Detecting Leadership in Online Multi-Party Discourse

Tomek Strzalkowski[1,2], George Aaron Broadwell[1], Jennifer Stromer-Galley[1], Samira Shaikh[1], Ting Liu[1], Sarah Taylor[3]

[1]State University of New York – University at Albany [2]Polish Academy of Sciences [3]Lockheed Martin IS&GS
tomek@albany.edu

Abstract

We present in this paper, the application of a novel approach to computational modeling, understanding and detection of social phenomena in online multi-party discourse. A two-tiered approach was developed to detect a collection of social phenomena deployed by participants, such as topic control, task control, disagreement and involvement. We discuss how the mid-level social phenomena can be reliably detected in discourse and these measures can be used to differentiate participants of online discourse. Our approach works across different types of online chat and we show results on two specific data sets.

Introduction

Social interaction in an increasingly online world provides a rich resource for research. The dynamics of small group interaction have been well studied for spoken and face-to-face conversation. However, for a reduced-cue environment such as online chat in a virtual chat room, these dynamics are obtained distinctly, and require explicit linguistic devices to convey social and cultural nuances. Indeed, the only means of expression is through discourse. In addition, participants customarily use esoteric lingo, a variety of emoticons (e.g. ☺, ☹) and unconventional grammar, which add to the challenge. The need arises for robust ways to reliably detect and model behaviors of discourse participants and group dynamics that rely entirely on language and its features.

We have developed a two-tier empirical approach that uses *observable* linguistic features of textual data. These features can be automatically extracted from dialogue to detect certain social phenomena. Our objective is to develop computational models of how social phenomena are manifested in language through the choice of linguistic, syntactic, semantic and conversational forms by discourse participants. We call these higher-level social phenomena such as leadership and group cohesion *social roles or states*, which are achieved by discourse participants through mid-level *social behaviors* such as topic and task control, disagreement and involvement. Given a representative segment of multiparty task-oriented dialogue, our prototype system automatically classifies all discourse participants by the degree to which they deploy selected *social behavior*. These are the mid-level social phenomena, which are deployed by discourse participants in order to achieve or assert higher-level *social roles or states*, including leadership. The high-level social phenomena are then inferred from a combination of social behaviors attributed to each discourse participant; for example, a high degree of topic control and a high degree of involvement by the same person may indicate a leadership role.

In this paper, we discuss the first tier: how to effectively model and classify social language uses in multi-party dialogue and the specific challenges that arise when dealing with them. We also present results of predicting the leader based on these behaviors when compared against the leader picked by participants themselves on post-session surveys. Due to space constraints, two of the behaviors are discussed. Our data comes from two sources – online chat data collected on different topics and chat data collected by participants playing quests in massively multi-player games. Our research so far is focused on the analysis of English language synchronous chat, however we have also begun analysis of Urdu and Mandarin discourse.

Related Research

Most current approaches to dialogue focus on information content and structural components (Blaylock 2002, Stolcke, et al. 2000); few take into account the effects that speech acts may have upon the social roles of discourse participants.

There is a body of literature in anthropology, linguistics, social psychology, and communication on the relationship between language and 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. 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). Other research on the analysis of online chat has focused on topic thread detection and extraction (Adams & Martell 2008, Bengel et al. 2004, Dong et al. 2006).

Data and Annotation

Our initial focus has been on on-line chat dialogues. Chat data is plentiful and readily available on-line, however, its adaptation for research purposes present a number of challenges that include users' privacy issues on the one hand, and their complete anonymity on the other. To derive complex models of conversational behavior such as we are interested in, we need information about the participants and about their interactions with each other. This information may be captured through questionnaires or interviews following each interaction session. The questions must be designed to reflect the aims of the study, which in our case include participants' assessment of their own behavior and roles in conversation as well those of the others. This crucial information is required to validate models and would be difficult, if not impossible, to obtain otherwise. Given the scarcity of available data resources, a new data collection process was required. This is still a fairly typical situation, particularly in the study of Internet chat, that new corpora are created 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 planned a series of experiments in which recruited subjects were invited to participate in a series of on-line chat sessions in specially designed secure environments. The experiments were carefully designed around topics, tasks, quests and games for the participants to engage in so that appropriate types of behavior, e.g., disagreement, power play, persuasion, etc. may emerge spontaneously. In separate data collection experiments, we collected a corpus of 50 hours of 90-minute chat sessions called the MPC corpus (Shaikh et al. 2010a) and a corpus of chat and movement data from 48 quests in Second Life[1] (Small et al. 2011). Figure 1 shows a fragment of conversation from the latter corpus, called the SCRIBE corpus, while Figure 2 shows a fragment from the MPC corpus. Note that given the distinct nature of the tasks and games, the datasets in these two corpora have different characteristics. The SCRIBE corpus has chat focused on successful completion of a quest, and as a result is quite task-focused. In the MPC corpus, while the participants were frequently given tasks, there are digressions and tangential conversations that are generally unfocused on any topic and more unstructured. Our approach works on both types of chat.

