MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse

Samira Shaikh[1], Tomek Strzalkowski[1,2], Sarah Taylor[3], and Nick Webb[1]

[1]ILS Institute, University at Albany, State University of New York
[2]Institute of Computer Science, Polish Academy of Sciences
[3]Advanced Technology Office, Lockheed Martin IS&GS

Abstract

In this paper, we describe our experience with collecting and creating an annotated corpus of multi-party online conversations in a chat-room environment. This effort is part of a larger project to develop computational models of social phenomena such as agenda control, influence, and leadership in on-line interactions. Such models will help capturing the dialogue dynamics that are essential for developing, among others, realistic human-machine dialogue systems, including autonomous virtual chat agents. In this paper we describe data collection method used and the characteristics of the initial dataset of English chat.

1 Introduction

Multi-party online conversation has become a pervasive form of communication within virtual communities. The popularity of social networking sites has made such communication ubiquitous across all age groups. This phenomenon creates a vast amount of conversational data, which may be utilized for studying a wide spectrum of linguistic and social phenomena and could be exploited in support of various NLP tasks. In particular, a great amount of communication within an online community occurs in virtual chat-rooms, where users log in and converse with other users who may be online at that time. Conversations are typically conducted using free form, highly informal text dialect.

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. Public chat-rooms may be excellent sources of data for studies involving on-line language usage (e.g., novel uses of vocabulary, syntax), general conversational etiquette, and related issues. However, for deriving more complex models of conversational behavior, we need the interaction to be reasonably focused on a task and/or social objectives within a group.

In order 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. Participants were selected from among current and past University students and staff based on their general level of experience with chat communication, but otherwise representing fairly diverse demographics and backgrounds. Whenever possible, we interviewed the candidates to make sure they would be comfortable with various roles we envisioned for them, including the nominal conversation lead, as well as with assuming any emergent and opportunistic roles, such as a challenger, a supporter, a disruptor, etc.

The purpose of this collection was two-fold: (1) understanding how certain social behaviors are reflected in language, and (2) building an automated chat agent that could effectively achieve certain (initially limited) social objectives in the chat-room. This required a careful design of the experiments around topics, tasks, and games for the participants to engage in so that appropriate types of behavior, e.g., dis-

agreement, power play, persuasion, etc. may emerge spontaneously.

Obtaining high-quality conversational corpora with such complex characteristics is inherently difficult. The foremost consideration here is to make sure that the conversation appears as natural as possible given that the entire setup is, in fact, artificial and that the subjects are well aware that their interactions are recorded. Moreover, there is a distinct lack of motivation or incentives for the subjects to engage in more effective but risky behaviors.

In order to mitigate these concerns we have devised a multi-tiered collection process in which the subjects start from simple, free-flowing conversations and progress towards more complex and structured interactions. In this paper we report on the first two stages of this process, which were recently completed. The third, large-scale collection effort is currently being designed. Details of the experimental design are discussed in section 3.

The initial two stages of data collection comprised of 14 sessions of English chat dialogue conducted in groups ranging in size from 3 to 8. We have also conducted 7 sessions of chat with participants conversing in Urdu, which constitutes only a part of the first stage. In this paper we discuss English chat data only.

All English dialogue has been annotated at three levels: communication links, dialogue acts, and topic and focus boundaries. Some details of these annotations will be discussed later in this paper, although a full description is impossible within the scope of this article. 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, for example dialogue act tagging.

2 Related Work

Much research has been undertaken to create corpora in support of dialogue research; however, most available collections are spoken language conversations involving two participants. 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.

Previous work on the study of dialogue phenomena concentrated on two person interactions, both task-focused, such as Map Task (Anderson et al., 1991) and open conversation, as in the annotated portion of the Switchboard corpus (Jurafsky et al., 1997), as well as languages other than English, such as Spanish CALLHOME (Levin et al., 1998) or the NESPOLE speech to speech translation corpus of German, French, Italian and English (Levin et al., 2003).

Recently, work has expanded to include multi-person meetings (such as the ICSI-MRDA corpus) and include a wider range of modalities. For example, the AMI corpus stems from a European research project centered on multi-modal meeting room technology. The AMI Meeting Corpus (Carletta, 2007) contains 100 hours of meetings captured using many synchronized recording devices.

All of these resources look at spoken language. There is a parallel interest in the online chat environment, although the development of useful resources has progressed less. The NPS chat corpus (Forsyth and Martell, 2007) is a corpus of around 10,000 postings from age specific chat rooms on the internet, which have been hand anonymized and labeled with part-of-speech tags and dialogue act labels. The NPS corpus is freely distributed as part of the Natural Language Toolkit (Bird et al., 2009).

