collections that exist were built primarily for the purpose of training dialogue act tagging and similar linguistic phenomena; few if any of these corpora are suitable for deriving pragmatic models of conversation, including socio-linguistic phenomena. Existing resources include a multi-person meeting corpus ICSI-MRDA and the AMI Meeting Corpus (Carletta, 2007), which contains 100 hours of meetings captured using synchronized recording devices. Still, all of these resources look at spoken language rather than on-line chat. There is a parallel interest in the online chat environment, although the development of useful resources has progressed less. Some corpora exist such as the NPS Internet chat corpus (Forsyth and Martell, 2007), which has been hand-anonymized and labeled with part-of-speech tags and dialogue act labels. The StrikeCom corpus (Twitchell et al., 2007) consists of 32 multi-person chat dialogues between players of a strategic game, where in 50% of the dialogues one participant has been asked to behave 'deceptively'.

It is thus more typical that those interested in the study of Internet chat compile their own corpus on an as needed basis, e.g., Wu et al. (2002), Khan et al. (2002), Kim et al. (2007).

Driven by the need to obtain a suitable dataset we designed a series of experiments in which recruited subjects were invited to participate in a series of on-line chat sessions in a specially designed secure chat-room. The experiments were carefully designed around topics, tasks, and games for the participants to engage in so that appropriate types of behavior, e.g., disagreement, power play, persuasion, etc. may emerge spontaneously. These experiments and the resulting corpus have been described elsewhere (Shaikh et al., 2010b), and we refer the reader to this source. Ultimately a corpus of 50 hours of English chat dialogue was collected comprising more than 20,000 turns and 120,000 words. In addition we also assembled a corpus of 20 hours of Urdu chat.

A subset of English language dataset has been annotated at four levels: communication links, dialogue acts, local topics and meso-topics (which are essentially the most persistent local topics). Although full details of these annotations are impossible to explain within the scope of this article, we briefly describe them below. Annotated datasets were used to de-

velop and train automatic modules that detect and classify social uses of language in discourse. It is important to note that the annotation has been developed to support the objectives of our project and does not necessarily conform to other similar annotation systems used in the past.

- *Communicative links*. In a multi-party dialogue an utterance may be directed towards a specific participant, a subgroup of participants or to everyone.
- *Dialogue Acts*. We developed a hierarchy of 15 dialogue acts for annotating the functional aspect of the utterance in discussion. The tagset we adopted is based on DAMSL (Allen & Core, 1997) and SWBD (Jurafsky et al., 1997), but compressed to 15 tags tuned significantly towards dialogue pragmatics and away from more surface characteristics of utterances (Shaikh et al., 2010a).
- *Local topics*. Local topics are defined as nouns or noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun.
- *Topic reference polarity*. Some topics, which we call *meso-topics*, persist through a number of turns in conversation. A selection of meso-topics is closely associated with the task in which the discourse participants are engaged. Meso-topics can be distinguished from the local topics because the speakers often make polarized statements about them.

4. Socio-linguistic Phenomena

We are interested in modeling the social phenomena of Leadership and Power in discourse. These high-level phenomena (or Social Roles, *SR*) will be detected and attributed to discourse participants based on their deployment of selected *Language Uses* (LU) in multi-party dialogue. Language Uses are mid-level socio-linguistic devices that link linguistic components deployed in discourse (from lexical to pragmatic) to social constructs obtaining for and between the participants. The language uses that we are currently studying are *Agenda Control*, *Disagreement*, and *Involvement* (Broadwell et al., 2010).

Our research so far is focused on the analysis of English-language synchronous chat, and we are looking for correlations between various metrics that can be used to detect LU in multiparty dialogue. We expect that some of these correlations may be culturally specific or language-specific, as we move into the analysis of Urdu and Mandarin discourse in the next phase of this project.

