VCA: An Experiment With A Multiparty Virtual Chat Agent

Samira Shaikh[1], Tomek Strzalkowski[1, 2], Sarah Taylor[3], Nick Webb[1]
[1]ILS Institute, University at Albany, State University of New York
[2]Institute of Computer Science, Polish Academy of Sciences
[3]Advancded Technology Office, Lockheed Martin IS&GS
E-mail: ss578726@albany.edu, tomek@albany.edu

Abstract

The purpose of this research was to advance the understanding of the behavior of small groups in online chat rooms. The research was conducted using Internet chat data collected through planned exercises with recruited participants. Analysis of the collected data led to construction of preliminary models of social behavior in online discourse. Some of these models, e.g., how to effectively change the topic of conversation, were subsequently implemented into an automated Virtual Chat Agent (VCA) prototype. VCA has been demonstrated to perform effectively and convincingly in Internet conversation in multiparty chat environments.

1 Introduction

Internet chat rooms provide a ready means of communication for people of most age groups these days. More often than not, these virtual chat rooms have multiple participants conversing on a wide variety of topics, using a highly informal and free-form text dialect. An increasing use of virtual chat rooms by a variety of demographics such as small children and impressionable youth leads to the risk of exploitation by deceitful individuals or organizations. Such risks might be reduced by presence of virtual chat agents that could keep conversations from progressing into certain topics by changing the topic of conversation.

Our aim was to study the behavior of small groups of online chat participants and derive models of social phenomena that occur frequently in a virtual chat environment. We used the MPC chat corpus (Shaikh et al., 2010), which is 20 hours of multi-party chat data collected through a series of carefully designed online chat sessions. Chat data collected from public chat rooms, while easily available, presents significant concerns regarding its adaptability for our research use. Publicly available chat data is completely anonymous, has a high level of noise and lack of focus, in addition to engendering user privacy issues for its use in modeling tasks. The MPC corpus was used in (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. A brief description of the MPC corpus and its relevant characteristics is given in Section 3 of this paper.

One specific phenomenon of social behavior 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 these actions was 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. Our analysis of the corpus for such phenomena led to the use of an annotation scheme that allows us to annotate for topic and focus change in conversation. We describe the annotation scheme used in Section 4.

We constructed an autonomous virtual chat agent (VCA) that could achieve initially limited social goals in a chat room with human participants. We used a novel approach of exploiting the topic of conversation underway to search the web and find related topics that could be inserted in the conversation to change its flow. We tested 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 effectively deployed by the most influential chat participants in the MPC corpus. The VCA design, architecture and mode of operation are described in detail in Section 5 of this paper.

2 Related Work

Automated dialogue agents such as the early ELIZA (Weizenbaum, 1966) and PARRY

(Colby, 1974) could conduct a one-on-one "conversation" with a human using rules and pattern-matching algorithms. More recently, the addition of heuristic pattern matching in A.L.I.C.E (Wallace, 2008) led to development of chat bots using AIML[1] and its variations, such as Project CyN[2]. Most of the work on conversational agents was limited to one-on-one situations, where a single agent converses with a human user, whether to perform a transaction (such as booking a flight or banking transactions) (Hardy et al., 2006) or for companionship (e.g., browsing of family photographs) (Wilks, 2010). Many of these systems were inspired by the challenge of the Turing Test or its more limited variants such as Loebner Prize.

Research in the field of developing a multi-user chat-room agent has been limited. This is somewhat surprising because a multi-user setting makes the agent's task of maintaining conversation far less onerous than in one-on-one situations. In a chat-room, with many users engaged in conversations, it is much easier for an agent to pass as just another user. Indeed, a skillfully designed agent may be able to influence an ongoing conversation.

3 MPC Chat Corpus

The MPC chat corpus is a collection of 20 hours of chat sessions with multiple participants (on average 4), conversing for about 90 minutes in a secure online chat room. The topics of conversation vary from free-flowing chat in the initial collection phase to allow participants to build comfortable a rapport with each other, to specific task-oriented dialogues in the latter phase; such as choosing the right candidate for a job interview from a list of given resumes. This corpus is suitable for our research purposes since the chat sessions were designed around enabling the social phenomena we were interested in modeling.

4 Annotation Scheme

We wished to annotate the data we collected to derive models from language use for social phenomena. These represent complex pragmatic concepts that are difficult to annotate directly, let alone detect automatically. Our approach was to build a multi-level annotation scheme.

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

4.1 Communicative Links

Annotators are asked to mark each utterance in one of three categories – utterance is addressed to a participant or a set of participants, it is in response to a specific prior utterance by another participant or it is a continuation of the participant's own prior utterance. By an utterance, we mean the set of words in a single turn by a participant. In multi-party chat, participants do not generally add addressing information in their utterances and it is often ambiguous to whom they are speaking. Communicative link annotation allows us to accurately map who is speaking to whom in the conversation, which is required for tracking social phenomena across participants.

4.2 Dialogue Acts

At this annotation level, we developed a hierarchy of 20 dialogue acts, based loosely on DAMSL (Allen & Core, 1997) and SWBD-DAMSL (Jurafsky et al., 1997), but greatly reduced and more tuned to dialogue pragmatics. For example, the utterance "It is cold here today" may function as a Response-Answer when given in response to a question about the weather, and would act as an Assertion-Opinion if it is evaluated alone. The dialogue acts, thus augmented, become an important feature in modeling participant behavior for our research purpose. A detailed description of the tags is beyond the scope of this paper.

4.3 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. Annotators typically marked the first mention of a substantive noun phrase as a topic or focus introduction.

[1] http://www.alicebot.org/aiml.html

[2] http://www.daxtron.com/123start.htm?Cyn

The effect of topic change is apparent when a subsequent utterance by another participant is about the same topic. This is a successful attempt at changing the topic. Shown in Figure 1 is an example of topic shift annotated in our data collection.

AA 1: did anyone watch the morning talk shows today (MTP, for example)?
KA 2: nope!
AA 3: I missed them – I was hoping someone else had.
AA 4: My kids tell me the band you're going to hear (dc for cutie) is great.
(TOPIC: music bands, FOCUS: dc for cutie)
KA 5: oh cool! Their lyrics are nice, I think.
(TOPIC: music bands, FOCUS: dc for cutie)
KA 6. what kind of music do you guys listen to?
(TOPIC: music, FOCUS: none)
KN 7: I don't really have a favorite genre….you on youtube right now?
(TOPIC: music, FOCUS: youtube)

Figure 1. A topic change in dialogue, with three participants (AA, KA and KN)

We found this model of topic change fairly consistently exhibited, where the participants would ask an open question, in order to get other participants to respond to them, thereby changing the course of conversation. We collected all utterances marked topic shifts and focus shifts and created a set of templates from them. These templates served as a model for the VCA to utilize when creating a response.

Another model of behavior that we found as a consequence of topic change is topic sustain. This is an instance where the utterance is marked to be on the same topic as the one currently being discussed, for example, utterance 5 in Figure 1. These may be in the form of offering support or agreement with a previous utterance or asking a question about a new in-topic aspect.

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. Of the 14 sessions we selected from the MPC corpus, we selected 10 for annotation, with at least 3 annotators for each session. In Table 1 some of the overall statistics computed from this set are shown. We computed inter-annotator agreement on all three levels of our annotation, i.e. Communication Links, Dialogue Acts and Topic/Focus Shifts. Topic and Focus shifts had the highest inter-annotator agreement scores on different measures such as Krippendorf's Alpha (Krippendorff, 1980) and Fliess' Kappa (Fliess, 1971). In Figure 2, we show inter-annotator agreement measures on Topic/Focus shift annotation for four of the annotated sessions. Krippendorff's Alpha and Fleiss' Kappa measures show inter-annotator agreement on topic shift alone, and Conflated Krippendorff's Alpha measures show the agreement when topic and focus are conflated as one category. With such high degree of agreement, we can reliably derive models of topic shift behavior from our annotated data.

Total Number of Sessions Annotated	10
Number of annotators per file	3
Total Utterances Annotated	4640
Average number of utterances per session	~520
Total topics identified per session	174
Total topic shifts identified per session	344

Table 1. Selected statistics from annotated data set

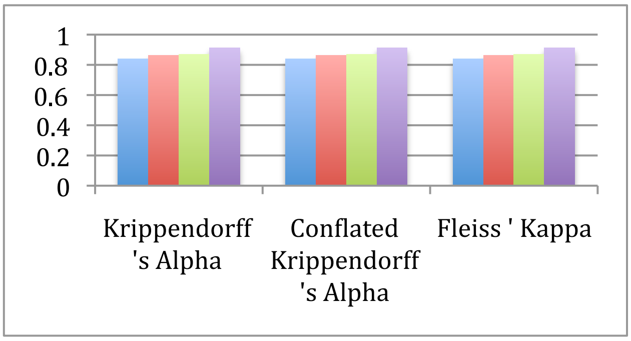


Figure 2. Inter-annotator agreement measures for Topic/Focus shifts

5 VCA Design

A virtual chat agent is an automated program with the ability to respond to utterances in chat. Our VCA is distinctive in its ability to participate in multi-party chat and manage to steer the flow of conversation to a new topic. We exploit the dialogue mechanism underlying HITIQA (Small et al. 2009) to drive the dialogue in VCA.

The topic as defined by the information contained in the participant's utterance is used to mine outside data sources (e.g., a corpus, the web) in order to locate and learn additional information about that topic. The objective is to identify some of the salient concepts that appear associated with the topic, but are not directly mentioned in the utterance. Such associations may be postulated because additional concepts are repeatedly found near the concepts mentioned in the utterance.

An illustrative example found in our annotated corpus is the utterance, "*Lars Ulrich might have a thing or two to say about technology.*" Here, the topic of conversation prior to this utterance was "*technology*" and it was changed to "*music*" after this utterance. Here, "*Lars Ulrich*" is the bridge that connects the two concepts "*technology*" and "*music*" together.

5.1 VCA Architecture

The VCA is composed of the following modules that interact as shown in Figure 3.

5.1.1 Chat Analyzer

Every utterance in chat is first analyzed by the Chat Analyzer component. This process removes stop words, emoticons and punctuation, as well as any participant nicknames from the utterance. We postulate that the remaining content bearing words in the utterance represent the topic of that utterance. We call this analyzed utterance our chat "query" which is sent in parallel to the Document Retrieval and NL Processing component.

5.1.2 Document Retrieval

The document retrieval process retrieves documents from either the web or a test document corpus. We use Google AJAX api for our web retrieval process and InQuery (Callan et al., 1992) retrieval engine for our offline mode of operation to retrieve documents from the test corpus. The test document corpus was collected by mining the web for all utterances in our data collection, creating a stable document set for experimental purposes. Currently, the document corpus contains about 1Gb of text data.

5.1.3 Clustering

We cluster the paragraphs in documents retrieved using clustering method in Hardy et al. (Hardy et al., 2009) This process groups the paragraphs containing salient entities into sets of closely associated concepts. From each cluster, we choose the most representative paragraph, usually called the "seed" paragraph for further NL processing. Each seed paragraph and the chat query undergo the same further NL processing sequence.

5.1.4 Natural Language Processing

We process each chat query by performing stemming, part-of-speech tagging and named-entity recognition on it. Each seed paragraph is also run through same three natural language processing tasks. We are using Stanford POS tagger for our part-of-speech tagging. For named entity recognition, we have the ability to choose between BBN's IdentiFinder and AeroText™ (Taylor, 2004).

5.1.5 Framing

We build frames from the entities and attributes found in both the chat query and the paragraphs.. This work extends the concept of framing developed for HITIQA (Small et al, 2009) and COLLANE (Strzalkowski, 2009). Framing provides an informative handle on text, which can be exploited to compare the underlying textual representations, as we explain in the next section.

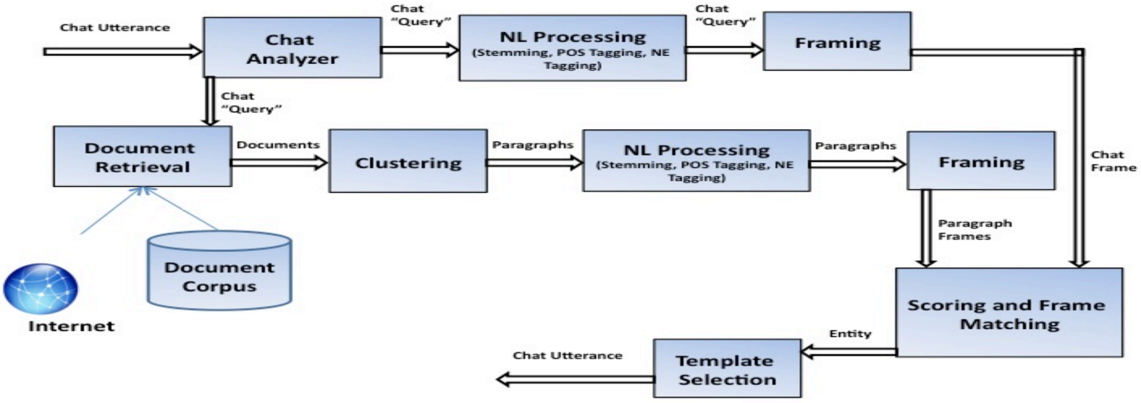


Figure 3. VCA Architecture

5.1.6 Scoring and Frame Matching

Using the information in the frames built in the previous step; we compare the chat query frame built from the chat query, to the frames created from the paragraphs, called paragraph frames. We assign a score for each paragraph frame based on how many attributes and their corresponding values match; in the current version of VCA a very basic approach to counting how many attribute-value pairs match is taken. Of all the paragraph frames we select the highest scoring frames and select the attribute-value pairs that are not part of the chat query frame. For example, as shown in Figure 4a below, the chat utterance “Aruba might be nice!” created the following chat query frame.

[POS]
NNP, Aruba
JJ, nice
[ENT] PLACE

a. Example chat query frame

Aruba Entity List:
VALUE = NASCAR and TYPE = ORGANIZATION and SCORE = 0
VALUE = **Dallas** and TYPE = PLACE and SCORE = 1
VALUE = Mateo and TYPE = PERSON and SCORE = 0

VCA: How about Dallas?

b. Frame Matching, Scoring and Template Selection

Figure 4. From frames to VCA responses

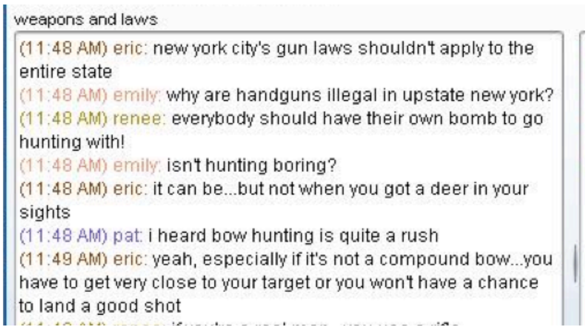
Correspondingly, we select all PLACE type entities from the highest-ranking paragraph frames. These are shown in Figure 4b as Aruba Entity list. The entities “*NASCAR*”, “*Women Seeking Men*” and “*Mateo*” are not of entity type – PLACE, we assign them a score of 0. The score is the frequency of occurrence of that entity in the paragraph; in this example it is found to be 1. Assigning scores by frequency of occurrence ensures that the most commonly occurring concept around the one that is being discussed in the chat query utterance will be used to respond with.

5.1.7 Template Selection

Once we have chosen the entity to respond with, we select a template from the set of templates for that entity. These are templates that are created based on the models created from topic change utterances annotated in our data set. For a select group of entities, which are quite frequently encountered in our data collection such as PLACE, PERSON, ORGANIZATION etc., we have a set of templates specific to that entity type. We also have several generic templates that may be used if the entity type does not match the ones that we have selected. For example, a PLACE specific template is “*Have you ever been to __?*” and a PERSON specific template is “*You heard about __?*”. Not all templates are formulated as questions. Another example of a generic template is “*__rules!*”.

6 Example of VCA Interaction

Figure 5 represents an example of the VCA in action in a simulated environment; the VCA is the participant “renee”. We can see how the conversation changes from “*gun laws*” to “*hunting*” after renee’s utterance at 11:48 AM.



weapons and laws

(11:48 AM) eric: new york city's gun laws shouldn't apply to the entire state
(11:48 AM) emily: why are handguns illegal in upstate new york?
(11:48 AM) renee: everybody should have their own bomb to go hunting with!
(11:48 AM) emily: isn't hunting boring?
(11:48 AM) eric: it can be...but not when you got a deer in your sights
(11:48 AM) pat: i heard bow hunting is quite a rush
(11:49 AM) eric: yeah, especially if it's not a compound bow...you have to get very close to your target or you won't have a chance to land a good shot

