Plans Toward Automated Chat Summarization

David C. Uthus
NRC/NRL Postdoctoral Fellow
Washington, DC 20375
david.uthus.ctr@nrl.navy.mil

David W. Aha
Naval Research Laboratory (Code 5514)
Washington, DC 20375
david.aha@nrl.navy.mil

Abstract

We describe the beginning stages of our work on summarizing chat, which is motivated by our observations concerning the information overload of US Navy watchstanders. We describe the challenges of summarizing chat and focus on two chat-specific types of summarizations we are interested in: thread summaries and temporal summaries. We then discuss our plans for addressing these challenges and evaluation issues.

1 Introduction

We are investigating methods to summarize real-time chat room messages to address a problem in the United States military: information overload and the need for automated techniques to analyze chat messages (Budlong et al., 2009). Chat has become a popular mode of communications in the military (Duffy, 2008; Eovito, 2006). On US Navy ships, watchstanders (i.e., personnel who continuously monitor and respond to situation updates during a ship's operation, Stavridis and Girrier (2007)) are responsible for numerous duties including monitoring multiple chat rooms. When a watchstander reports to duty or returns from an interruption, they have to familiarize themselves with the current situation, including what is taking place in the chat rooms. This is difficult with the multiple chat rooms opened simultaneously and new messages continuously arriving. Similarly, Boiney et al. (2008) observed that with US Air Force operators, when they returned to duty from an interruption, another operator in the same room verbally updates them with a summary of what had recently taken place in the chat rooms and where they can find the important information. Both of these situations are motivations for chat summarization, since watchstanders and operators could use automatically generated summaries to quickly orient themselves with the current situation.

While our motivation is from a military perspective, chat summarization is also applicable to other domains. For example, chat is used for communication in multinational companies (Handel and Herbsleb, 2002), open source meetings (Shihab et al., 2009; Zhou and Hovy, 2005), and distance learning (Osman and Herring, 2007). Summarization could aid people who missed meetings or students who wish to study past material in a summarized format.

Even though chat summarization has many potential uses, there has been little research on this topic (Section 3). One possible reason for this is that chat is a difficult medium to analyze: its characteristics make it difficult to apply traditional NLP techniques. It has uncommon features such as frequent use of abbreviations, acronyms, deletion of subject pronouns, use of emoticons, abbreviation of nicknames, and stripping of vowels from words to reduce number of keystrokes (Werry, 1996). Chat is also characterized by conversation threads becoming entangled due to multiple conversations taking place simultaneously in *multiparticipant chat*, i.e., chat composed of three or more users within the same chat room (Herring, 1999; Herring, 2010). The interwoven threads then make it more difficult to comprehend individual conversations.

The rest of this paper describes our challenges

Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pages 1–7,
Portland, Oregon, June 23, 2011.

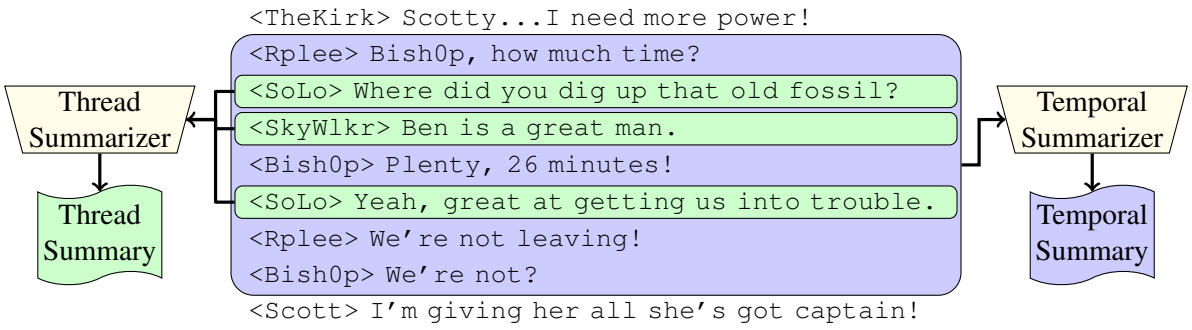

Figure 1: Process for generating thread and temporal summaries from a chat log.

in chat summarization. We define two chat-related types of summarizations we are investigating (Section 2) and describe related work (Section 3). Furthermore, we give an overview of our planned approach to these challenges (Section 4) and also address relevant evaluation issues (Section 5).