4.1 Agenda Control in Dialogue

Agenda Control is defined as efforts by a member or members of the group to advance the group's task or goal. This is a complex LU that we will model along two dimensions: (1) *Topic Control* and (2) *Task Control*. Topic Control refers to attempts by any discourse participants to impose the topic of conversation. Task Control, on the other hand, is an effort by some members of the group to define the group's project or goal and/or steer the group towards that goal. We believe that both behaviors can be detected using scalar measures per participant based on certain linguistic features of their utterances.

For example, one hypothesis is that topic control is indicated by the rate of *local topic introductions* (LTI) per participant (Givon, 1983). Local topics may be defined quite simply as noun phrases introduced into discourse, which are subsequently mentioned again via repetition, synonym, pronoun, or other form of co-reference. Thus, one measure of topic control is the number of local topics introduced by each participant as percentage of all local topics in a discourse.

Using an LTI index we can construct assertions about topic control in a discourse. For example, suppose the following information is discovered about the speaker LE in a multi-party discussion *dialogue-1*[1] where 90 local topics are identified:

1. LE introduces 23/90 (25.6%) of local topics in this dialogue.
2. The mean rate of local topic introductions is this dialogue is 14.29%, and standard deviation is 8.01.
3. LE is in the top quintile of participants for introducing new local topics

We can now claim the following, with a degree of confidence (to be determined):

$$TopicControl_{LTI}(LE, 5, dialogue\text{-}1)$$

We read this as follows: *speaker LE exerts the highest degree of topic control in dialogue-1*.

[1] Dialogue-1 refers to an actual dataset of 90-minute chat among 7 participants, covering approximately 700 turns. The task is to select a candidate for a job given a set of resumes.

Of course, LTI is just one source of evidence and we developed other metrics to complement it. We mention three of them here:

- *SMT Index*. This is a measure of topic control suggested in (Givon, 1983) and it is based on subsequent mentions of already introduced local topics. Speakers who introduce topics that are discussed at length by the group tend to control the topic of the discussion. The *subsequent mentions of local topics* (SMT) index calculates the percentage of second and subsequent references to the local topics, by repetition, synonym, or pronoun, relative to the speakers who introduced them.
- *Cite Score*. This index measures the extent to which other participants discuss topics introduced by that speaker. The difference between SMT and CiteScore is that the latter reflect to what degree a speaker's efforts to control the topic are assented to by other participants in a conversation.
- *TL Index* (TL). This index stipulates that more influential speakers take longer turns than those who are less influential. The TL index is defined as the average number of words per turn for each speaker. Turn length also reflects the extent to which other participants are willing to 'yield the floor' in conversation.

Like LTI, all the above indices are mapped into a degree of topic control, based on quintiles in normal distribution (Table 1).

	LTI	**SMT**	**CS**	**TL**	AVG
LE	5	5	5	5	5.00
JR	4	4	4	3	3.75
KI	4	3	3	1	2.75
KN	3	5	4	4	4.00
KA	2	2	2	4	2.50
CS	2	2	2	2	2.00
JY	1	1	1	2	1.25

Table 1: Topic Control distribution in dialogue-1. Each row represents a speaker in the group (LE, JR, etc.). Columns show indices used, with degrees per speaker on 5-point scale based on quintiles in normal distribution, and the average value.

Ideally, all the above indices (and others yet to be defined) should predict the same outcome, i.e., for each dialogue participant they should assign the same degree of topic control, relative to other speakers. This is not always the case, and where the indices divert in their predictions, our level of confidence in the generated claims decreases. We are currently working on how these different metrics correlate to each other and how they should be weighted to maximize accuracy of making Topic Control claims. Nonetheless, we can already output a Topic Control map (shown in Table 1) that captures a sense of internal social dynamics within the group.