Figure 5. Topic change example

7 Evaluation

We ran two tests of this initial VCA prototype in a public chat-room. VCA was inserted into a public chat-room with multiple participants on two separate occasions. The general topic of discussion during both instances was “*anime*”. We have developed an evaluation protocol in order to test the effectiveness of the VCA prototype in a realistic setting. The initial metric of VCA effectiveness is the rate of involvement measured in the number of utterances generated by the VCA during the test period. These utterances are subsequently judged for appropriateness using the metric developed for the Companions Project (Webb, 2010). The actual appropriateness annotation scheme can be quite involved, but for this simple test we reduced the coding to only binary assessment, so that the VCA utterances were annotated as either appropriate or inappropriate, given the content of the utterance and the flow of dialogue thus far. Using this coarse grain evaluation on a live chat segment we noted that the VCA made 9 appropriate utterances and 7 inappropriate utterances, which gives the appropriateness score of 56%. While some of VCA utterances seem inappropriate (i.e., not related to the conversation topic), we noted also that other posters generally tolerated these inappropriate utterances that occurred early in the dialogue. Moreover, these early inappropriate utterances did generate appropriate responses from the human users. This "positive" dynamic changed gradually as the dialogue progressed, when the participants began to ignore VCA's utterances.

While this coarse grained evaluation is useful, our plan is to conduct evaluation experiments by recruiting subjects for chat sessions and inserting the VCA in the discussion. We will measure the impact of the VCA in the chat session by having participants fill out post-session questionnaires, which can elicit their responses regarding (a) if they detect presence of a VCA at any time during the dialogue; (b) who was the VCA; (c) who changed the topic of conversation most often; and so on. Another metric of interest is the level of engagement of the VCA, which can be measured by the number of direct responses to an utterance by the VCA. We are developing the evaluation process, and report on the results in a separate publication.

References

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

Callan, J. P., W. B. Croft, and S. M. Harding. 1992. *The INQUERY Retrieval System*, in Proceedings of the 3rd Inter- national Conference on Database and Expert Systems.

Colby, K.M, Hilf, F.D, and S. Weber. 1972. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. In: *Artificial Intelligence* , Vol. 3, p. 199-221.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 74(5):378{382.

Hardy, Hilda, Nobuyuki Shimizu, Tomek Strzalkowski, Ting Liu, Bowden Wise and Xinyang Zhang. 2002. Cross-document summarization by concept classification. In *Proceedings of ACM SIGIR '02 Conference,* pages 121-128, Tampere, Finland.

Hardy, H., A Biermann, R. Bryce Inouye, A. McKenzie, T. Strzalkowski, C. Ursu, N. Webb and M. Wu. 2006. The AMITIES System: Data-Driven Techniques for Automated Dialogue. In Speech Communication 48 (3-4), pages 354-373. Elsevier.

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

Krippendorff, Klaus. 1980. Content Analysis, an Introduction to its Methodology. Sage Publications, Thousand Oaks, CA.

Samira S., Tomek Strzalkowski, Sarah Taylor and Jonathan Smith (2009) Comparing an Integrated QA system performance - A Preliminary Model. Proceedings of PACLING Conference, Sapporo, Japan.

Shaikh, S., Strzalkowski, T., Broadwell, A., Stromer-Galley, J., Taylor, Sarah and Webb, N. 2010. Proceedings of LREC Conference, Malta.

Sharon Small and Tomek Strzalkowski. 2009. HITIQA: High-Quality Intelligence through Interactive Question Answering. *Journal of Natural Language Engineering*, Vol. 15 (1), pp. 31—54. Cambridge.

Tomek Strzalkowski, Sarah Taylor, Samira Shaikh, Ben-Ami Lipetz, Hilda Hardy, Nick Webb, Tony Cresswell, Min Wu, Yu Zhan, Ting Liu, and Song Chen. 2009. COLLANE: An experiment in computer-mediated tacit collaboration. In Aspects of Natural Language Processing (M. Marciniak and A. Mykowiecka, editors). Springer.

Taylor, Sarah M. 2004. "Information Extraction Tools: Deciphering Human Language." IT Professional. Vol. 06, no. 6, pages: 28-34. November/December, 2004. Online. http://ieeexplore.ieee.org/iel5/6294/30282/01390870.pdf?tp=&arnumber=1390870&isnumber=30282

Wallace, R. 2008. The Anatomy of A.L.I.C.E. In Parsing the Turing Test. (Robert Epstein, Gary Roberts and Grace Beber, editors). Springer.

Webb, N., D. Benyon, P. Hansen and O. Mival. 2010. Evaluating Human-Machine Conversation for Appropriateness. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta.

Weizenbaum, Joseph. January 1966. "ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36–45.

Wilks, Y. 2010. Artificial Companions. In: Y.Wilks (ed.) Close Engagement with Companions: scientific, economic, psychological and philosophical perspectives. John Benjamins: Amsterdam.