An annotation scheme was developed to support the objectives of our project and does not necessarily conform to other similar annotation systems used in the past. A sizeable subset of the English language dataset has been annotated at various levels and we briefly describe three of them below:

Communicative links. Who speaks to whom in a multi-party discourse.

Dialogue Acts. We developed a hierarchy of 15 dialogue acts tuned towards dialogue pragmatics and away from more surface characteristics (Shaikh et al. 2010b). The tagset adopted in this work is based on DAMSL (Allen and Core 1997) and SWBD-DAMSL (Jurafsky et al. 1997).

Local topics. Local topics are defined as nouns or noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun.

Average inter-annotator agreement was 0.78 and above on annotation categories calculated using Krippendorf's (2005) alpha. Annotated datasets were used to develop and train automatic modules that detect and classify social phenomena in discourse. These modules include among others local topic and topic co-reference detection, dialogue act tagging and communicative links classification.

1. **SR**: *who's got the square gear?*
2. **KS**: *i do, but I'm stuck*
3. **SR**: *can you send it to me?*
4. **KS**: *i don't know how*
5. **SR**: *open your inventory, click and drag it to me.*

Figure 1. Fragment of dialogue from SCRIBE corpus.

Topic Control in Discourse

In this section, we describe two mid-level behaviors: Topic Control and Task Control in detail. Details of all language behaviors will be presented in a future, larger publication. Topic Control refers to attempts by any discourse participants to impose the topic of conversation. One hypothesis is that topic control is indicated by the rate of Local Topic Introductions 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. Consider the fragment of conversation in Figure 2.

[1] An online Virtual World developed and launched in 2003, by Linden Lab, San Francisco, CA. http://secondlife.com

1. **JR**: *wanna go thru carlas resume first?*
2. **KN**: *sure*
3. **LE**: *Sure.*
4. **KN**: *i wonder how old carla is*
5. **LE**: *Ha, yeah, when I hear nanny I think someone older.*
6. **KN**: *she's got a perfect driving record and rides **horses**! coincidence?*
7. **JR**: *'06 high school grad*
8. **LE**: *i think she rides a horse and not a car!*

Figure 2. Fragment of conversation from MPC corpus with the local topics highlighted.

This fragment is from a session of 90-minute chat among 7 participants, covering ~700 turns regarding selecting the best candidate for a job interview in the MPC corpus; we will call it Dialogue-1. In this fragment, *carla, resume, nanny* and *horses* are among the local topics introduced. Local topic *carla* introduced in turn 1 by Speaker JR is subsequently mentioned by Speaker KN in turns 4 and 6 and by Speaker LE in turn 8.

Using a Local Topic Introductions 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 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 introduces the highest number of local topics.

Using this information we can assert that Speaker LE exerts the highest degree of Topic Control in Dialogue-1. This index is just one source of evidence; we have developed other indices to complement it. Three of those are:

Subsequent Mentions of Local Topics 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 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 Subsequent Mentions and Cite Score 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.

Turn Length Index. This index stipulates that more influential speakers take longer turns than those who are less influential with topic control. The Turn Length 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.

Figure 3 shows the correlation of the four indices for six participants (LE, DE, KI, KN, DK and JR) for a dialogue. We find that the indices correlate quite well (correlation co-efficient 0.97). 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 determining how they should be weighed to maximize accuracy of making Topic Control claims. Topic Control claims for participants can be made by taking a linear combination of all index scores for that participant. This is indicated by the TCM (Topic Control Measure) entry in Figure 3. We see that Speakers LE scores the highest on the combined Topic Control Measure.