The other aspect of Agenda Control phenomenon is Task Control. It is defined as an effort to determine the group's goal and/or steer the group towards that goal. Unlike Topic Control, which is imposed by influencing the subject of conversation, Task Control is gained by directing other participants to perform certain tasks or accept certain opinions. Consequently, Task Control is detected by observing the usage of certain dialogue acts, including Action-Directive, Agree-Accept, Disagree-Reject, and related categories. Here again, we define several indices that allow us to compute a degree of Task Control in dialogue for each participant:

- *Directive Index* (DI). The participant who directs others is attempting to control the course of the task that the group is performing. We count the number of directives, i.e., utterances classified as Action-Directive, made by each participant as a percentage of all directives in discourse.
- *Directed Topic Shift Index* (DTSI). When a participant who controls the task offers a directive on the task, then the topic of conversation shifts. In order to detect this condition, we calculate the ratio of coincidence of directive dialogue acts by each participant with topic shifts following them.
- *Process Management index* (PMI). Another measure of Task Control is the proportion of turns each participant has that explicitly address the problem solving process. This includes utterances that involve coordinating the activities of the participants, planning the order of activities, etc. These fall into the category of Task (or Process) Management in most DA tagging systems.
- *Process Management Success Index* (PMSI). This index measures the degree of success by each speaker at controlling the task. A credit is given to the speaker whose suggested curse of action is supported by

other speakers for each response that supports the suggestion.[2] Conversely, a credit is taken away for each response that rejects or qualifies the suggestion. PMSI is computed as distribution of task management credits among the participants over all dialogue utterances classified as Task/Process Management.

As an example, let's consider the following information computed for the PMI index over *dialogue-1*:

1. *Dialogue-1* contains 246 utterances classified as Task/Process Management rather than doing the task.
2. Speaker KI makes 65 of these utterances for a PMI of 26.4%.
3. Mean PMI for participants is 14.3%; 80th percentile is >21.2%. PMI for KI is in the top quintile for all participants.

Based on this evidence we may claim (with yet to be determined confidence) that:

$TaskControl_{PMI}$(*KI, 5, dialogue-1*)

This may be read as follows: *speaker KI exerts the highest degree of Task Control in dialogue-1*. We note that Task Control and Topic Control do not coincide in this discourse, at least based on the PMI index. Other index values for Task Control may be computed and tabulated in a way similar to LTI in Table 1. We omit these here due to space limitations.

4.2 Disagreement in Dialogue

Disagreement is another language use that correlates with speaker's power and leadership. There are two ways in which disagreement is realized: *expressive disagreement* and *topical disagreement* (Stromer-Galley, 2007; Price, 2002). Both can be detected using scalar measures applied to subsets of participants, typically any two participants. In addition, we can also measure for each participant the rate with which he or she generates disagreement (with any and all other speakers). *Expressive Disagreement* is normally understood at the level of dialogue acts, i.e., when discourse participants make explicit utterances of disagreement, disapproval, or rejection in response to a prior speaker's utterance. Here is an example (KI and KA are two speakers in a multiparty dialogue in which participants discuss candidates for a youth counselor job):

KA: *CARLA... women are always better with kids*
KI: *That's not true!*
KI: *Men can be good with kids too*

While such exchanges are vivid examples of expressive disagreement, we are interested in more sustained phenomenon where two speakers repeatedly disagree, thus revealing a social relationship between them. Therefore, one measure of Expressive Disagreement that we consider is the number of Disagree-Reject dialogue acts between any two speakers as a percentage of all utterances exchanged between these two speakers. This becomes a basis for the *Disagree-Reject Index* (DRX). In dialogue-1 we have:

1. Speakers KI and KA have 47 turns between them. Among these there are 8 turns classified as Disagree-Reject, for the DRX of 15.7%.
2. *The mean DRX for speakers who make any Disagree-Reject utterances is 9.5%. The pair of speakers KI-KA is in the top quintile (>13.6%).*

Based on this evidence we can conclude the following:

$ExpDisagreement_{DRX}$ (*KI,KA, 5, dialogue-1*)

which may be read as follows: *speakers KI and KA have the highest level of expressive disagreement in dialogue-1*. This measure is complemented by a *Cumulative Disagreement Index* (CDX), which is computed for each speaker as a percentage of all Disagree-Reject utterances in the discourse that are made by this speaker. Unlike DRX, which is computed for pairs of speakers, the CDX values are assigned to each group participant and indicate the degree of disagreement that each person generates.