2 Our Summarization Challenge

Our research goal is to summarize chat in real-time. Summaries need to be updated with every new chat message that arrives, which can be difficult in high-tempo situations. For these summarizations, we seek an abstract, compact format, allowing watchstanders to quickly situate themselves with the current situation.

We are investigating two types of summarization: *thread summaries* and *temporal summaries*. These allow a user to actively decide how much summarization they need. This can be useful when a user needs a summary of a long, important conversation, or when they need a summary of what has taken place since they stopped monitoring a chat room.

2.1 Thread Summarization

The first type of summarization we are investigating is a thread summary. This level of summarization targets individual conversation threads. An example of this is shown in Figure 1, where a summary would be generated of the messages highlighted in green, which all belong to the same conversation. An example output summary may then be:

> SoLo and SkyWlkr are talking about Ben. SkyWlkr thinks he's great, SoLo thinks he causes trouble.

As shown, this will allow for a summarization to focus solely on messages within a conversation between users. A good summary for thread summarization will answer three questions: *who* is conversing, *what* they are conversing about, and what is the *result* of their conversation. With our example, the summary answers all three questions: it identifies the two speakers SoLo and SkyWlkr, it identifies that they are talking about Ben, and that the result is SkyWlkr thinks Ben is great while SoLo thinks Ben causes trouble.

The key challenge to thread summarization will be finding, extracting, and summarizing the individual conversation threads. This requires the ability to detect and extract threads, which has become of great interest in recent research (Duchon and Jackson, 2010; Elsner and Charniak, 2010; Elsner and Schudy, 2009; Ramachandran et al., 2010; Wang and Oard, 2009). Thread disentanglement and summarization will have to be done online, with conversation threads being updated every time a new message appears. Another challenge will be processing incomplete conversations, since some messages may be incorrectly classified into the wrong conversation threads. These issues will need to be addressed as this research progresses.

2.2 Temporal Summarization

The other form of summarization we seek is a temporal summary. We want to allow users to dynamically specify the temporal interval of summarization needed. In addition, a user will be able to specify the level of detail of the summary, which will be explained further later in this section. An example of a user selecting a temporal summary can be seen in

Figure 1. A summary will be generated of only the text that the user selected, which is shaded in blue. An example output summary may then be:

> Rplee and Bish0p disagree if there is enough time to stay. SoLo and SkyWlkr are talking about Ben.

A good summary for this task will answer the following question: what conversations have taken place within the specified temporal interval. In some cases depending on the user's preference, not all conversations will be included in the summary. When not all conversations are included, then a good summary will consist of the most important conversations and exclude those which are deemed less important. The amount of detail to be presented for each individual conversation will be determined by the temporal interval and the level of detail requested by the user, which is discussed later in this section.

The summaries will need to be generated after a user selects the temporal interval. To aid in this, we envision that the summarizer will leverage the thread summaries. Conversations threads, along with their abstracts, will be stored in memory, and these will be updated every time a new message is received. The temporal summarizer can then use the thread summaries to generate the temporal summaries.

A user will also be able to specify the level of detail in the summary in addition to the temporal interval. When generating a temporal summary, a higher level of detail will result in a longer summary, with the highest level of detail resulting in a summary consisting of all the thread summaries within the temporal interval. In the case of a lower level of detail, the summarizer will have to determine which threads are important to include, and further abstract them to create a smaller summary. The benefit of allowing the user to specify the level of detail is so that they can determine how much detail they need based on personal requirements. For example, if someone only has a short amount of time to read a summary, then they can specify a low level of detail to quickly understand the important points discussed within the temporal interval they want covered.