We use Stanford (Klein and Manning 2003) part-of-speech tagger to automatically detect nouns and noun phrases in dialogue and select those of import as local topics. These are nouns or noun phrases that are mentioned subsequently in the discourse. We have developed modules for anaphora resolution; in addition we use Wordnet (Fellbaum et al. 2006) to detect synonyms and other related words.

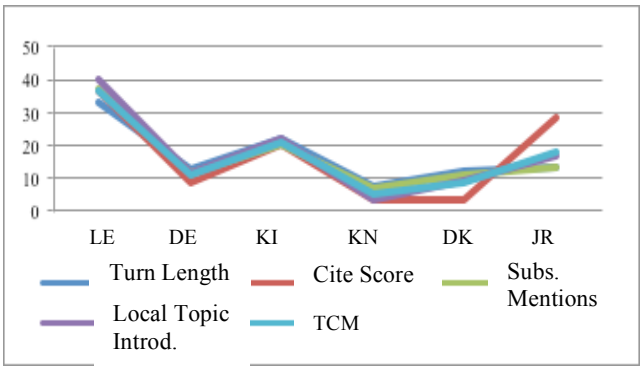

Figure 3: Correlation of Topic Control indices on a sample dialogue.

Task Control in Discourse

Task Control 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. 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.

Task Control is typically realized on the level of dialogue acts, including Action-Directive, Agree-Accept, Disagree-Reject, as well as any Process Management statements. As an example, consider turn 1 by Speaker JR

in Figure 2 above. We define several indices that allow us to compute a degree of Task Control in dialogue for each participant. Two of them are:

Directive Index. 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.

Process Management index. 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 Process (or Task) Management in most dialogue tagging systems.

Let us consider the following information computed for the PMI index over Dialogue-1:

1. Dialogue-1 contains 246 utterances classified as Process Management rather than doing the task.
2. Speaker KI makes 65 of these utterances for a Process-Management Index of 26.4%.
3. Mean Process Management Index for participants is 14.3%.
4. KI has the highest Process Management Index out of all participants.

Using this information, we can assert that Speaker KI exerts the highest degree of Task Control in Dialogue-1.

As with Topic Control, we use a linear combination of index values to arrive at a single measure of Task Control called the Skilled Control Measure (SCM). Task Control mainly relies on dialogue act classification of participant utterance. In our prototype system, dialogue acts are tagged based on presence of certain cue phrases derived from an external training corpus (Webb and Ferguson 2010), tuned against the annotated portion of our corpus. Figure 4 shows the combined SCM measure for 7 participants of SCRIBE quest chat dialogue (Dialogue-2). We see that Speakers SS and BK score higher than the rest of the participants on SCM measure.

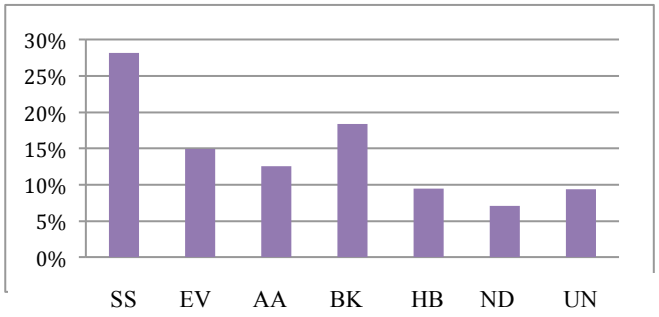

Figure 4: Skilled Control Measure scores for participants of Dialogue-2.

Combining Measures to compute Leadership

Our automated system comprises of a series of modules that create automated annotation of the source dialogue for computation of the various indices defined in previous sections. Automatically annotated dialogue is then used to compute index values from which claims of socio-linguistic behaviors are derived.

Index values of each behavior are combined into a single measure to elicit a ranking for the participants on that behavior. As discussed in section on Topic Control, automatically derived topic control indices are combined to arrive at a measure of Topic Control (TCM) for participants in the discourse. Indices of other behaviors are combined in a similar fashion to produce rankings on Task Control, Disagreement and Involvement. The combined measure for Task Control is called SCM (Skilled Control Measure)

In order to evaluate accuracy of the automated process we compared the claims generated automatically by the system to assessments provided by discourse participants to questions regarding social phenomena. Following each session, participants were instructed to answer a survey aimed at eliciting responses regarding the interaction they had freshly taken part in. Survey questions were carefully designed, directing participants to give their reaction without being overtly suggestive and to acquire their response on the specific behaviors we are interested in. Only some of the questions from the survey are listed. Participants rated each other, as well as themselves on an unnumbered 10-point scale (except Question 3 below).