While Expressive Disagreement is based on the use of more overt linguistic devices, *Topical Disagreement* is defined as a difference in referential *valence* in utterances (statements, opinions, questions, etc.) made on a topic. Referential valence of an utterance is determined by the type of statement made about the topic in question, which can be positive (+), negative (–), or neutral (0). A positive statement is one in favor of (*express*

[2] The exact structure of the credit function is still being determined experimentally. For example, more credit may be given to first supporting response and less for subsequent responses; more credit may be given for unprompted suggestions than for those that were responding to questions from others.

advocacy) or in support of (*supporting information*) the topic being discussed. A negative statement is one that is against or negative on the topic being discussed. A neutral statement is one that does not indicate the speaker's position on the topic. Here is an example of opposing polarity statements about the same topic in discourse:

Sp-1: *I like that he mentions "Volunteerism and Leadership"*

Sp-2: *but if they're looking for someone who is experienced then I'd cross him off*

Detecting topical disagreement in discourse is more complicated because its strength may vary from one topic in a conversation to the next. A reasonable approach is thus to measure the degree of disagreement between two speakers on one topic first, and then extrapolate over the entire discourse. Accordingly, our measure of topical disagreement is valuation differential between any two speakers as expressed in their utterances about a topic. Here, the topic (or an "issue") is understood more narrowly than the local topic defined in the previous section (as used in Topic Control, for example), and may be assumed to cover only the most persistent local topics, i.e., topics with the largest number of references in dialogue, or what we call the *meso-topics*. For example, in a discussion of job applicants, each of the applicants becomes a meso-topic, and there may be additional meso-topics present, such as qualifications required, etc.

The resulting *Topical Disagreement Metric* (TDM) captures the degree to which any two speakers advocate the opposite sides of a meso-topic. TDM is computed as an average of *P*-valuation differential for one speaker (advocating *for* a meso-topic) and (*–P*)-valuation differential for the other speaker (advocating *against* the meso-topic).

Using TDM we can construct claims related to disagreement in a given multiparty dialogue of sufficient duration (exactly what constitutes a sufficient duration is still being researched). Below is an example based on a 90-minute chat dialogue-1 about several job candidates for a youth counselor. The discussion involved 7 participants, including KI and KA. Topical disagreement is measured on 5 points scale (corresponding to quintiles in normal distribution):

$TpDisAgree_{TDM}$(*KI,KA,"Carla",4,dialogue-1*)

This may be read as follows: speakers KI and KA *topically disagree to degree 4* on topic [job candidate] "Carla" in dialogue-1. In order to calculate this we compute the value of TDM index between these two speakers. We find that KA makes 30% of all positive utterances made by anyone about Carla (40), while KI makes 45% of all negative utterances against Carla. This places these two speakers in the top quintiles in the "for Carla" polarity distribution and "against Carla" distribution, respectively. Taking into account any opposing polarity statements made by KA against Carla and any statements made by KI for Carla, we calculate the level of topical disagreement between KA and KI to be 4 on the 1-5 scale.

TDM allows us to compute topical disagreement between any two speakers in a discourse, which may also be represented in a 2-dimensional table revealing another interesting aspect of internal group dynamics.

4.3 Involvement in Dialogue

The third type of social language use that we discuss in this paper is Involvement. Involvement is defined as a degree of engagement or participation in the discussion of a group. It is an important element of leadership, although its importance is expected to differ between cultures; in Western cultures, high involvement and influence (topic control) often correlates with group leadership.