The StrikeCom corpus (Twitchell et al., 2007) is a corpus 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 more typical that those interested in the study of Internet chat compile their own corpus on an as needed basis, such as the work of Wu et al. (2002) and Khan et al. (2002) on IRC chat rooms, the work of Kim et al. (2007) on student discussion boards or the study of online conversations between two people, a customer and a shopping assistant, used in the dialogue act annotation effort of Ivanovic (2005).

3 Experiment Design

We collected approximately 20 hours of chat dialogue spread out over 14 sessions of 90 minutes each, amounting to a total of 7317 individual utterances. There were, on average, 5 participants present in each session. In this section we describe how the data collection process was accomplished. In the next section we discuss the characteristics of the dataset collected.

3.1 Subjects

Subjects were recruited from within the University community, and consisted of students and alumni, as well as research staff, including junior faculty. We sent out an email recruiting messages on the mailing lists for a few departments including Computer Science, Information Science and Communication, following the guidelines set out by the University IRB protocol regarding human subject experiments. For the purposes of our research, we wanted to have a minimum of 4 participants for each chat session. We started with a pool of 13 respondents to our initial recruitment message. Participant age varied from 22 years to 55 years with average age being 34 years, with 7 males and 6 females. Participants were compensated for their time.

3.2 The Chat-room Setup

We developed a chat server and client for the purpose of this data collection. The communication between the server and the client is over simple HTTP protocol. Both programs are written in Java, where the chat client is a Java applet, with an interface that is similar to popular chat clients, and can be accessed using any web browser. We have since replaced this chat environment with an XMPP based client-server setup, which we will be using for further data collection. Participants were assigned unique nicknames and given secure login access to the chat server, which could be accessed via the web from any remote location.

3.3 Chat Sessions

We conducted a series of 14 chat sessions divided into two phases. Each phase consisted of 7 sessions, or approximately 10 hours of discourse. We posted a chat session schedule and participants would sign up for as many sessions as were convenient to them. For each session we had an average of 5 users online at the same time, including a nominal "leader" who was responsible for keeping the discussion on a particular topic for the duration of the session (90 minutes). During the first phase of collection, we let the participants to volunteer as leaders and to choose any topic they wished to discuss, but beyond that they were free to converse in any way that felt most comfortable. This phase produced some good, lively conversations that helped to establish initial relationships between the participants, and set the ground for more structured discourse in the second phase.

During the second phase, we gave the participants a specific topic to discuss or a task to perform in each session. For example, the topic could be "Should Dick Cheney and others be prosecuted for their role in using torture?" or "What is your opinion of the government bailout of the American auto-industry?" For a specific task based dialogue, we had the participants form a search committee and select the best candidate for a job from a list of fictional resumes. As in Phase 1, there was a nominal leader whose job was to keep conversation on the topic, but now this assignment required significantly more skill. In Phase 2 dialogues, we observed a marked increase of social phenomena including disagreements, agenda control, and varying degrees of involvement among the participants. In some sessions, alliances formed and discussion leaders emerged quite separate from the nominal chat leads. We are currently analyzing this data towards a formal assessment on how frequently and in which manner these social phenomena occur.

One specific phenomenon we wanted to model was an effective change of conversation topic, when a participant or a group of participants deliberately (if perhaps only temporarily) shift the discussion to a different, possibly related topic. Both success and failure of this action were of interest because the outcome depended upon the choice of utterance, the persons to whom it was addressed, their reaction, and the time when it was produced. In a few dialogues, we gave selected participants "hidden" roles. One role, which we may call a "disruptor" was to opportunistically introduce a secondary topic, somewhat related to the discussion topic, but not directly. Another possible hidden role was that of a consensus breaker where the purpose was to split the group into camps.

We gave the participants who were selected for the particular roles such as leader, disruptor, and consensus breaker only a general outline of what these roles should accomplish. The participants were free to play out these roles in any manner they wished, and only when a suitable opportunity presented itself.

4 Data Statistics

The basic statistical information about the collected data set, which we shall refer to as Multi-Party Chat (MPC) corpus, is given in Table 1 and Table 2. As already indicated earlier, the current

data set represents only a fraction of a larger corpus currently under development.

Total turns in chat corpus	Total words in corpus	Average Words/Turn	Total Emoticons/Misspellings/ Abbreviations
7317	58175	8	241/683/1362

Table 1. Turn total statistics from 14 sessions

Avg. Participants per session	Average Turns per session	Average Turns Per User	Maximum/Minimum Turns per session
5	520	100	165/47

Table 2. Turn average statistics from 14 sessions

In Table 1, we use emoticons to mean a sequence of characters commonly used to signify emotions in chat, such as a smiley face. Misspellings are different from chat-speak, and can be a result of typing errors and non-standard abbreviations.