Temporal summaries present additional challenges to address. The primary one is determining which conversation threads to include in the summary, which require a ranking metric. Additionally, there is an issue of whether to include a conversation thread if all messages do not all fall within the temporal interval. For example, if there is a long conversation composed of many messages, and only one message falls within the temporal interval, should it then be included or discarded? These issues will also need to be addressed as this research progresses.

2.3 Chat Corpora

An additional challenge of this work is finding a suitable chat corpus that can be used for testing and evaluating summarization applications. Most chat corpora do not have any summaries associated with them to use for a gold standard, making evaluations difficult. This evaluation difficulty is described further in Section 5.

Currently, we are aware of two publicly available chat logs with associated summaries. One of these is the GNUe Traffic archive[1], which contains human-created summaries in the form of a newsletter based primarily on Internet Relay Chat (IRC) logs. Working with these chat logs requires abstractive (i.e., summaries consisting of system-generated text) and extractive (i.e., summaries consisting of text copied from source material) applications (Lin, 2009), as the summaries are composed of both human narration and quotes from the chat logs.

The other corpus is composed of chat logs and summaries of a group of users roleplaying a fantasy game over IRC.[2] The summaries are of an abstractive form. Creating summaries for these logs is more difficult since the summaries take on different styles. Some summarize the events of each character (e.g., their actions during a battle), while others are more elaborate in describing the chat events using a strong fantasy style.

3 Related Work

Summarization has been applied to many different media (Lin, 2009; Spärck Jones, 2007), but only Zhou and Hovy (2005) have worked on summarizing chat. They investigated summarizing chat logs in order to create summaries comparable to the human-made GNUe Traffic digests, which were described in Section 2.3. Their approach clustered partial mes-

[1] http://kt.earth.li/GNUe/index.html

[2] http://www.bluearch.net/night/history.html

sages under identified topics, then created a collection of summaries, with one summary for each topic. In their work, they were using an extractive form of summarization. For evaluation, they rewrote the GNUe Traffic digests to partition the summaries into summaries for each topic, making it easier to compare with their system-produced summaries. Their approach performed well, outperforming a baseline approach and achieving an F-score of 0.52.

There has also been work on summarization of media which share some similarities to chat. For example, Zechner (2002) examined summarization of multiparty dialogues and Murray et al. (2005) examined summarization of meeting recordings. Both of these media share in common with chat the difficulty of summarizing conversations with multiple participants. A difference with chat is that both of these publications focused on one conversation sequentially while chat is characterized by multiple, unrelated conversations taking place simultaneously. Newman and Blitzer (2003) described the beginning stages of their work on summarizing archived discussions of newsgroups and mailing lists. This has some similarity with conversations, but a difference is that newsgroups and mailing lists have metadata to help differentiate the threaded conversations. Additional differences between chat and these other media can be seen in the unusual features not found in other forms of written texts, as described earlier in Section 1.

4 Planned Approach

We envision taking a three step approach to achieve our goals for this research. We will abstract this to a non-military domain, so that it is more accessible to the research community.

4.1 Foundation

The first step is to focus on improving techniques for summarizing chat logs in general to create a foundation for future extensions. With the only approach so far having been by Zhou and Hovy (2005), it is unknown whether this is the best path for chat summarization, nor is it known how well it would work for real-time chat. Also, since its publication, new techniques for analyzing multiparticipant chat have been introduced, particularly in thread disentanglement, which could improve chat summarization.

We hypothesize that constructing an approach that incorporates new techniques and ideas, while addressing lessons learned by Zhou and Hovy (2005), can result in a more robust chat summarizer that can generate summaries online. A part of this process will include examining other techniques for summarization, drawing on ideas from related work discussed in Section 3, such as leveraging latent semantic analysis (Murray et al., 2005). Furthermore, we will incorporate past work on dialogue act tagging in chat (Wu et al., 2005) to both improve summarization and create a framework for the next two steps. However, there is one limitation with their work: the templates used for tagging were manually created, which is both time-intensive and fragile. To overcome this, we plan to use an unsupervised learning approach to discover dialogue acts (Ritter et al., 2010).