In order to measure Involvement we designed several indices based on turn characteristics for each speaker. Four of the indices are briefly explained below:

- The *NP index* (NPI) is a measure of gross informational content contributed by each speaker in discourse. *NPI* counts the ratio of third-person nouns and pronouns used by a speaker to the total number of nouns and pronouns in the discourse.
- The *Turn index* (*TI*) is a measure of *interactional frequency*; it counts the ratio of turns per participant to the total number of turns in the discourse.
- The *Topic Chain Index* (*TCI*) counts the degree to which participants discuss of the most persistent topics. In order to calculate TCI values, we define a *topic chains* for all local topics. We compute frequency of mentions of these longest topics for each participant.

- The *Allotopicality Index* (*ATP*) counts the number of mentions of local topics that were introduced by other participants. An ATP value is the proportion of a speaker's allotopical mentions, i.e., excluding "self-citations", to all allotopical mentions in a discourse.

As an example, we may consider the following situation in *dialogue-1*:

1. *Dialogue-1* contains 796 third person nouns and pronouns, excluding mentions of participants' names.
2. Speaker JR uses 180 nouns and pronouns for an NPI of 22.6%.
3. The median NPI is 14.3%; JR are in the upper quintile of participants (> 19.9%).

From the above evidence we can draw the following claim:

$$Involvement_{NPI}(JR, 5, dialogue\text{-}1)$$

This may be read as: *speaker JR is the most involved participant in dialogue-1*.

As with other language uses, multiple indices for Involvement can be combined into a 2-dimensional map capturing the group internal dynamics.

5. Implementation & Evaluation

We developed a prototype automated DSARMD system that comprises a series of modules that create automated annotation of the source dialogue for all the language elements discussed above, including communicative links, dialogue acts, local/meso topics, and polarity. Automatically annotated dialogue is then used to generate language use degree claims. In order to evaluate accuracy of the automated process we conducted a preliminary evaluation comparing the LU claims generated from automatically annotated data to the claims generated from manually coded dialogues. Below we briefly describe the methodology and metrics used.

Each language use is asserted per a participant in a discourse (or per each pair of participants, e.g., for Disagreement) on a 5-point "strength" scale. This can be represented as an ordered sequence $LU_X(d_1, d_2, \ldots d_n)$, where LU is the language use being asserted, X is the index used, d_i is the degree of LU attributed to speaker i. This assignment is therefore a 5-way classification of all discourse participants and its correctness is measured by dividing the number of correct assignments by the total number of elements to be classified, which gives the micro-averaged precision. The accuracy metric is computed with several variants as follows:

1. *Strict mapping*: each complete match is counted as 1; all mismatches are counted as 0. For example, the outputs LU_X (5,4,3,2,1) and LU_X (4,5,3,1,1) produce two exact matches (for the third and the last speaker) for a precision of 0.4.
2. *Weighted mapping*: since each degree value d_i in $LU_X(d_1, d_2, \ldots d_n)$ represents a quintile in normal distribution, we consider the position of the value within the quintile. If two mismatched values are less than ½ quintile apart we assign a partial credit (currently 0.5).
3. *Highest – Rest*: we measure accuracy with which the highest LU degree (but not necessarily the same degree) is assigned to the right speaker vs. any other score. This results in binary classification of scores. The sequences in (1) produce 0.6 match score.
4. *High – Low*: An alternative binary classification where scores 5 and 4 are considered High, while the remaining scores are considered Low. Under this metric, the two sequences in (1) match with 100% precision.

The process of automatic assignment of language uses derived from automatically processed dialogues was evaluated against the control set of assignments based on human-annotated data. In order to obtain a reliable "ground truth", each test dialogue was annotated by at least three human coders (linguistics and communication graduate students, trained). Since human annotation was done at the linguistic component level, a strict inter-annotator agreement was not required; instead, we were interested whether in each case a comparable statistical distribution of the corresponding LU index was obtained. Annotations that produced index distributions dissimilar from the majority were eliminated.