While we are analyzing the data in detail for the kinds of social phenomena reflected in language use, we collected various statistics on the linguistic, syntactic and grammatical properties of the utterances in our corpus. In particular, we were interested in the rates of use of emoticons, chat-speak (words that are part of chat room jargon such as 'imho' or 'lol'), punctuation, as well as presence of misspellings, ratio of content words to stop words, and number of words per utterance. Figure 1 shows some of these features for each participant averaged across all sessions. Figure 1 shows an interesting trend of how the use of emoticons may be related to the occurrence of misspellings for a user. This trend holds true for all participants except participant P6.

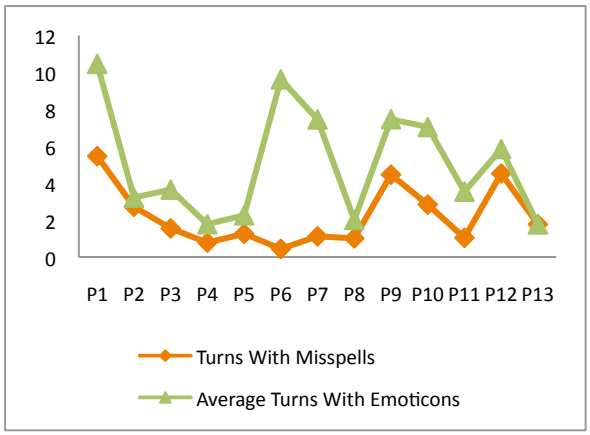

Figure 1. Turns with misspelling versus turns with emoticons

We also compared some simple characteristics of nominal conversation leaders against those of other participants in the discussion. For example, we measured the leader verbosity as the number of turns multiplied by the amount of words in a turn. The chart in Figure 2 indicates that there may be a correlation between verbosity and a leading role in a discussion. We use the term verbosity to be a measure of turn length times the number of turns for that user.

While these superficial statistics are comparatively easy to compute, we are interested in assessing their correlation with more advanced language use factors, such dialogue acts, communicative links and topic and focus changes that are known to be predictive of the types of social phenomena we wish to detect. In the next section we briefly outline the corpus annotation process applied to the MPC corpus.

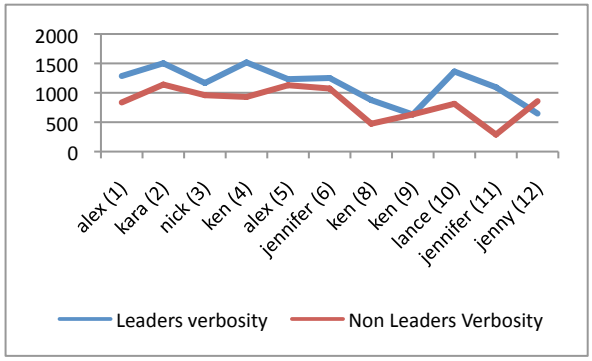

Figure 2. Leader versus non-leader verbosity

5 Annotations

We wish to annotate the data we collected to derive models from language use for social actions such as disagreement, influence, agenda control, and eventually for social roles such as leadership. All of the above represent complex pragmatic concepts that are difficult to annotate directly, let alone detect automatically. Our approach is thus to build a multi-level annotation scheme, where each lower (component) level annotation supplies evidence that supports a claim that some higher-level phenomenon is present.

In this paper we briefly outline only basic component level annotation that consists of three interleaved layers: communicative links, dialogue acts, and topic/focus changes. A more detailed description of the annotation scheme will be presented in a future publication.

5.1 Communicative Links

One of the challenges in multi-party dialogue is to establish which user an utterance is directed towards. Users do not typically add addressing

information in their utterances, which leads to ambiguity while creating a communication link between users. With this annotation level, we asked the annotators to determine whether each utterance was addressed to some user, in which case they were asked to mark which specific user it was addressed to; was in response to another prior utterance by a different user which required marking the specific utterance responded to; or a continuation of the user's own prior utterance.

CL annotation allows for accurate mapping of dialogue dynamics in the multiparty setting, and is a critical component of tracking such social phenomena as disagreement and speaker power.

5.2 Dialogue Acts

We developed a hierarchy of 20 dialogue acts for annotating the functional aspect of the utterance in discussion. The tagset we adopted is loosely based on DAMSL (Allen & Core, 1997) and SWBD (Jurafsky et al., 1997), but greatly reduced and also tuned significantly towards dialogue pragmatics and away from more surface characteristics of utterances. In particular, we ask our annotators what is the pragmatic function of each utterance within the dialogue, a decision that often depends upon how earlier utterances were classified. Thus augmented, DA tags become an important source of evidence for detecting such social phenomena as disagreement between dialogue participants, as well as leadership.

Using the augmented DA tagset also presents a fairly challenging task to our annotators, who need to be trained for many hours before an acceptable rate of inter-annotator agreement is achieved. For this reason, we consider our current DA tagging as a work in progress.

5.3 Topic and Focus boundaries

The flow of discussion in chat shifts quite rapidly from one topic to another. Furthermore, within each topic (e.g., *music bands*) the focus of conversation (e.g., *dc for cutie*) moves just as rapidly. We distinguish between topic and focus to accommodate both broader thematic shifts and more narrow aspect changes of the topic being discussed. For example, participants might discuss the topic of healthcare reform, by focusing on President Obama, and then switch the focus to some particulars of the reform, such as the "public option". Similarly, topics may shift while the focus remains the same (e.g., moving on to Obama's economic policies), although such changes are less common.

We gave our annotators a fair amount of leverage on how to label the topics and how to recognize the focus. Our primary interest was in an accurate detection of topic/focus boundaries and shifts, since these are critical evidential components of the agenda control and leadership social phenomena.

Table 3 summarizes the statistics of the annotated MPC corpus. The first column is the number of sessions we have annotated so far, and the total number of utterances in those sessions. The most frequent communicative link assigned by the annotators was a 'response-to' CL, as listed in the second column (256 times per session). The third column shows that of the 20 dialogue act tags we have developed for this corpus, the most frequently assigned tag was the 'Assertion-Opinion' tag (222 average frequency). The fourth column in Table 3 shows the average number of topic and focus shifts identified in a session by the annotators. Table 4 shows DA tag distribution; besides Assertion-Opinion, other frequent tags include 'Acknowledge', 'Agree-Accept', 'Information-Request' that were assigned to utterances on average 50 times per session.

Total sessions annotated/total utterances annotated	Most frequent CL per session	Most frequent dialogue act per session	Topic/focus changes per session
8/4640	Response-to (256)	Assertion-opinion (222)	14/43

Table 3. Annotation statistics for MPC corpus.

Dialogue Act	Total
Assertion-Opinion	5346
Acknowledge	1315
Information-Request	984
Agree-Accept	966
Positive-Answer	944
Explanation	741
Confirmation-Request	593
Communication-Management	508

Table 4. Total frequency of dialogue acts annotated in the MPC corpus.

6 Building a Virtual Chat Agent

One of our research objectives is to construct an autonomous virtual chat agent (VCA) that could achieve initially limited social goals in a chat room with human participants. We are currently testing the first prototype with the capability to opportunistically change to topic of conversation using a combination of linguistic, dialogic, and topic reference devices, which we observed ef-

fectively deployed by the most influential chat participants in the MPC corpus. A detailed description of the VCA will be the subject of a separate publication once this work is completed.

7 Discussion

In this paper we described the first two phases of building a chat corpus for specific research goals. This is a work in progress. While we continue to refine the experimental design for data collection we are encouraged by the properties of the emerging corpus. Our intention is to make this corpus available to the research community once the collection and annotation process is complete.

Acknowledgments
Development of MPC corpus had been supported in part by grants from the Intelligence Advanced Research Projects Activity (IARPA), the National Institute of Justice, and Lockheed Martin Corporation.

References

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

Anderson, A., M. Bader, E. Bard, E.Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson and R. Weinert. 1991, "The HCRC Map Task Corpus", *Language and Speech* 34(4), 351--366.

Bird, Steven, Klien, Ewan and Loper, Edward. 2009, "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit." O'Reilly Media, 2009

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

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

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

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., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1997. "Automatic detection of discourse structure for speech recognition and understanding." In Proceedings of IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.

Khan, Faisal M., Todd A. Fisher, Lori Shuler, Tianhao Wu and William M. Pottenger, 2002, "Mining Chat-room Conversations for Social and Semantic Interactions." Computer Science and Engineering, Lehigh University.

Kim, Jihie., Erin Shaw, Grace Chern, and Donghui Feng. 2007, "An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions" In Proceedings of the AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants

Levin, Lori, Ann Thyme-Gobbel, Alon Lavie, Klaus Ries, and Klaus Zechner. 1998. "A discourse coding scheme for conversational Spanish." In Proceedings of the International Conference on Speech and Language Processing.

Levin, L., C. Langley, A. Lavie, D. Gates, and D. Wallace. 2003. "Domain specific speech acts for spoken language translation." In Proceedings of 4th SIGdial Workshop on Discourse and Dialogue.

Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler and William M. Pottenger. 2002. "Posting Act Tagging Using Transformation-Based Learning," In the Proceedings of the Workshop on 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", Proceedings of Intelligence and Security Informatics, *Lecture Notes in Computer Science,* Vol. 3073