Extending the MPC corpus to Chinese and Urdu - A Multiparty Multi-Lingual Chat Corpus for Modeling Social Phenomena in Language

Ting Liu[1], Samira Shaikh[1], Tomek Strzalkowski[1, 2], Aaron Broadwell[1], Jennifer Stromer-Galley[1], Sarah Taylor[3], Umit Boz[1], Xiaoai Ren[1], Jingsi Wu[1]

1 University at Albany, State University of New York

2 Institute of Computer Science, Polish Academy of Sciences

3 Sarah M. Taylor Consulting, LLC

E-mail: tliu@albany.edu, tomek@albany.edu

Abstract

In this paper, we report our efforts in building a multi-lingual multi-party online chat corpus (MMPC) in order to develop a firm understanding in a set of social constructs such as agenda control, influence, and leadership as well as to computationally model such constructs in online interactions. These automated models will help capture the dialogue dynamics that are essential for developing, among others, realistic human-machine dialogue systems, including autonomous virtual chat agents. In this paper, we first introduce our experiment design and data collection method in Chinese and Urdu, and then report on the current stage of our data collection. We annotated the collected corpus on four levels: communication links, dialogue acts, local topics, and meso-topics. Results from the analyses of annotated data on different languages indicate some interesting phenomena, which are reported in this paper.

Keywords: Multi-lingual Multi-party online-chat, annotation, social phenomena, post-session survey

1. Introduction

In this paper, we describe our work on building a multilingual corpus for multiparty discourse. We aim to model complex social phenomena, such as leadership, influence, and group cohesion in multiparty discourse based on a set of language uses by discourse participants and to apply our models in multiple languages. Our previous work was focused on English corpus. The current work extends the collection of multiparty conversation corpus to include Chinese and Urdu data.

Multi-party online conversations are particularly interesting to examine not only because they are a relatively common means of communication through the Internet, but also because the reduced cue environment implies that the only ways for group dynamics to unfold is through discourse. However, its adaptation for research purposes presents a number of challenges in that most data 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.

Therefore, we designed a series of experiments to expand our online chat corpus (Shaikh et al., 2010b) by collecting Chinese and Urdu data. We had three major goals through the collection,

1. To build a corpus that reflects the social behavior in the Chinese community and Urdu community.
2. To build a corpus that conveys the cultural differences between the different language communities.
3. To develop new models for DSARMD system (Browdwell et al., 2012) that are applicable to multiple languages.

The corpus collection design involved carefully setting up the experiments, recruiting subjects, selecting proper topics, and designing post-session questionnaire. We also had a system (DSARMD) training phase in which linguistic experts conducted a set of annotations and analyses to further validate our sociolinguistic measures.

This article is structured as follows: Section 2 presents the related research in the literature. Section 3 gives the design of this experiment. Section 4 compares the three languages. Section 5 discusses the annotation and Section 6 describes the post-session survey.

2. Related Work

The current research on natural language processing for Chinese still focuses on POS tagging, parsing (Xue et al. 2005), translation (Song, et al. 2010), etc. So the data collections created by researchers are not suitable for our interests in the study of social phenomena. Especially, no effort has been done on collection of Chinese multiparty online-chat corpus. In the Urdu language as well, the focus has thus far been on developing natural language tools such as parsers (W. Ali, 2010) and part-of-speech taggers (Muaz et al., 2009). Thus, there was a need to collect a corpus of online multi-party conversation in

these languages, to build models and to be able to compare them across cultures and languages.

Also, the existence of useful resources is limited. CallHome and CallFriends[1] are corpora that consists unscripted telephone conversations. The conversations are produced by overseas Chinese students calling home or friends in China. Another corpus is Chinese Broadcast Conversation Parallel Text[2]. In this corpus, eight different conversation programs (total 20.4hrs) were selected from CCTV and Phoenix TV[3].

One weakness of these existing resources is that the number of the participants in conversations is too small, either 2 or 3 in each session. So it may not be able to convey the social behavior that we are interested in, such as Leadership, Influencer, group cohesion etc. Another weakness is the limited amount of information available about the dialogue participants. Such information may be captured through questionnaires or interviews following each data collection experiment designed to reflect the aims of the study. The resulting participant data, which in our case must include participants' assessment of their behavior and roles in conversation, is critical for model validation that would be difficult to obtain otherwise.

Furthermore, in (Broadwell et al., 2012), we demonstrated the great success in modeling the social construct in English online chat corpus built by us (Shaikh et al. 2010b). This gave us strong motivation on data collection for Chinese and Urdu.

3. Design of Experiment

3.1 Subjects

The subjects were recruited from within the University community, Chinese community and Indian and Pakistani community, because we needed the participants to be native language speakers. For the purposes of our research, we wanted to have a minimum of 4 and maximum of 12 participants for each chat session. For Chinese data collection, we recruited 45 native Chinese speakers. Since most of them are international graduate students, the age range is from 20 to 29. We interviewed these candidates to make sure that they're fully understand this experiment and also comfortable if we assign them specific role, such as leader, challenger, or supporter, during the discussion.

3.2 Chat Sessions

We set two phases for our collection. The first phase involved two sessions, which are training sessions. In this phase, we provide hot topics in China, such as "360 v. Tencent" and same for Urdu speaking participants – "Politics of Pakistan under Prime Minister Zardari. However, participants can also discuss whatever they are interested. Through the training sessions, participants will be familiar with the environment of our chat room and get to know each other and built initial relations.

Starting in the second phase, we provide each group either a specific topic to discuss or a task to perform for each session and the participants have to focus on the task until they finished it. For example, we would have participants form a search committee and select the best candidate for a job from a list of fictional resumes. For each session, we assign one participant as a leader who will make sure the discussion is on track and a conclusion will be given at the of the discussion.

For our research purpose, we set the discussion in two ways. One is that we ask multiple groups to work on same topic to see how different groups achieve their conclusion and whether they will come out same conclusion. Another is to choose the identical topics for all three languages' discussion. Therefor, we could compare the social behaviors across different languages and cultures. So far, we have collected 19.5 hours of Chinese chat dialogue spread out over 13 sessions of 90 minutes each and 20.5 hours Urdu chat data in 14 sessions of 90 minutes each.

After each session, participants were instructed to answer a survey aimed at eliciting responses regarding the interaction they had freshly completed. Survey questions were carefully designed by social science standards, requesting participants' reactions without being overtly suggestive. Participants rated each other, as well as themselves, for these questions on an unnumbered 10-point scale.

4. Data analysis

Language	Avg. Participants per session	Avg. Turns per session	Avg. Turns Per User	Avg. Word per Turn
English	5	520	104	8
Chinese	10	1189	119	10
Urdu	4	520	130	8

Table 1: Statistics from 14 sessions of English corpus and 13 sessions of Chinese corpus, and Urdu corpus

[1] http://www.ldc.upenn.edu/

[2] http://www.ldc.upenn.edu/

[3] CCTV is a broadcaster from Mainland China and Phoenix TV is a Hong Kong -based satellite TV station

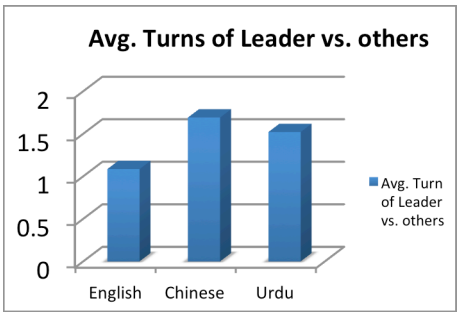

Figure 1: Average number of turns generated by Leaders against average number of turns generated by other group members

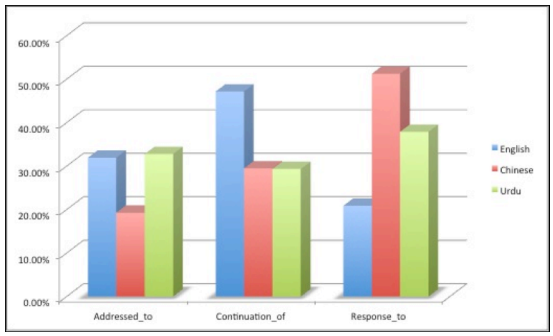

Figure 2: The frequency of Communication Links in annotated MMPC corpus

Since we selected the identical topics for English, and Chinese and Urdu discussions, we can compare their activities. Table 1 shows that for the number of average turns per participant in Chinese corpus is 14% more than those in English corpus. It does not indicate that the participants in Chinese corpus are more active and involved in the discussion, because in each Chinese discussion session, the average number of participants is doubled comparing with the average number of participants in English discussion session. Thus, the participants in Chinese discussion sessions have more opportunities to chat with different people and therefore produce more turns. However, the size of Urdu discussion group is similar as the size of English group, but each participant contributes 25% more utterances, which suggest that Urdu participants are more active during the discussion.

Since each session had a leader assigned, we'd like to see whether the leader's behavior is different from other group participants. One of the measures is to compare the number of turns that participants contributed to the discussion. Figure 2 shows the average number of turns generated by Leader against the average number of turns generated by other participants. In English corpus, Leader only had 8% more turns comparing with the averages. However, the assigned leader in Chinese corpus produced 69% more turns comparing with other participants and Same as the leader in Urdu data, who contributed 52% more turns. That could be an indicator that the assigned leaders in Chinese groups and Urdu groups were more responsible and put more efforts in managing the discussion.

5. Annotation

We have also annotated a significant portion of the data we collected in order to train our models for language use related to disagreement, sociability, agenda control, and eventually for social roles and phenomena, such as leadership, influence and group cohesion. In this paper we briefly outline only basic component level annotation that consists of four interleaved layers: communicative links, dialogue acts, local topic tracking, and meso-topic valences. A more detailed description of the annotation scheme is available in (Shaikh et al., 2010a).

5.1 Communicative links

In a multi-party dialogue an utterance may be directed towards a specific participant, a subgroup of participants or to everyone, we call this category addressed-to; an utterance may be a continuation of a prior utterance by the same participant (continuation-of); or it may respond to a prior utterance by a different speaker (response-to). Figure 2 shows the percentage of 3 types of communication-links in annotated MMPC corpus. When looking at English corpus, we could see it has significantly higher frequency of continuation-of utterances, which imply the participants in English group intended to convey more information in one expression and had to split into 2 or more turns. Another observation from English corpus is that the percentage of response-to utterances is dramatically lower comparing with the other two sets. It may suggest that the participants in English groups were rather to express their own opinion instead of following other people's discussion.

However, Chinese corpus shows the opposite way. It has significantly higher frequency of response-to utterances and lower frequency of address-to utterances. Apparently, the participants in Chinese group are more cooperative and willing to follow other people's discussion. On the other hand, the frequency of communication link types in Urdu corpus is in the middle of English corpus and Chinese corpus.

5.2 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 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

Corpus Type	Assertion-Opinion	Acknowledge	Information-Request	Agree-Accept	Action-Directive	Disagree-Reject
English	33.47%	5.78%	9.36%	16.5%	2.27%	8.23%
Chinese	43.96%	4.53%	8.75%	12.03%	**4.93%**	8.04%
Urdu	33.6%	2.56%	7.86%	18.71%	2.32%	7.74%

Table 2. The frequency of some dialogue act tags in annotated MMPC corpus

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 language uses and such social phenomena as leadership. Examples of dialogue act tags include Assertion-Opinion, Action-Directive, Acknowledge, Information-Request, and Confirmation-Request.

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.

Table 2 provides the frequency of part dialogue act tags in annotated MPC corpus. The Chinese corpus has the highest frequency of Assertion-opinion and Action-Directive. Especially the frequency of Action-Directive is twice as the frequency of the others. It demonstrated the management effort made by the leaders in Chinese groups during the discussion.

5.3 Local Topics

Local topics are defined as nouns or noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun. Any content-bearing noun or noun phrase can be used to introduce a new local topic, and there may be one of more local topics introduced in each dialogue turn. Tracking local topics and their subsequent mentions is constructive in detecting such social language uses as Topic Control and Involvement

(19:0:2 PM) guihua: ***个人意见***, 我不喜欢***苹果***
(19:0:2 PM) guihua: ***Personal opinion***, I don't like ***Apple***
(19:0:12 PM) weifang: ***苹果***有啥不好呀
(19:0:12 PM) weifang: What's wrong with ***Apple***
(19:0:13 PM) chunzhuo: ***苹果***不错
(19:0:13 PM) chunzhuo: ***Apple*** is not bad
(19:0:14 PM) chuqing: ***苹果***也不喜欢你
(19:0:14 PM) chuqing: ***Apple*** doesn't like you too

Figure 3: A piece of on-line chat from MMPC training session

Figure 3 displays an example (with the translation) that from a training session in Chines corpus. In this example, the first utterance mentioned two first appeared noun phrases, ***个人意见***(***personal opinion***) and ***苹果***(***Apple***). Both of them are New Local Topics. Then in the following utterances, ***苹果***(***Apple***) appears repeatedly and all the appearances are the sub-sequent mentions of the first ***苹果***(***Apple***).

During the annotation, we are excluding 1st and 2nd person pronouns, such as 我 (I), 我的 (my), 我们(we), and 你 (you) and names of the participants in the dialogue from this coding. So if the participants in the chat are named Guihua, Weifang, and Chunzhuo, we are not marking them as local topics.

5.4 Meso-Topics and their Valences

Some local 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. A selection of meso-topics is closely associated with the task in which the discourse participants are engaged. For example, when the task is to select a candidate for a job, the name of each applicant becomes a meso-topic. Meso-topics can be distinguished from the local topics because the speakers often make polarized statements about them. An utterance is polarized if it expresses sentiment or valence that a speaker assigns to the meso-topic. Valence can be positive or negative, or in absence of any obvious polarity, it may be neutral. A positive polarity tag is used when an utterance is expressly in favor of the meso-topic, or if it supplies favorable or supporting information about it. A negative polarity tag is used when an utterance is expressly against the meso-topic, or if it supplies unfavorable or negative information about it. If an utterance is neither positive nor negative the neutral polarity tag is used.

6. Post Discussion Survey

After each chat session, participants were required to fill a post-session questionnaire, which took 5 to 10 minutes. The questionnaire contain 11 questions, which split into two parts,

6.1 Evaluations on their own participation

In post-session survey, we have 5 questions for participants to give estimations of performance on their own participation. From these questions, we hope to collect the information that how comfortable they were during the discussion and how much they achieved.

Figure 4: One example question on participant's own evaluation

Figure 4 shows an example of such questions. The participants can pick a score from 5 scale points based on his/her experience during the discussion. The higher the score, the more positive that this participant had a better communication.

6.2 Evaluations on other participants

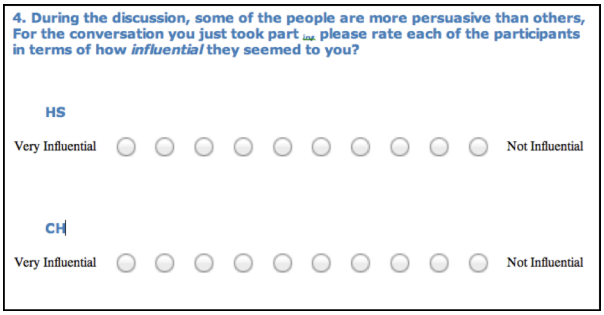

Figure 5: One example question for cross evaluation

In this part, we designed 6 questions to ask participants to evaluate other participants' performance to elicit responses for socio-linguistic behavior. Figure 5 shows one example of our questions. From this question, we hope to get which participant is the influential person during the discussion. We asked participants to give scores (from 1 to 10) to all participants including themselves. The lower score that one participant get, the higher influence the participant is.

Figure 6 displays the average influential score of each participant after cross evaluation. It is clear to show that

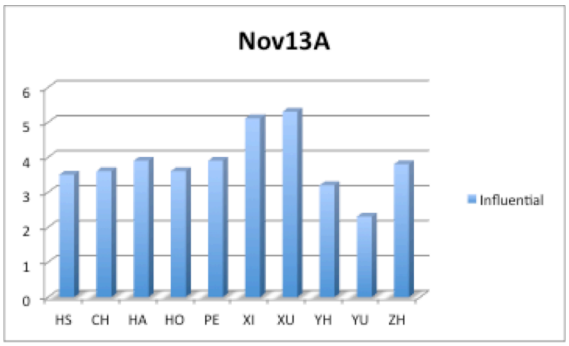

Figure 6: The Average scores of participants on Influential cross evaluation

YU is the most influential person out of the ten participants. Therefore, we could have an extra set of human assessment, which helps to evaluate our results from automatic system.

7. Conclusion

In this paper, we describe a multi-lingual multi-party chat corpus in English (earlier publication), Chinese and Urdu, which is an extension of our previous corpus collected in English called the MPC corpus. We describe the corpus characteristics and the collection method and give brief details about the annotated portion of the corpus. The aim of the MMPC corpus is to build a resource that would allow for automated modeling of a set of socio-linguistic behavior in multi-party discourse. We have developed an annotation scheme specifically geared towards annotating linguistic and syntactic cues that would aid in building automatic models and have put in place a participant survey instrument to elicit responses from the participants in order to obtain ground truth about these models.

8. Acknowledgement

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

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

G. A. Broadwell, J. Stromer-Galley, Strzalkowski, T., S. Shaikh, S. Taylor, T. Liu, U. Boz, A. Ella, L. Jiao, N. Webb. 2012. Modeling Sociocultural phenomena in discourse. Natural Language Engineering: page 1 of 45. Cambridge University Press 2012

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

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 Proc. of IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.

Muaz, A., Ali, A., Hussain, S. (2009). Analysis and Development of Urdu POS Tagged Corpora. In the

Proceedings of the 7th Workshop on Asian Language Resources, IJCNLP'09, Suntec City, Singapore.

Shaikh, S., Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Webb, N., Boz, U., and Elia, A. 2010a. DSARMD annotation guidelines version 2.5. Technical Report 014, ILS, SUNY, Albany, New York.

Shaikh, S., Strzalkowski, T., Taylor, S., and Webb, N. 2010b. MPC: a multi-partychat corpus for modeling social phenomena in discourse. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.

Song Z., S. Strassel, G. Krug, and K. Maeda. (2010). Enhanced Infrastructure for Creation and Collection of Translation Resources. In the Proceedings of the Seventh conference on International Language Resources and Evaluation. Valletta, Malta

Twitchell, Douglas P., Jay F. Nunamaker Jr., and Judee K. Burgoon. (2004). Using Speech Act Profiling for Deception Detection. In Proceedings of Intelligence and Security Informatics, Lecture Notes in Computer Science, Vol. 3073

Wajid Ali and Sarmad Hussain. (2010). Urdu dependency parser: A data-driven approach. In proceedings of the Conference on Language and Technology (CLT'10), Islamabad, Pakistan.