Automated dialogue processing involved the following modules:

- *Local topics detection* identifies first mentions by tracking occurrences of noun phrases. Subsequent mentions are identified using fairly simple pronoun resolution (based mostly on lexical features), with Wordnet used to identify synonyms, etc.
- *Meso-topics* are identified as longest-chain

local topics. Their *polarity* is assessed at the utterance level by noting presence of positive or negative cue words and phrases.

- *Dialogue acts* are tagged based on presence of certain cue phrases derived from a training corpus (Webb et al., 2008).
- *Communicative links* are mapped by computing inter-utterance similarity based on n-gram overlap.

Preliminary evaluation results are shown in Tables 3-5 with average performance over 3 chat sessions (approx 4.5 hours) involving three groups of speakers and different tasks (job candidates, political issues). Topic Control and Involvement tables show average accuracy per index. For example, the LTI index, computed over automatically extracted local topics, produces Topic Control assignments with the average precision of 80% when compared to assignments derived from human-annotated data using the strict accuracy metric. However, automated prediction of Involvement based on NPI index is far less reliable, although we can still pick the most involved speaker with 67% accuracy. We omit the indices based on turn length (TL) and turn count (TI) because their values are trivially computed. At this time we do not combine indices into a single LU prediction. Additional experiments are needed to determine how much each of these indices contributes to LU prediction.

Topic Control	LTI	SMT	CS
Strict	0.80	0.40	0.40
Weighted	0.90	0.53	0.53
Highest-Rest	0.90	0.67	0.67
High-Low	1.00	0.84	0.90

Table 3: Topic Control LU assignment performance averages of selected indices over a subset of data covering three dialogues with combined duration of 4.5 hours with total of 19 participants (7, 5, 7 per session).

Involvement	NPI	TCI	ATP
Strict	0.31	0.42	0.39
Weighted	0.46	0.49	0.42
Highest-Rest	0.67	0.77	0.68
High-Low	0.58	0.74	0.48

Table 4: Involvement LU assignment performance averages for selected indices over the same subset of data as in Table 3.

Topical Disagreement performance is shown in Table 5. We calculated precision and recall of assigning a correct degree of disagreement to each pair of speakers who are members of a group. Precision and recall averages are then computed over all meso-topics identified in the test dataset, which consists of three separate 90-minute dialogues involving 7, 5 and 7 speakers, respectively. Our calculation includes the cases where different sets of meso-topics were identified by the system and by the human coder. A strict mapping of levels of disagreement between speakers is hard to compute accurately; however, finding the speakers who disagree the most, or the least, is significantly more robust.

Topical Disagreement	Prec.	Recall
Strict	0.33	0.32
Weighted	0.54	0.54
Highest-Rest	0.89	0.85
High-Low	0.77	0.73

Table 5: Topical Disagreement LU assignment performance averages over 13 meso-topics discussed in three dialogues with combined duration of 4.5 hours with total of 19 participants (7, 5, and 7 per session).

6. Conclusion

In this paper we presented a preliminary design for modeling certain types of social phenomena in multi-party on-line dialogues. Initial, limited-scale evaluation indicates that the model can be effectively automated. Much work lies ahead, including large scale evaluation, testing index stability and resilience to NL component level error. Current performance of the system is based on only preliminary versions of linguistic modules (topic extraction, polarity assignments, etc.) which perform at only 70-80% accuracy, so these need to be improved as well. Research on Urdu and Chinese dialogues is just starting.

Acknowledgements

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the U.S. Army Research Lab. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

References

Agar, Michael. 1994. Language Shock, Understanding the Culture of Conversation. Quill, William Morrow, New York.

Allen, J. M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. www.cs. rochester.edu/research/cisd/resources/damsl/

Anderson, A., et al. 1991. The HCRC Map Task Corpus. Language and Speech 34(4), 351--366.

Austin, J. L. 1962. *How to do Things with Words*. Clarendon Press, Oxford.

Bird, Steven, et al. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Blaylock, Nate. 2002. Managing Communicative Intentions in Dialogue Using a Collaborative Problem-Solving Model. Technical Report 774, University of Rochester, CS Dept.

Broadwell, G. A et al. (2010). Social Phenomena and Language Use. ILS Technical report.

Carberry, Sandra and Lynn Lambert. 1999. A Process Model for Recognizing Communicative Acts and Modeling Negotiation Dialogue. *Computational Linguistics*, 25(1), pp. 1-53.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation Journal 41(2): 181-190

Carlson, Lauri. 1983. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel.

Eric N. Forsyth and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26.

Field, D., et al. 2008. Automatic Induction of Dialogue Structure from the Companions Dialogue Corpus, 4th Int. Workshop on Human-Computer Conversation, Bellagio.

Givon, Talmy. 1983. Topic continuity in discourse: A quantitative cross-language study. Amsterdam: John Benjamins.

Ivanovic, Edward. 2005. Dialogue Act Tagging for Instant Messaging Chat Sessions. In Proceedings of the ACL Student Research Workshop. 79–84. Ann Arbor, Michigan.

Ji, Gang Jeff Bilmes. 2006. Backoff Model Training using Partially Observed Data: Application to Dialog Act Tagging. HLT-NAACL

Jurafsky, Dan, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. http://stripe.colorado.edu/~jurafsky/ manual.august1.html

Jurafsky, D., et al. 1997. Automatic detection of discourse structure for speech recognition and understanding. IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.

Khan, Faisal M., et al. 2002. Mining Chat-room Conversations for Social and Semantic Interactions. Computer Science and Engineering, Lehigh University.

Kim, Jihie., et al. 2007. An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions. AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants

Levin, L., et al. (1998). A discourse coding scheme for conversational Spanish. International Conference on Speech and Language Processing.

Levin, L., et al. (2003). Domain specific speech acts for spoken language translation. 4th SIGdial Workshop on Discourse and Dialogue.

Linell, Per. 1990. The power of dialogue dynamics. In Ivana Markov´a and Klaus Foppa, editors, *The Dynamics of Dialogue*. Harvester, 147–177.

Poesio, Massimo and Andrei Mikheev. 1998. The predictive power of game structure in dialogue act recognition. International Conference on Speech and Language Processing *(ICSLP-98)*.

Price, V., Capella, J. N., & Nir, L. (2002). Does disagreement contribute to more deliberative opinion? Political Communication, 19, 95-112.

Sacks, H. and Schegloff, E., Jefferson, G. 1974. A simplest systematic for the organization of turn-taking for conversation. In: *Language* 50(4), 696-735.

Samuel, K. et al. 1998. Dialogue Act Tagging with Transformation-Based Learning. 36th Annual Meeting of the ACL.

Scollon, Ron and Suzanne W. Scollon. 2001. *Intercultural Communication, A Discourse Approach*. Blackwell Publishing, Second Edition.

Searle, J. R. 1969. *Speech Acts*. Cambridge University Press, London-New York.

Shaikh, S. et al. 2010. DSARMD Annotation Guidelines, V. 2.5. ILS Technical Report.

Shaikh S. et al. 2010. MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse, Proc. LREC-2010, Malta.

Stolcke, Andreas et al. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3).

Stromer-Galley, J. 2007. Measuring deliberation's content: A coding scheme. Journal of Public Deliberation, 3(1).

Tianhao Wu, et al. 2002. Posting Act Tagging Using Transformation-Based Learning. Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining

Twitchell, Douglas P., Jay F. Nunamaker Jr., and Judee K. Burgoon. 2004. Using Speech Act Profiling for Deception Detection. Intelligence and Security Informatics, LNCS, Vol. 3073

Webb, N., T. Liu, M. Hepple and Y. Wilks. 2008. Cross-Domain Dialogue Act Tagging. 6th International Conference on Language Resources and Evaluation (LREC-2008), Marrakech.