4.2 Thread Extension

The second step will be to extend summarization to thread summaries. This will require leveraging thread disentanglement techniques, with the possibility of using multiple techniques to improve the capability of finding whole conversation threads. For the summary generations, we will first create extractive summaries before extending the summarizer to generate abstractive summaries. In addition, we will address the problem of incomplete conversations for the cases when not all messages can be extracted correctly, or when not all the messages of a conversation are available due to joining a chat room in the middle of a conversation.

Another task will be the creation of a suitable corpus for this work. As discussed in Section 2.3, there are only two known corpora with associated summaries. Neither of these corpora are well suited for thread summarization since the summaries are not targeted towards answering specific questions (see Section 2.1), making evaluations difficult. We plan on creating a corpus by extending an existing thread disentanglement corpus (Elsner and Charniak, 2010). This corpus consists of technical chat on IRC related to Linux, and has been annotated by humans for conversation threads. We will expand this corpus to include both extractive and abstractive summaries for each of the threads. The advantage

of using this corpus, beyond the annotations, is that it is topic-focused, which is a closer match of what one would expect to see in the military domain compared to social chat.

4.3 Temporal Extension

The third and final step will be to extend summarization to temporal summaries. The key point of this will be to extend the summarization capability so that a user can specify the level of detail within the summary, which will then determine the length of the summary and how much to include from the thread summaries. This will then involve creating a ranking metric for the different conversations. Unlike the thread extension, no additional abstraction will be needed. Instead, the temporal extension will reuse the thread summaries, and reduce their length by ranking the sentences within the individual summaries as done with traditional text summarization. Additionally, the problem of conversation threads containing messages both inside and outside the temporal interval will need to be addressed.

As with the thread extension, a corpora will need to be created for this work. We expect that this will build on the corpora used for the thread extension. This will then require additional summaries to be created for different levels of temporal intervals and detail. To make this task feasible, we will restrict the number of possible temporal intervals and levels of detail to only a few options.

5 Evaluation Issues

A major issue in summarization is evaluation (Spärck Jones, 2007), which is also a concern for this work. One problem for evaluation is the lack of suitable gold standards, as described in Section 2.3. Another problem is that we plan on working with abstractive forms in the future.

For the foundation step, we can follow the same procedures as Zhou and Hovy (2005), which would allow us to compare our results with theirs. This would restrict the work to only an extractive form for comparisons, though it is possible to extend to abstract comparisons due to the gold standards being composed of both extractive and abstractive means.

Evaluation for the thread and temporal extensions will require additional work due to both the lack of suitable gold standards and our need for abstractive summaries instead of extractive summaries. The evaluations will include both intrinsic (i.e., how well the summarizer is able to meet its objectives) and extrinsic evaluations (i.e., how well the summaries allow the user to perform their task, Spärck Jones (2007)). For the intrinsic evaluations, we will use both automated techniques (e.g., ROUGE[3]) and human assessors for evaluating both the thread and temporal summarizations. Some concerns for evaluation is that with the thread summaries, evaluation will be impacted by how accurately conversation threads can be extracted. With the temporal summaries, the temporal intervals and the level of detail determines the length and detail of the summary.

For the extrinsic evaluations, this research will be evaluated as part of a larger project, which will include human subject studies. Subjects will be situated in a simulated watchstander environment, must monitor three computer monitors simultaneously (one of which will contain live chat) while also listening to radio communications. Testing of our chat summarization methods will be done in collaboration with testing on 3D audio cueing to investigate and evaluate whether these technologies can help watchstanders combat information overload.

6 Conclusion

We have presented the challenges we face in chat summarization. Our goal for this research is that it will result in a robust chat summarizer which is able to generate abstract summaries in real-time. This is a difficult, exciting domain, with many possible applications. We have shown that the difficulties are due to the chat medium itself, lack of suitable data, and difficulties of evaluation.

Acknowledgements

Thanks to NRL for funding this research and to the reviewers for their valuable feedback. David Uthus performed this work while an NRC postdoctoral fellow located at the Naval Research Laboratory. The views and opinions contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of NRL or the DoD.

[3] http://berouge.com/default.aspx

References

Lindsley G. Boiney, Bradley Goodman, Robert Gaimari, Jeffrey Zarrella, Christopher Berube, and Janet Hitzeman. 2008. Taming multiple chat room collaboration: Real-time visual cues to social networks and emerging threads. In *Proceedings of the Fifth International ISCRAM Conference*, pages 660–668. ISCRAM.

Emily R. Budlong, Sharon M. Walter, and Ozgur Yilmazel. 2009. Recognizing connotative meaning in military chat communications. In *Proceedings of Evolutionary and Bio-Inspired Computation: Theory and Applications III*. SPIE.

Andrew Duchon and Cullen Jackson. 2010. Chat analysis for after action review. In *Proceedings of the Interservice/Industry Training, Simulation & Education Conference*. I/ITSEC.

LorRaine T. Duffy. 2008. DoD collaboration and chat systems: Current status and way ahead. In *Proceedings of the International Conference on Semantic Computing*, pages 573–576. IEEE Computer Society.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.

Micha Elsner and Warren Schudy. 2009. Bounding and comparing methods for correlation clustering beyond ILP. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 19–27. ACL.

Bryan A. Eovito. 2006. The impact of synchronous text-based chat on military command and control. In *Proceedings of the Command and Control Research and Technology Symposium*. CCRP.

Mark Handel and James D. Herbsleb. 2002. What is chat doing in the workplace? In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, pages 1–10. ACM.

Susan C. Herring. 1999. Interactional coherence in CMC. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences*. IEEE Computer Society.

Susan C. Herring. 2010. Computer-mediated conversation: Introduction and overview. *Language@Internet*, 7. Article 2.

Jimmy Lin. 2009. Summarization. In M. Tamer Özsu and Ling Liu, editors, *Encyclopedia of Database Systems*. Springer.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 33–40. ACL.

Paula S. Newman and John C. Blitzer. 2003. Summarizing archived discussions: A beginning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 273–276. ACM.

Gihan Osman and Susan C. Herring. 2007. Interaction, facilitation, and deep learning in cross-cultural chat: A case study. *The Internet and Higher Education*, 10(2):125–141.

Sowmya Ramachandran, Randy Jensen, Oscar Bascara, Todd Denning, and Shaun Sucillon. 2010. Automated chat thread analysis: Untangling the web. In *Proceedings of the Interservice/Industry Training, Simulation & Education Conference*. I/ITSEC.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. ACL.

Emad Shihab, Zhen Ming Jiang, and Ahmed E. Hassan. 2009. Studying the use of developer IRC meetings in open source projects. In *Proceedings of the IEEE International Conference on Software Maintenance*, pages 147–156. IEEE Computer Society.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

James Stavridis and Robert Girrier. 2007. *Watch Officer's Guide: A Handbook for All Deck Watch Officers*. Naval Institute Press, fifteenth edition.

Lidan Wang and Douglas W. Oard. 2009. Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 200–208. ACL.

Christopher C. Werry. 1996. Linguistic and interactional features of Internet Relay Chat. In Susan C. Herring, editor, *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, pages 47–64. John Benjamins.

Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler, and William M. Pottenger. 2005. Posting act tagging using transformation-based learning. In Tsau Young Lin, Setsuo Ohsuga, Churn-Jung Liau, Xiaohua Hu, and Shusaku Tsumoto, editors, *Foundations of Data Mining and Knowledge Discovery*, volume 6 of *Studies in Computational Intelligence*, pages 321–331. Springer Berlin / Heidelberg.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Liang Zhou and Eduard Hovy. 2005. Digesting virtual "geek" culture: The summarization of technical Internet Relay Chats. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 298–305. ACL.