1. During the discussion, some of the people are more influential than others. For the conversation you just took part in, please rate each of the participants in terms of how influential they seemed to you? (Scale: Very Influential-Not Influential)

2. During the discussion, some of the people have greater effect on the group's decision than others. For the conversation you just took part in, please rate each of the participants in terms of how much they affected the group's decision? (Scale: Very Effective-Not Effective)

3. Below is a list of participants including you. Please rank order the participants with 1 being the leader, 2 being a leader but not so much as 1, and so on.

The accuracy metric is computed by taking the ranking obtained by automated process and comparing it against the ranking obtained from post-session survey responses. For example, compare a system generated ranking ({A, B}, C, D) with a participant generated ranking (A, {B, C}, D). The set of preference relations induced by this ordering is {A>B, A>C, A>D, B=C, B>D, C>D}. The preference set produced by the system is {A=B, A>C, A>D, B>C, B>D, C>D}. The accuracy of the system ordering is 4/6 or .66. Using this metric, we computed accuracy of Topic Control

(TCM) and Task Control (SCM) for a subset of 10 datasets in our corpus; shown in Figure 5. Average accuracy of both measures is ~70%. Here F19B, F26A, F27B and so on indicate names of individual chat sessions in the corpus. We can similarly measure performance of our other behaviors.

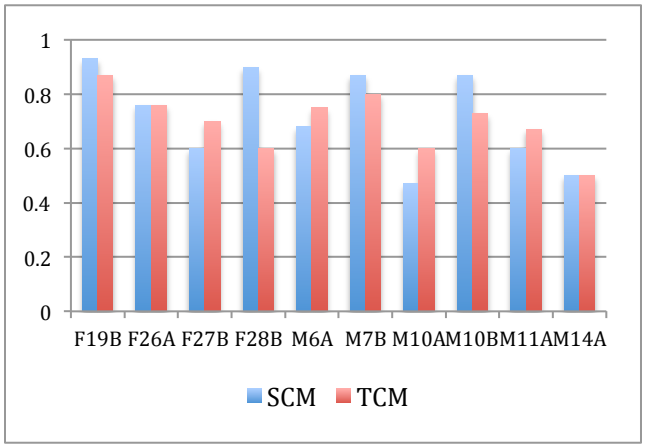

Figure 5: Performance of combined Topic Control (TCM) and combined Task Control (SCM) measures on a subset of data.

We also included a question on the survey to elicit responses regarding leadership (question 3 above). Participants ranked others as well as themselves on a leadership scale. We take the average score for each participant given on this question to find who the leader was for that session. Participants themselves are the best assessors of what transpires in the session and eliciting their responses freshly after the session provides us with a practical basis to compare against.

For example, consider Dialogue-2, an actual data set from our SCRIBE corpus. Speaker SS received the highest number of votes by all participants to be the leader for this session. Using our combined measures for the mid-level behaviors, we can arrive at an automatic leadership calculation as well. Topic Control, Task Control, Disagreement and Involvement measures are combined for a leadership ranking for each participant. We have found, through empirical data that Topic Control and Task Control have the highest correlation with Leadership in the dialogues we have seen. Consequently, Topic Control and Task Control are used as the primary indicators of leadership and we weigh them the highest. Disagreement and Involvement serve as secondary indicators that provide supplemental evidence of leadership. In Dialogue-2, Speaker SS had the highest SCM score (as evinced in Figure 4) and also the highest TCM score (graph omitted due to space constraints). In addition, Speaker SS scored higher than most on our Disagreement and Involvement measures. Using this evidence, we can rank Speaker SS as well the other participants on a leadership scale, shown in Figure 6. We are currently experimenting with weights and combinations of behaviors to maximize accuracy of leadership ranking.

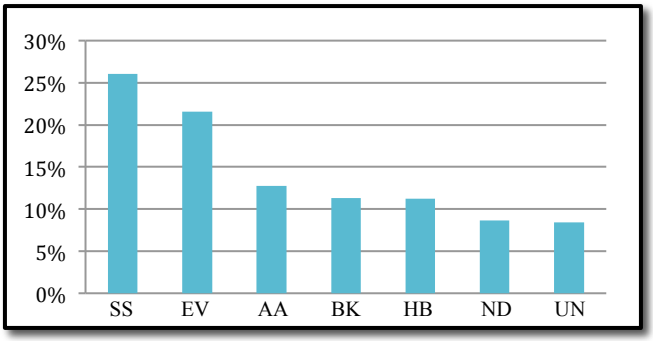

Figure 6: Leadership ranking for participants of Dialogue-2.

Since we are interested in finding who is the leader given a representative set of dialogue, the order and ranking of participants at positions below a certain rank, say rank 3 becomes less relevant. For instance, accurately predicting the participant who ranks fifth or fourth is less valuable than predicting the leader at the first and second positions. For that reason, our accuracy metric takes into account correctly choosing the participant who was also chosen by the participants (by the voting from the post session questionnaire) at either the first or second position. Using this metric, our performance is 74% on average across substantive subsets of both corpora for predicting the leader at first rank, while it is 94% for predicting the leader at either the first or second ranks. Table 1 shows the individual accuracy scores for both the MPC corpus and SCRIBE corpus using both indicators.

	Leader at 1st rank	Leader at 1st or 2nd rank
MPC Corpus	69%	88%
SCRIBE Corpus	80%	100%

Table 1: Accuracy of leadership scores across corpora.

Conclusions and Future Work

We have shown a linguistic approach that relies entirely on language to model behaviors of participants in online chat dialogues. Our approach works on chat dialogues from corpora with distinct characteristics. We use observable linguistic features that are automatically extracted from text to obtain measures of certain mid-level social behaviors, namely Topic Control, Task Control, Disagreement and Involvement. These measures are then combined to predict higher-level social phenomena such as

leadership. Current performance results are very encouraging, both at the socio-linguistic behavior level and the leadership level.

Future research includes work on improving the performance of mid-level behaviors, testing stability of indices and improving the performance of language processing components. We are also interested in exploring additional higher-level social phenomena such as group cohesion and stability; these take into account the distribution of social behaviors across participants, rather than individual rankings.

References

Adams, P.H., & Martell, C.H. 2008. Topic detection and extraction in chat. Proceedings of the IEEE International Conference on Semantic Computing (pp. 581-588). Santa Clara, CA: IEEE Press.

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

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

Bengel, J., Gauch, S., Mittur, E., & Vijayaraghavan, R. 2004. ChatTrack: Chat room topic detection using classification. Intelligence and Security Informatics, 3073, 266-277.

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

Dong, H., Hui, S.C., & He, Y. (2006). Structural analysis of chat messages for topic detection. Online Information Review, 30(5), 496-516.

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

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

Khan, F. M., T. A. Fisher, L. Shuler, T. Wu and W. M. Pottenger 2002. Mining Chat-room Conversations for Social and Semantic Interactions. Computer Science and Engineering, Lehigh University Technical Report, LU-CSE-02-011.

Kim, J., E. Shaw, G. Chern and D. Feng. 2007. An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions. AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Krippendorff, K. 2005. Computing Krippendorff's alpha-reliability. Technical Report. University of Pennsylvania. PA. http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf

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.

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

Shaikh, S., T. Strzalkowski, S. Taylor and N. Webb. 2010a. MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse, in proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta. 2010.

Shaikh, S., T. Strzalkowski, G. A. Broadwell, J. Stromer-Galley, N. Webb, U. Boz and A. Elia. 2010b. DSARMD Annotation Guidelines Version 2.5, ILS Technical Report 014.

Small, S., Stromer-Galley, Jennifer., Strzalkowski, T. 2011. Multi-modal Annotation of Quest Games in Second Life. In Proceedings of ACL 2011, Portland, Oregon.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. Computational Linguistics 26(3), 339–373.

Webb, N. and M. Ferguson. 2010. Automatic Extraction of Cue Phrases for Cross-Corpus Dialogue Act Classification, in the proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), Beijing, China. 2010.

Wu, T., F. M. Khan, T. A. Fisher, L. A. Shuler and W. M. Pottenger. 2002. Posting Act Tagging Using Transformation-Based Learning. In Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining.