Making Sense of Group Chat through Collaborative Tagging and Summarization

AMY X. ZHANG, MIT CSAIL, USA
JUSTIN CRANSHAW, Microsoft Research, USA

While group chat is becoming increasingly popular for team collaboration, these systems generate long streams of unstructured back-and-forth discussion that are difficult to comprehend. In this work, we investigate ways to enrich the representation of chat conversations, using techniques such as tagging and summarization, to enable users to better make sense of chat. Through needfinding interviews with 15 active group chat users, who were shown mock-up alternative chat designs, we found the importance of structured representations, including signals such as discourse acts. We then developed Tilda, a prototype system that enables people to collaboratively enrich their chat conversation while conversing. From lab evaluations, we examined the ease of marking up chat using Tilda as well as the effectiveness of Tilda-enabled summaries for getting an overview. From a field deployment, we found that teams actively engaged with Tilda both for marking up their chat as well as catching up on chat.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**;

Additional Key Words and Phrases: group chat; online discussions; online communities; collaboration; summarization; tagging; annotation; knowledge management; sensemaking

ACM Reference Format:
Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (November 2018), 27 pages. https://doi.org/10.1145/3274465

1 INTRODUCTION

Group chat applications have seen considerable growth in recent years, especially for coordinating information work. By enabling quick, team-wide message exchange in different channels, these applications promise to minimize the frictions of group communication, particularly for distributed and remote teams. Many organizations use systems such as Slack [6], HipChat [3], Internet Relay Chat (IRC) [59], Microsoft Teams [5], and Google Hangouts Chat [2] to make decisions, answer questions, troubleshoot problems, coordinate activity, and socialize. As of 2016, Slack alone reported having over 4 million daily users [47].

However, chat systems can have a number of downsides. Unlike email or forums, chat is predominantly synchronous, with a heightened expectation for quick responses and a high volume of back-and-forth messages exchanged in rapid succession [10]. As a result, chat logs are often comprised of a great many short messages forming multiple distinct yet intertwined conversation threads, with little distinction made between messages that are important and those that are not.

Authors' addresses: Amy X. Zhang, MIT CSAIL, 32 Vassar St. Cambridge, MA, 02139, USA, axz@mit.edu; Justin Cranshaw, Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA, justincr@microsoft.com.

2573-0142/2018/11-ART196 $15.00
https://doi.org/10.1145/3274465

This can make it difficult for members of the group who are not present in the conversation in real-time to make sense of it after the fact—for example, when someone falls behind, goes on vacation, revisits old chat logs, or is a newcomer to the group. Perhaps because of this burden of sifting through chat conversations, users have criticized group chat as encouraging an overwhelming "always on" culture, and some organizations have chosen to cast it aside [38, 41].

To make group chat conversations more comprehensible, we can build off of sensemaking affordances designed for other textual domains, such as email, online forums [53], or documents and general information management [31]. For instance, *tags* can be added to important messages to contextualize them or differentiate them from unimportant messages, similar to labels in emails or highlighted sentences in long documents. Furthermore, adding *structure* to the conversation could allow related messages to be grouped, much like distinct threads in an email inbox. Finally, both of these affordances facilitate the *summarization* of long back-and-forth conversations into a condensed format, much like notetaking in live meetings. Although these approaches to sensemaking have been explored in asynchronous discussion [36, 57, 89–91], little work has explored how to enrich synchronous chat conversations, which has additional challenges.

In this work, we consider how to apply these techniques *in situ*, enabling participants to enrich their discussions *while they are conversing*. We explore a variety of ways chat participants can mark up portions of their chat to create enriched, structured representations that allow users to get a high level overview of a full conversation and to dive in to parts of interest. Furthermore, our approach does not require a dedicated notetaker, allowing our design to conform to the spontaneous nature of group chat discussions. We conduct our analysis through an iterative design process, beginning with needfinding interviews and design mock-ups, and culminating in lab studies and a field study of a prototype system.

From interviews, we learned about the information management practices of 15 active group chat users, finding that many interviewees have trouble keeping up with chat and often miss important messages while scrolling up to read through their backlog. To ground the interviews, we created mock-up illustrations of different synthesized representations of a chat conversation, each emphasizing different information extracted from the conversation and varying degrees of structure. Some designs made use of tags on individual messages, others focused on extraction of important quotes, while still others involved written abstractive summaries. From showing the designs to our interviewees, we found a preference for more structured designs as well as signals such as major *discourse acts* [68] in a conversation, where discourse acts are categories of statements that characterize their role in the discussion (e.g. "question" or "answer").

Based on these findings, we developed Tilda, a prototype system built for Slack that allows discussion participants to collectively tag, group, link, and summarize chat messages in a variety of ways, such as by adding emoji reactions to messages or leaving written notes. Tilda then uses the markup left by participants to structure the chat stream into a skimmable summary view accessible within the chat interface. The summaries become live artifacts that can be edited, referenced, and posted to particular channels and individuals. Users can dive in to points of interest by following links in a summary to its place in the original chat stream.

We evaluated Tilda through three studies. First, we performed a within-subjects experiment to measure the effort required for groups to mark their chat while executing a shared task. We compared Tilda to using Slack alone and using Slack with a shared online document for notetaking. From 18 participants, we found evidence that Tilda was the better tool for taking notes while participating in the conversation. In a second experiment, we used the artifacts created in the first study to investigate the effort for a newcomer to comprehend the past conversations. From 82 participants, we found that users looking over summaries and chat logs enriched by Tilda felt less hurried when catching up compared to the other conditions. Additionally, those who utilized the links within Tilda summaries to dive into specific chat messages had a lower mental load and performed better at finding information from the chat log while still taking less time overall. Finally, we conducted a week-long field study of Tilda within 4 active Slack teams of 16 users total, and observed that teams actively used Tilda to mark up content and also found Tilda to be effective for catching up or looking back at old content.

2 RELATED WORK

2.0.1 Group Chat and Instant Messaging. The first group chat was developed at University of Illinois to connect users of an instructional system [82]. Since then, group chat, and its close relative instant messaging, have amassed billions of users world-wide [16, 18, 47, 83]. Chat has been extensively studied in numerous application areas, including how it can foster intimacy among friends and family [37, 81], how social norms form in online chat communities [65], how firms collaborate with chat in the workplace [32, 35, 63], how open source software developers coordinate in distributed teams [71], and how chat can lead to unintended consequences, such as a reduction in face-to-face communication, and increased interruption and distraction [10, 21, 29, 39]. Echoing this work, we also find unintended consequences and side effects in today's group chat, in particular, that people like the simplicity of having a single tool for rich interactions with their entire team, but struggle to keep up with the demands of information management.

2.0.2 Sensemaking of Online Conversations. Due to issues of scale and the lack of structure in online discussion, researchers have developed tools for making sense of large conversations, including tools to produce more structured and rich representations of discussion as well as tools for higher level synthesis. Techniques such as highlighting [91] or tagging [89] can assist with "foraging loops" during the sensemaking process [62], by providing more contextual signposts to readers while navigating through and diving in to the discussion space. Tools for higher level synthesis include visualization, clustering, and summarization techniques to more easily gain an overview. Some work has focused on the production side, including tools to turn groups or threads of discussion into short summaries [57, 90], or organize comments into topics [36]. Others automatically present more visual representations of the discussion, such as displaying opinions in a two-dimensional space [24] or portraying relationships between discussants [79], temporal activity [23], or reply structure [43]. However, most of these tools have focused on threaded discussion forums, while few exist for discussions with no reply structure. When it comes to chat, some work focuses on new chat representations, such as allowing people to have threaded conversations in chat [73] or time-based views [28]. Researchers have also looked at high-volume chat feeds as backchannels to live events, exploring how restricting feed visibility to virtual neighborhoods can help people manage the volume, enabling more participation [54]. However, such chats are rarely maintained as archival knowledge for the group, which is the scenario we study.

2.0.3 Notetaking and Live Meeting Notes. A common technique for synthesis when it comes to synchronous conversations in particular is the practice of notetaking during meetings. Research has demonstrated that notetaking is beneficial both to individuals, in improving learning and comprehension [34, 44], and to teams and organizations, in improving knowledge management practices and fostering collaboration [52]. During live meetings, it is common for teams and organizations to assign someone the role of designated notetaker [26], who may find it difficult to participate in the conversation due to the cognitive effort and split attention required to take notes [60, 61, 85]. Due to the cognitive load of synthesizing conversation, we consider how more lightweight techniques such as tagging or inline notes in the chat could make notetaking easier. We also consider how the work could be broken down and distributed among participants, both to lower individual load and spread the benefits of participation.

Many tools have been developed to improve notetaking in live meetings and lectures, including tools that enable participants to collaborate with shared notes [22, 42, 48, 66], tools for embedding notetaking within multimedia experiences [14, 15], and tools for leveraging meeting recordings to bootstrap notetaking [30, 55]. However, little research has been done looking specifically at notetaking during group chat, where conversations can often occur spontaneously.

2.0.4 Conversational User Experiences. In order to integrate seamlessly into chat conversations as they are ongoing, our Tilda prototype is developed as a Slack bot [51], exposing its functionality to the participants within their conversation. Chatbots have a long history in research [69], from initial explorations for fun and entertainment [80], to modern assistants offering a conversational interface to complex tasks [8, 11, 20, 25, 78]. Our system differs from many of these bots, in that it does not rely on natural language understanding [70], and is instead command driven, reacting only to specific user-input commands and actions. Several command-driven chatbots initially gained popularity in IRC communities [9], including Debian MeetBot [4], which is still actively used by organizations such as Ubuntu and Wikimedia to take notes during IRC meetings, or Zakim [7], which is in use at the W3C. MeetBot allows the leader of a chat meeting to designate the start and end of the meeting and enables participants to add different forms of notes to a running list of notes using hashtag commands. Similarly, Zakim is used during meetings for setting agenda items, reminders, speaking queues, and meeting scribes. While inspired by MeetBot, our prototype tool does not require scheduled meetings but can be used for more informal group chat conversations, with topics shifting continuously and people coming in and out throughout the day.

2.0.5 Microtasks. Microtask workflows are an effective way to break down complex tasks into manageable, independently executable subtasks that can be distributed to others and executed over time [46, 74]. They have been successfully used for taxonomy-creation [13], writing [76, 77], holding a conversation [50], transcription [49], and scheduling meetings [20]. In examining sensemaking of chat conversations, we were inspired to embed the concept of a microtask as way to “teamsource” the synthesis of a long chat conversation, a difficult task that often takes a dedicated notetaker.

2.0.6 Automatic Summarization. Finally, there is a long history of natural language processing research on automatic summarization [58]. While less work has focused on group chat, several projects look at summarization, including work on summarizing threads [64, 86], extracting important information from email conversations [67, 84], and analyzing audio conversations [56]. Building on this work, we provide insights into the level and type of structure people desire in synthesized representations of chat conversations. Our work points to the importance of *discourse act* tags for providing structure and context to chat messages. This finding has implications for prior work towards characterizing discussion using common sequences of major discourse acts [88]. Other work has looked at automatic extraction of major discourse acts such as questions and answer pairs [72] or tasks [19] from email, forums [17, 45], and chat [27]. However, a great deal of prior work builds models from data labeled by dedicated paid annotators. In this work, we examine lightweight ways to collect discourse act and other signals while collectively chatting, which could be used as richer training data towards automatic summarization of chat.

3 NEEDFINDING INTERVIEWS FOR MAKING SENSE OF GROUP CHAT

We began by interviewing active group chat users to understand how, why, and how often they go through prior chat conversations, and their strategies for and frustrations with making sense of long streams of chat messages.

3.1 Methodology

We conducted semi-structured interviews with 15 people who use group chat on a daily basis (6 female, 9 male, average age of 30.0). Interviewees were recruited through social media postings, email lists, and word-of-mouth, and were compensated $20 for their time. Individuals came from a diverse set of group chat teams, including tech companies, research groups, communities of interest, and co-working spaces. Groups ranged from as small as 4 people to over 500 people and from exchanging a few messages a day to thousands. Interviewees used a multitude of applications for group chat, including 11 on Slack, 4 on Microsoft Teams, 1 on HipChat, and 1 on WeChat.

We began by asking interviewees to open up the chat application for their most active chat group. We asked about how interviewees access their chat, their frustrations with using group chat, and their practices for managing the volume of chat messages they receive. We next sought to understand what content interviewees found important within their chat and which signals determine that importance. We asked interviewees to find an important conversation in their chat of which they were not a part and explain how they determined it was important and what they wished to glean from it. We then presented mock-up designs showing four different synthesized versions of the same conversation to them in randomized order, to probe their opinions about the type of information shown and the presentation of that information.

Interviews were conducted remotely by the first author and lasted 45-90 minutes. They were recorded and then transcribed using a paid transcription service. Then, the first author went through the transcripts and coded them for themes using an open coding approach [12]. Through multiple iterations along with periodic discussions with the rest of the research team, the coding led to 71 codes, from which the following major themes were selected. Because of the low number of interviewees, our interview findings should be regarded as indicative.

3.2 Current Experiences with and Strategies for Managing Group Chat

3.2.1 Participants have an "Always On" Mentality but Still Fall Behind. Almost all (14/15) interviewees kept their group chat application open on their computer or phone the entire day, echoing reports that users of Slack have it open 10 hours on average per weekday [40]. Interviewees cited many reasons for being continually present, including being "on call" to answer urgent messages, seeking to gain an ambient awareness of others' activities, a concern about "missing out", and disliking having to deal with a backlog of missed conversations. But several interviewees acknowledged downsides of continually being on chat, with one saying, "*I think there's a lot of content that I don't need to consume. I've read [that] content switching is distracting and bad for productivity...But I hate having unread notifications.*" Most interviewees (11/15) also mentioned checking chat while not working or on vacation, and checking it more often than they would have liked. Despite their efforts, falling behind was a common occurrence (13/15 interviewees). Some interviewees blamed the volume of messages while others had trouble distinguishing relevant information: "*There are so many things happening at the same time...I had a very hard time [determining] what are relevant for me, and what are the things I don't really need to care about at all.*" Still others purposefully let messages go unread in certain channels or groups because the ratio of important to unimportant messages was low or they had only a passing interest in the topic.

3.2.2 Newcomers are Overwhelmed by Chat History. Besides active members, newcomers are another population that may desire to go through concluded conversations. A few interviewees (4/15) talked specifically about the newcomer experience of joining a chat group. They described it as overwhelming and tedious, but they still felt compelled to go back over the chat history to get better acquainted with the team and the work. For instance, one interviewee said "*...there was a whole history of stuff that I wanted to know about so that we could not reinvent the wheel, so that we* *could understand where ideas are coming from...It was not so much about missing stuff. It was more coming into a new thing...wanting to know what is it? Because you just can't read back through it all.*"

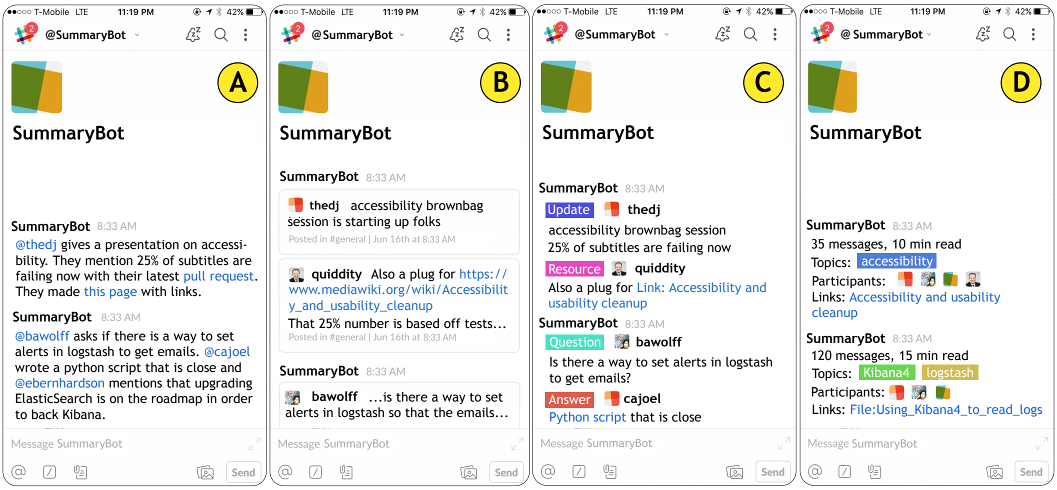
3.2.3 Strategies for Catching Up are Unsatisfactory. When looking back through chat history, either to catch up or to get acquainted with a group, we found that the dominant strategy (9/15) was to skim content by scrolling up in their chat window. However, several expressed frustration with this strategy, with one interviewee saying, "*Scrolling is basically the big issue, which is that you've got this giant timeline of stuff...You can only scroll and skim back through so many views on the viewport before you start getting tired of looking.*" Other interviewees echoed this sentiment, pointing to how chat logs are poorly segmented by discussion topic, contain a great deal of back-and-forth before reaching a conclusion, and intersperse important information with humor or chit-chat, providing little ability to distinguish the two. One interviewee said "*...there's a lot of conversation, and it all concludes with some result...all I want is results...then I wouldn't have to read 300 back-and-forths.*"

When falling behind, several interviewees (6/15) also simply chose to ignore missed messages, assuming they were irrelevant by then or that important information would reach them eventually, such as by email. This strategy exacerbated issues such as questions that were continually re-asked, or important requests that went unanswered. One interviewee said, "*[Someone] was requesting help for something...I knew when I read it that everyone was going to ignore it because it was going to get lost in the Slack channel...it was a really important thing but it was just a lot easier to ignore...it just sort of gets pushed up...*" Even though interviewees felt that important information would eventually reach them, several (5/15) could remember specific instances when they had missed important information that they wished they had known about. In these cases, someone neglected to mention them in the conversation, or an announcement was made that got lost among other messages, or they had a passing interest in a channel but no way of occasionally dipping in to catch up on important happenings.

3.2.4 Recalling or Re-finding Information is Hard. Another way to explore a long chat stream is to use search to filter for specific conversations. Half of the interviewees (7/15) had trouble searching back over chat conversations to find information. Interviewees, when trying to recall conversations they were a part of, needed to remember particular phrases that were said or other details, with one saying "*...if you don't know exactly what you're looking for, or if you misremembered the word...search begins to be fairly limited...Usually you'd need two to three bits of information. A word, a person...[otherwise] there might be months' worth of stuff...*" Interviewees who couldn't pinpoint information with search resorted to scrolling in the surrounding area of the search results, encountered the same issues with scrolling mentioned earlier.

Related to the strategy of expecting important information to arrive through multiple avenues, a few interviewees (4/15) also described conversations spilling over from chat into email, making it harder to retrace what happened. One interviewee said "*It's especially annoying if this conversation started here and then there was an email thread, and it was hard to interlace the two chronologically.*" Another interviewee, catching up from vacation, made a note to respond to an unanswered request in chat but missed that it had been responded to in an email. Thus, using multiple channels for pushing out information may make it difficult to recall where conversations took place.

3.2.5 Existing Processes for Organizing Information are Cumbersome. In response to difficulties with finding or catching up with chat conversations, some interviewees described policies the group had instated to collect knowledge. However, many of these were unsuccessful because of the cumbersome nature of the process, leading to lack of adherence to the policy or lack of maintenance over time. For instance, several interviewees (5/15) had a separate knowledge store, such as a community Q&A site, collaborative documents, or a wiki. One interviewee, discussing finding



A

SummaryBot 8:33 AM
@thedj gives a presentation on accessibility. They mention 25% of subtitles are failing now with their latest pull request. They made this page with links.

SummaryBot 8:33 AM
@bawolff asks if there is a way to set alerts in logstash to get emails. @cajoel wrote a python script that is close and @ebernhardson mentions that upgrading ElasticSearch is on the roadmap in order to back Kibana.

B

SummaryBot 8:33 AM
thedj accessibility brownbag session is starting up folks
Posted in #general | Jun 16th at 8:33 AM

quiddity Also a plug for https://www.mediawiki.org/wiki/Accessibility_and_usability_cleanup
That 25% number is based off tests…
Posted in #general | Jun 16th at 8:33 AM

SummaryBot 8:33 AM
bawolff …is there a way to set alerts in logstash so that the emails…

C

SummaryBot 8:33 AM
Update thedj
accessibility brownbag session
25% of subtitles are failing now
Resource quiddity
Also a plug for Link: Accessibility and usability cleanup

SummaryBot 8:33 AM
Question bawolff
Is there a way to set alerts in logstash to get emails?
Answer cajoel
Python script that is close

D

SummaryBot 8:33 AM
35 messages, 10 min read
Topics: accessibility
Participants:
Links: Accessibility and usability cleanup

SummaryBot 8:33 AM
120 messages, 15 min read
Topics: Kibana4 logstash
Participants:
Links: File:Using_Kibana4_to_read_logs

Fig. 1. Some examples of mock-ups shown to interviewees to compare and contrast different synthesized chat designs: A) abstractive, B) extractive, C) discourse act labels, D) high level signals.

answers to questions, said he preferred to search the chat history instead of his company's internal community Q&A site because people often failed to post to the Q&A site or update their post with the answer. This was considered a documentation chore, uncoupled to the main goal of getting the question answered, despite being considered a good practice in the team. Two interviewees also mentioned how people summarized accumulated pinned messages in Slack into Google Docs files; however, the files were rarely used and quickly forgotten due to their lack of visibility in the chat system. Another interviewee complained about how it always fell to the same people to organize information from chat, highlighting the diffusion of responsibility due to the group setting.

3.2.6 Summary. We found that many interviewees spend a significant amount of time scrolling through their chat history, despite being continuously available on chat, and face frustrations with differentiating content when doing so, leading to missed important information. We also saw how conversations that start in chat sometimes get picked up in email or vice versa, making them hard to follow and re-find. This suggests that tools could better bridge and link more synchronous communication systems such as chat to more asynchronous ones such as email. Similarly, we saw that attempts to synthesize information from chat failed because they were poorly integrated, due to being in a separate location and with a workflow separate from chatting. This suggests that tools for enriching or synthesizing chat should be tightly integrated into the chat environment, and any artifacts created should also be coupled to the original discussion.

3.3 Preferences for the Content and Presentation of Synthesized Chat Designs

Next, we sought to learn from our interviewees what information from a chat conversation is useful for determining importance, as well as what presentation of that information is best for gaining an overview quickly. We did this by asking interviewees to find an important chat conversation from their chat history to talk about as well as give their impression of four different design mock-ups that we prepared beforehand. We presented the design mock-ups to interviewees in a randomized order, and for each, asked interviewees what aspects they liked and disliked. At the end, we asked interviewees to compare the designs and discuss which ones they preferred and why.

The mock-ups were conceived by surveying existing applications for enriching or synthesizing conversations. They were also chosen to encompass a diversity of types of information, from excerpts to topics to discourse acts, as well as a range of presentations, from less structured to more structured, to elicit interviewees' reactions. Figure 1 shows examples of the four mock-up types. Design A presents a written abstractive summary of the discussion in the form of short sentences, inspired by the practice of notetaking in meetings. Design B is an extractive summary made up of important excerpts taken directly from the chat log, inspired by tools like Digest.AI[1] or Slack's Highlights feature [1]. Design C augments excerpts of the conversation by tagging them with major discourse acts, similar to tools like Debian MeetBot [4]. Finally, Design D showcases high level signals, such as main participants, number of messages, topic tags, and a subject line, inspired by affordances in major email clients. We created two examples for each design, with conversations taken from the same chat from a Wikipedia IRC chat log. We asked interviewees to assume that all designs are manually created to sidestep concerns about perceived feasibility of automation.

3.3.1 A Purely Extractive Approach Lacks Context. Only one interviewee preferred a purely extractive approach (Figure 1-B) for getting an overview, stating that she preferred to read people's contributions in their own voice. However, most interviewees did not like this design because of the loss of context, with one interviewee stating, "*A lot of these messages are very much conversational, and so unlike an email where everything is self contained, it's a flow. So just pulling out a single message does lose some of that important context.*" This was surprising given the number of existing tools that use an extractive approach. Two interviewees were aware of the Slack Highlights feature [1] that shows automatically extracted important messages, but expressed the same concern.

3.3.2 A Purely Abstractive Approach Lacks Structure. Alternatively, only 3/15 interviewees liked the purely abstractive approach (Figure 1-A). This was also surprising given that abstractive summaries of a conversation would likely be the most labor-intensive to create and is often considered a gold standard in summarization tasks. The interviewees that liked this design liked that it was possible to gain a comprehensive understanding of what happened, while other designs offered an indication but would need further investigation. However, most interviewees objected to this design because they found it too difficult to skim due to the lack of structure. One interviewee said "*I have no way of knowing almost until I finished this thing whether or not I'm interested. It doesn't save me any time triaging.*" Two interviewees also mentioned needing to trust the writer of the summary and were concerned about variability in quality.

3.3.3 Signals about Topic, People, and Volume are Informative and Easy to Skim. Eight interviewees liked the design exploring different high-level signals about a conversation (Figure 1-D), with most commenting on the additional structure provided. One interviewee said "*I can decide on the outset if I care about the thing that was discussed or not, and if I don't care, then I move on. I don't like the clutter of having long or multiple messages.*" Many interviewees found signals such as topic keywords, a main subject line, major participants, and the number of messages or an estimate of reading time to be informative.

3.3.4 Discourse Act Tags Add Context to Extracted Messages. Finally, the design exploring the use of major discourse acts as labels to group notes was by far the most popular, with 14/15 interviewees preferring this design (Figure 1-C). Given the additional structure, interviewees felt they had a greater ability to skim and home in on specific categories of interest, such as unanswered questions, which was difficult in the abstractive or extractive designs. But unlike the design with only high-level signals, this design still provided information about what occurred in the discussion.

[1] https://slackdigest.com/

One interviewee said "*I love the tags. I love the fact that sometimes you have a question and now the question leads to an answer...It tells me how to read the content.*" The improvement over a purely extractive approach was the ability for the discourse acts and links between them to provide a narrative for the extracted messages.

Given the emphasis that interviewees placed on major actions over the course of a conversation, we asked interviewees to consider what kinds of discourse acts they would want to have highlighted. The following discourse types were mentioned:

- **Action items**: Several interviewees mentioned wanting a way to track assigned action items or any follow-up to-dos that resulted from any kind of discussion.
- **Troubleshooting**: Several interviewees also mentioned the importance of marking problem statements, the resolution of troubleshooting discussions, as well as suggestions or ideas to solve them. Interviewees also wanted to easily see which problems were still ongoing.
- **Deliberation**: Interviewees mentioned having many scheduling discussions or debates. They thought of labeling these with a problem statement along with a decision marking the outcome or pros and cons labeled separately.
- **Questions and answers**: Similarly to problems and solutions, interviewees wanted to highlight questions, along with their answers, as well as any unanswered questions.
- **Announcements, links, tips**: Finally, interviewees saw a use case for labeling announcements and links to items, as well as observations, tips, or other useful one-off information.

3.3.5 Hierarchical Exploration Manages Volume and Provides Agency. Finally, interviewees described how they would prefer to interact with synthesized representations of chat. Some interviewees (4/15) desired some sort of ability to explore hierarchically, whether that be from the summary to the original discussion or from a shorter summary with high-level signals, to a longer summary that contained excerpts. One interviewee stressed the importance of controlling exploration, saying "*I want to scroll through it and zoom in and out of it...skim, but skim with a bit more intent. I might be more likely to use...something a bit more interactive. I don't want to just be told...I want to be helped.*" Another interviewee wanted a different level of depth depending on how much conversation they had missed; the more they missed, the shorter each individual summary should be.

3.3.6 Summary. From the feedback that the mock-ups prompted, we found that interviewees preferred a high degree of structure to aid their sensemaking. At the same time, they were interested in cues that could provide context about what happened in the discussion. This feedback suggests that a hybrid approach combining structured high-level signals about a conversation with important excerpts marked with their discourse act could be both easily skimmable yet contextual. Finally, we found that interviewees desired the ability to use summaries to guide deeper exploration. This suggests that summary views could have different hierarchies of synthesis, with a shorter initial representation leading to a longer one, eventually leading to the original discussion.

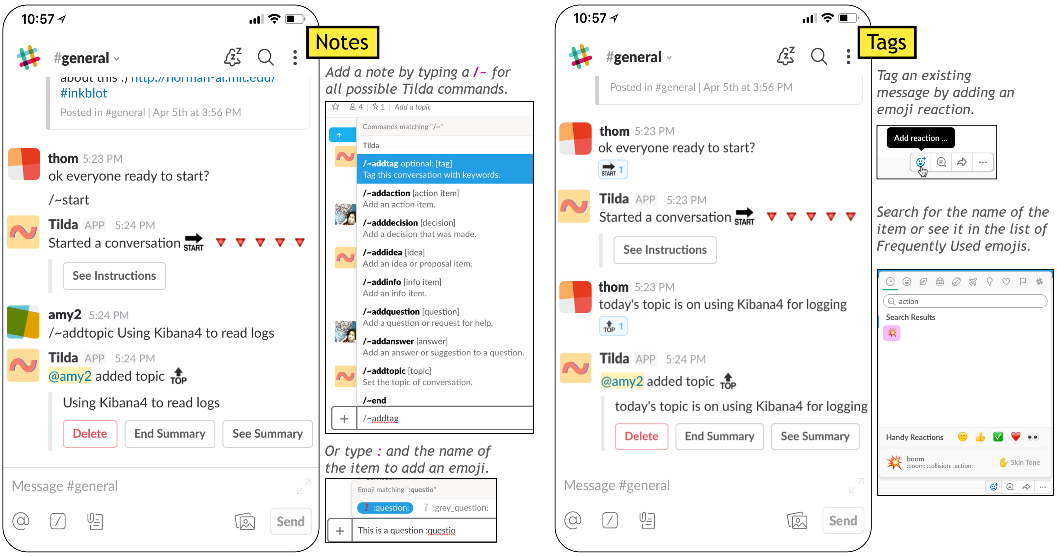
4 TILDA: A TOOL FOR COLLABORATIVE SENSEMAKING OF CHAT

Building on the findings of our interviews, we developed a prototype system called Tilda[2], instantiated as a Slack application, for participants in a group chat conversation to collectively mark up their chat to create structured summaries, using lightweight affordances within Slack.

4.1 Enriching Chat Conversations using Notes and Tags

4.1.1 Techniques for Enriching Chat. Tilda provides two main techniques for enriching a chat conversation, as shown in Figure 2. The first way is through inserting a **note** while in the course of

[2] Visit tildachat.com. Tilda sounds somewhat like pronouncing "TL;DR" (too long; didn't read). The logo for Tilda is a tilde.



about this :) http://norman-ai.mit.edu/
#inkblot
Posted in #general | Apr 5th at 3:56 PM

thom 5:23 PM
ok everyone ready to start?
/~start

Tilda APP 5:24 PM
Started a conversation START ▼ ▼ ▼ ▼ ▼

amy2 5:24 PM
/~addtopic Using Kibana4 to read logs

Tilda APP 5:24 PM
@amy2 added topic TOP
Using Kibana4 to read logs

Add a note by typing a /~ for all possible Tilda commands.

Add a topic

Commands matching "/~"

Tilda

/~addtag optional: [tag]
Tag this conversation with keywords.

/~addaction [action item]
Add an action item.

/~adddecision [decision]
Add a decision that was made.

/~addidea [idea]
Add an idea or proposal item.

/~addinfo [info item]
Add an info item.

/~addquestion [question]
Add a question or request for help.

/~addanswer [answer]
Add an answer or suggestion to a question.

/~addtopic [topic]
Set the topic of conversation.

/~end

/~addtag

Or type : and the name of the item to add an emoji.

Emoji matching ":questio"

:question: :grey_question:

This is a question :questio

Posted in #general | Apr 5th at 3:56 PM

thom 5:23 PM
ok everyone ready to start?
START 1

Tilda APP 5:23 PM
Started a conversation START ▼ ▼ ▼ ▼ ▼

thom 5:23 PM
today's topic is on using Kibana4 for logging
TOP 1

Tilda APP 5:24 PM
@amy2 added topic TOP
today's topic is on using Kibana4 for logging

Tag an existing message by adding an emoji reaction.

Search for the name of the item or see it in the list of Frequently Used emojis.

action

Search Results

Handy Reactions

boom
:boom: :collision: :action:

Fig. 2. The main techniques for adding metadata in Tilda include notes and tags. On the left, the chat is enriched in real time by injecting notes using slash commands or inline emojis. On the right, the chat is marked up by adding tags to pre-existing messages using emoji reactions.

conversation. A user may add a note by using a custom *slash command*, Slack's feature for invoking commands within the dialog box, or by adding a custom *inline emoji* to the text of their message. Slash commands allow users to type a slash in order to pull up an auto-completed list of commands. For this reason, all Tilda commands are prepended with a tilde. Some types of notes consist solely of the command, such as a note to designate the start or end a conversation. Other notes contain textual content, such as the marking of a conversation's topic or the addition of a question. Each note gets added as a chat message to the transcript of the chat log when they are created.

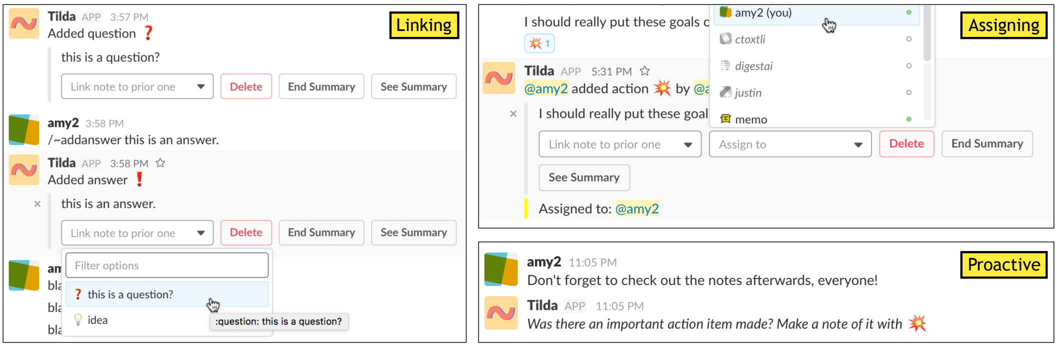
The second way is through **tagging** of existing chat messages using custom *emoji reactions*, a feature in Slack, as well as common in other messaging systems such as Facebook Messenger, where any user can attach an emoji to the bottom of an existing message. Users can use this method to tag any pre-existing message going back in time, and so can choose to mark up an old conversation or one as it is ongoing. Users can use tags to designate messages as the start or end of a conversation or mark messages with their discourse act, such as a question or an answer. Unlike slash commands and inline emojis, one can add an emoji reaction to anyone's chat message, not just their own.

For each item added, whether by note or tag, the Tilda application posts a message in the chat documenting the action and allows the user to undo their action, toggle to see the current state of the items in the conversation, or interact with the items in other ways.

4.1.2 Categories of Tags or Notes. Using either of these two techniques, users can add a variety of metadata to their chat conversation (see Table 1 for a complete list). First, as mentioned above, users can mark the beginning and end of conversations as a way to segment the chat stream and **group** a series of items together. This can be done using either the note or tag technique. For convenience, conversations also automatically start whenever a new piece of metadata is added to the chat, and they automatically end if there is no activity for 20 minutes, though this can be undone if it was premature. In between start and end markers, users can mark up the chat by

Label	Command	Emoji	Function
Action	\~addaction	💥	Add action item
Answer	\~addanswer	❗	Add answer item
Decision	\~adddecision	🤝	Add decision item
Idea	\~addidea	💡	Add idea item
Question	\~addquestion	❓	Add question item
Topic	\~addtopic	🔝	Add topic of conversation
Tag	\~addtag		Add custom tag to conversation
Start	\~start	START	Start a new conversation
End	\~end	🔚	End current conversation

Table 1. List of discourse act items and their commands and emojis, as well commands and emojis related to conversation-level markup, including adding a topic or custom tag and starting or ending the conversation.



Tilda APP 3:57 PM
Added question ❓

this is a question?

amy2 3:58 PM
/~addanswer this is an answer.

Tilda APP 3:58 PM
Added answer ❗

this is an answer.

❓ this is a question?

💡 idea

I should really put these goals c

💥 1

Tilda APP 5:31 PM
@amy2 added action 💥 by @a

I should really put these goal

Assigned to: @amy2

amy2 11:05 PM
Don't forget to check out the notes afterwards, everyone!

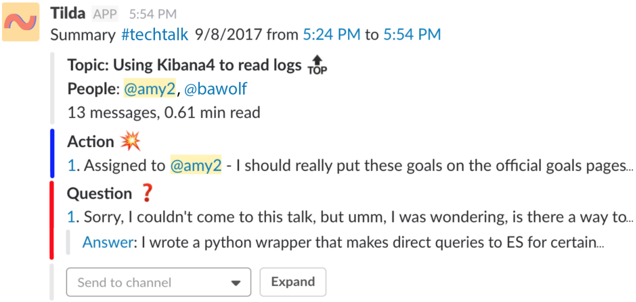
Tilda APP 11:05 PM
Was there an important action item made? Make a note of it with 💥

Fig. 3. Examples of linking a Tilda item to a prior item, assigning an Action item to a member, and getting a proactive nudge to annotate a message.

contributing Tilda items to an ongoing summary of a conversation. The possible discourse acts, as seen in the first five rows in Table 1, correspond to the types of discussion actions that interviewees wished to have highlighted. In addition, users can add a topic sentence to a conversation or add a custom topic tag to the conversation, two signals our interviewees found informative.

4.1.3 *Adding Additional Context.* In addition, we provided other abilities to add structure based on findings from our interviews. First, a user can **link** a Tilda item to a prior one, as shown in Figure 3. This can be used when an item should be seen in context with another item for it to be better understood. For instance, an Answer item could be linked to its corresponding Question item. Linking is facilitated by a dropdown menu in the chat message that the Tilda application posts. For Action items in particular, users can also **assign** the item to a person who is a member of the channel. This was added because several interviewees were interested in tracking to-dos that arose due to discussion. Any user can assign the Action item or re-assign it at a later point in time.

4.1.4 *Encouraging Participation.* Finally, to encourage or remind users about notetaking, Tilda proactively posts suggestions to add a tag when it notices certain activities, as seen in Figure 3. These activities were determined manually and encoded in Tilda as explicit rules. For instance, if a user *stars* a recent message, a feature in Slack to private pin messages to a list, Tilda will post a suggestion to annotate it with a discourse act. Second, we manually devised a number of



Tilda APP 5:54 PM
Summary #techtalk 9/8/2017 from 5:24 PM to 5:54 PM

Topic: Using Kibana4 to read logs 🔝
People: @amy2, @bawolf
13 messages, 0.61 min read

Action 💥
1. Assigned to @amy2 - I should really put these goals on the official goals pages...

Question ❓
1. Sorry, I couldn't come to this talk, but umm, I was wondering, is there a way to...
Answer: I wrote a python wrapper that makes direct queries to ES for certain...

Fig. 4. Example Tilda summary generated from user tags and notes. The summary is grouped by discourse act, expandable, and each note is linked to its place in the original chat.

phrases associated with each discourse act type based off of conversations we saw in pilots, such as “remember to” with “Action”. When Tilda sees such a phrase, it posts a suggestion to annotate the message with the corresponding discourse act. In the future with more data, one could imagine moving to machine learned suggestions trained by prior tagged messages.

4.2 Synthesizing Chat Conversations using Structured Summaries

The notes and tags that users leave behind using Tilda can be immediately used by readers scrolling up through the chat log. Tilda also gathers them into structured summaries that allow a reader to get an overview of a discussion as well as dive in to the original chat.

4.2.1 Presentation of Summaries. Figure 4 shows an example of a summary in Tilda. Based off of feedback from interviewees, the summary includes signals about the conversation, such as number of messages and estimated read time, major participants, any custom tags that users have added to the conversation, and a topic sentence if it exists. It also presents the items that users added grouped and colored by their discourse act type. If an item was linked to another item, it appears underneath and indented to the right. Because users may leave many items in a single conversation, we only show a subset in the summary with the ability to expand to see all. The subset is determined using a heuristic that prioritizes categories like Topic and Action and limits each category to the first two items left chronologically. Upon expanding, users may sort all the items chronologically or grouped by category. Each item is also preceded by a hyperlink pointing to where it originally took place in the chat log, providing the hierarchical exploration and deep integration between summary and discussion that interviewees desired. In addition, because all items in the summary originate as markup in the original chat log, any edits to the content or markup in the original discussion automatically updates the summary, making it a live artifact wherever it is displayed.

4.2.2 Delivery of Summaries. One way that summaries can be delivered is through **following** the summaries of a particular channel. Any user can, in another public or private channel, set that space to follow the summaries of a public channel using the slash command `\~followchannel #channelname`. Users can specify parameters in the `followchannel` command to limit summaries to only those containing a particular participant or tag. From then on, all summaries matching the parameters and generated in the original channel will get posted to its designated places. In this way, Tilda could be used to take discussions from a smaller or private channel and have them summarized to a larger or public one.

One potential way to set up Tilda is to create team-wide "summary channels" that follow the summaries of one or more other channels. Another more personalized way to use Tilda is for a user to subscribe to the summaries from one or more channels in their direct message with Tilda. Finally, users also have the ability to selectively send a single summary to a channel using a dropdown, as seen in Figure 4.

4.3 System Implementation and Considerations

Tilda is implemented as a Slack application, with messages from Tilda arriving in the chat log, similarly to a chatbot. It is built on top of the Microsoft Bot Framework, an SDK that allows one to develop chatbots for a number of applications at once, and the Slack API. The backend server is built in Node.js and interfaces with a MongoDB database.

Several considerations went into the implementation of Tilda. First, we chose to develop a Slack application over developing a separate chat system or a browser extension because Slack applications can be quickly installed to any team already on Slack using OAuth authentication. Additionally, users can use it in any browser of their choosing on mobile, tablet, or desktop. We chose to implement for Slack over other chat systems such as IRC because of the ability to use Slack-specific features such as custom slash commands and emoji reactions, as well as create interactive and dynamic prompts within chat messages. Finally, we chose to build sensemaking capabilities into a chat system as opposed to designing a separate system that imports chat messages. We chose this direction after encountering difficulties with understanding chat after the fact, which we uncovered while piloting interfaces and workflows for marking up a pre-existing chat log.

However, these decisions also required us to make some trade-offs due to the limitations of Slack's API. For instance, the only way to communicate with users or add affordances beyond commands and emojis is to post a message in the chat as a bot. But due to the space they take up, messages posted by Tilda could pollute the chat stream. Additionally, summaries can only be presented via a chat message, which may be difficult for users already juggling multiple channels. A more integrated approach might have summaries overlaid on top of or directly alongside the original discussion. In the future, these ideas could be explored in a novel chat system or an extension that can alter the existing interface. In the meantime, our prototype allows us to quickly experiment with and deploy techniques for enriching and synthesizing chat in real-world settings.

5 EVALUATION

We conducted two lab evaluations of Tilda to study how easy it is to enrich chat conversations while chatting as well as to study the experience of catching up on missed conversations using structured summaries. While these lab studies enabled us to examine specific facets of Tilda usage in detail, they were necessarily conducted under artificially constrained setting. To examine Tilda in more naturalistic chat settings, we also conducted a field study, where we observed expert Slack users from real organization use Tilda while they conducted their normal activities.

5.1 Study 1: Marking Up Chat Conversations While Chatting

In the first lab study, we considered the common scenario where chat participants wish to make note of important discussion items while they also actively conversing. We conducted a within-subjects experiment that compared using Tilda for keeping notes to more traditional methods such as collaborating on a shared online document for notes, or not taking notes at all.

While it is common for group chat conversations in real organizations to be partially asynchronous, focusing on notetaking during *active* discussions enabled us to explore the cognitive load and cost of switching contexts between participating in chat and marking content with Tilda, as it compares to using an online collaborative document. We were also interested in understanding whether any benefits from notetaking would justify the added overhead of keeping notes.

We recruited 18 participants (mean age 36.6, 8 female, 10 male) from UpWork, a paid freelancing platform, at the rate of $20 per hour, with each participant working around 2.5 to 3 hours in total depending on their assigned conditions. Participants were all based out of the U.S., native or fluent English speakers, and somewhat or very tech-savvy, though 6 participants were new to Slack. Participants were placed randomly into groups of 3, with 6 groups total.

5.1.1 Discussion Tasks. We devised two collaborative tasks that each group would perform together. The tasks were chosen because they were comprised of many smaller parts that needed to fit together, and they involved deliberation as opposed to simply compiling or coordinating information. The tasks were:

- **Story**: Collectively come up with a new T.V. show based on the show Friends. Participants were asked to come up with the cast, location, and the plot of a 5-episode season.
- **Travel**: Plan a month-long cross-country roadtrip in the U.S. Participants were asked to pick 5 major cities and national parks and other landmarks to visit, as well as the route, transportation, and accommodations.

5.1.2 Experiment Design. Every group of 3 completed the Story task first, followed by the Travel task. Each task was completed in one of the following three conditions:

- **Tilda**, where the group used Slack with Tilda to discuss the task and mark up their chat,
- **Doc**, where the group used Slack to discuss the task and take notes using a shared Google Doc, and
- **None**, where the group used Slack to only discuss the task.

Since there were two tasks per group, each group participated in a pair of conditions. Thus, for every pair of conditions, two groups out of the six groups total were assigned that pair. To account for ordering effects, we counterbalanced the condition order, so groups with the same pair of conditions received a different condition first.

To start a study session, we invited everyone in a group to a Slack channel, where we spent 30 minutes on an icebreaker and a tutorial on Slack administered via a Word document shared with the group. Then users worked on their first task for 45 minutes and completed a post-task survey rating their experience. They were then invited to a different Slack channel where they worked on the second task for 45 minutes, completed the same survey, and then completed a survey comparing the two conditions. They then collectively participated in a debriefing discussion in Slack with the authors about their experience where we asked them to compare conditions.

Before the Tilda condition, we gave users a 30 minute tutorial covering advanced Slack features and Tilda, again using a Word document shared with the group. During this session, users got acquainted with Slack slash commands, inline emojis, and emoji reactions, as most of our subjects did not have much familiarity with these features. In addition, the tutorial provided a basic overview of Tilda, covering the different types of notes and tags one could leave using Tilda. Before the Doc condition, we gave users access to a shared Google Doc for notetaking. There was no tutorial for Google Docs as all our users stated they were experienced Google Docs users.

In the Doc and Tilda conditions, we required users to keep track of their conversation using the provided tools. Users were also told before the study that they would debrief the authors afterwards about what they decided so as to motivate them to keep better notes.

5.1.3 Results. We compare each condition against each other. Due to the small sample size, the results are not statistically significant, except where indicated otherwise. Instead, we present more qualitative findings and observations that should be regarded as indicative.

Tilda versus Doc: All 6 users that were in both Tilda and Doc conditions marked Tilda as substantially better at keeping track of what happened in the discussion. Additionally, 4/6 users thought Tilda was somewhat or a lot better for participating in the discussion, and most preferred to use Tilda for the same task again (5/6). One user said "*Honestly now that I know about Tilda I would never use Google Docs for brainstorming ideas with others. Tilda is way simpler.*" Other users talked about being more organized with Tilda: "*...the Google Doc was hard to follow if you didn't know what it was already about but I feel Tilda kept all of our ideas organized and made it easier to follow*", with 5/6 users marking that Tilda was a lot better for looking back over the discussion. However, 3/6 users found Google Docs to be easier to use for notetaking than Tilda. This may partially be because they just learned Tilda but were experienced Google Docs users: "*I think because I use Google Docs regularly, it makes more sense to me. But Tilda captures a conversation better.*" We also analyzed post-task survey ratings of all Tilda conditions and all Doc conditions, finding that people in Tilda conditions rated themselves on a 5-pt Likert scale as more successful in accomplishing their task (N=12, 4.25 vs. 3.58, $p < 0.1$).

Tilda versus None: For the users that compared using Tilda versus using only Slack, 4/6 found Tilda to be better for keeping track of what happened during the discussion, and 5/6 found Tilda to be better for looking back over the discussion. One participant mentioned the hyperlinks in the summaries, saying "*I loved how Tilda let you click on links to go back to the original messages instead of having to manually scroll through myself.*" However, only half found the Tilda condition better for participating in the discussion, and 5/6 users found None easier to use. Users in the post-task surveys also rated Tilda as more mentally demanding than using Slack alone (3.83 vs. 2.83, $p < 0.05$). This is not surprising given that the Tilda condition explicitly involves doing more than the None condition. As to whether the benefits of Tilda outweigh the costs, 3/6 stated they would use Tilda again for the same task while 2/6 preferred just using Slack. One participant said "*Slack...is easier to use just because there is less to keep track of, but for organization, Tilda is the way to go*", which suggests that the cognitive load introduced by Tilda might be worth it for more demanding tasks. Another user said "*If I was working in a corporate or work environment and in project management, Tilda would be perfect.*"

Doc versus None: In comparison, the 6 participants in Doc and None conditions overall rated Google Docs more poorly, with only 1/6 users preferring the Doc condition for participating in the discussion, 3/6 for keeping track of what happened, and 3/6 for looking back over the discussion. Only one user preferred to use Google Docs and Slack again for the same task while 3/6 preferred to use just Slack. In discussions, users complained about fragmented attention in the Doc condition, with one person saying "*If you have multiple tools open then it's not clear where all of the people, where their focus is directed to.*" Users also disliked how information was scattered in both the Google Doc notes and the chat log, saying: "*If I come back to many ideas I don't remember where they came from. It causes mental distress.*" Indeed, we observed some participants actually having some discussions in the Google Doc as they were editing it in real-time. We also noticed participants using copy-and-paste often to transfer messages from the chat log to Google Docs.

5.2 Study 2: Using Structured Summaries to Catch Up on Missed Conversations

In the second lab study, we conducted a between-subjects experiment to compare catching up on concluded conversations using Tilda summaries versus Google Docs notes or just the Slack chat log. To do this, we used the 12 artifacts created in the first study, including original chat logs as well as any accompanying Tilda summaries or Google Docs notes, and recruited new participants to look them over and answer comprehension questions about the discussions. We recruited 82 users (mean age 35, 28 female, 54 male) from Mechanical Turk, an online microtasking platform. Users were paid $3.25 per task and were required to have a 97% acceptance rate and 1,000 accepted tasks.

	None	Doc	Tilda
Time Spent (min)	11:12 (5:32)	12:12 (8:12)	12:55 (6:25)
Grade Received (out of 7)	5.79 (1.07)	5.89 (1.05)	5.88 (1.03)
Experience (5=Very Good)	3.14 (1.03)	3.59 (1.15)	3.83 (1.01)
Felt Rushed (5=Very High)	2.57 (1.02)	**3.04 (1.19)**	**2.08 (1.21)**

Table 2. Results from Study 2, where new users familiarized themselves with conversations from Study 1 using the artifacts created, broken down by the three conditions. We report the average and σ for time taken on the overall task, grade that users received from completing comprehension questions, self-reported experience on a post-task survey, and self-reported feelings of being rushed on a post-task survey. Statistically significant differences are in bold.

5.2.1 Experiment Design. There were 28 users for each of the three conditions of None, Doc, and Tilda, with half reviewing the Travel task from Study 1 and half the Story task. Before the study began, the first author used the task descriptions from Study 1 to create 7 comprehension questions for each task without looking at any artifacts, and then created a rubric for each of the 12 artifacts from Study 1. Users were given access to the corresponding Slack group and the Google Docs notes or Tilda summaries, which were located in a separate channel in the same group, if they existed. Users were not taught about Tilda except for an explanation that the hyperlinks in the Tilda summaries pointed to messages in the original chat log. At the same time, users were also given the 7 comprehension questions to answer in a survey form. There was no time limit for users nor instructions to spend a particular amount of time. After answering the questions, users filled out a separate survey about their experience, including NASA TLX questions about task load [33]. After the study, the first author graded each response out of 7 based on the rubric while blind to the condition. Two responses were discarded due to a score of 1/7 or lower, and three Tilda responses were discarded for self-reporting they were unaware of the hyperlinking feature despite the instructions.

5.2.2 Results. **Tilda users felt less rushed than Doc users**. We calculated how long users took by looking at time spent filling out the comprehension questions, with results in Table 2. While users overall spent the most time in Tilda and the least time in None, a one-way analysis of variance (ANOVA) test found that these differences were not significant (F=1.15, p=0.32), due to the high variation in time spent. From surveying users about their experience, users rated Tilda the highest, though these differences were not significant as well (F=2.41, p=0.09). Finally, we asked users about their task load, including the question of "*How hurried or rushed was the pace of the task?*" on a 5-pt Likert scale, where 5 is "very high". An ANOVA test yielded significant difference between the conditions (F=5.54, $p < 0.01$). Using a post hoc Tukey HSD test, we found that Tilda and Doc are significantly different at $p < 0.005$, with Tilda users feeling less rushed. In post-study comments, users described what they found hard, with one user saying *"I would have used Google Docs exclusively to answer the questions, but not all the information in Slack was there (and vice-versa)"* in the Doc condition, and another user saying *"The conversation seemed to be all over the place, there was no structure other than a group randomly chatting"* in the None condition.

People who used Tilda hyperlinks had lower load. Since we did not give a tutorial on Tilda, we were interested to see whether and how users in the Tilda condition would choose to use Tilda summaries. Only 4 of the users in the Tilda condition chose not to click the hyperlinks at all (No-Link), while 10 users used links often (Heavy-Link), according to self-reports. The remaining users said they used the links a few times. Heavy-Link users reported somewhat lower mental load (3 vs. 4.25, $p < 0.01$), feeling a great deal less rushed (1.3 vs. 4, $p < 0.001$), and feeling a great

	Active Users	Total Days Active	Channels with Tilda	All Chat Messages	Num Tilda Summaries	Total Num Tilda Items	Avg Tilda Items Per Summary	Avg Tilda Items Added Per User
Team A	3	6	6	277	15	53	4.5 (5.4)	20.3 (3.8)
Team B	6	9	4	870	40	220	5.5 (6.3)	35.7 (10.5)
Team C	4	5	6	478	22	101	5.8 (5.9)	31.3 (23.3)
Team D	3	8	9	373	36	51	1.5 (1.9)	17.7 (11.4)

Table 3. Overall usage statistics for the 4 Slack teams in the field study. Teams had variable usage of Slack as well as Tilda, with Team B as the most active overall.

deal less irritated and stressed (1.6 vs. 4, $p < 0.005$). Heavy-Link users also rated their experience as better, and spent less time overall yet still received a higher grade than No-Link users, though these differences were not significant. However, it is possible that our findings could be due to self-selection bias as opposed to solely due to using links in the summary to dive into the chat log.

5.3 Field Study

We conducted a week-long field study with four teams that use Slack to have work-related discussions. This field study allowed us to observe how Tilda is used in practice by real organizations.

We recruited teams by posting to social media and asking colleagues to distribute our call for participation. For the study, we aimed to recruit a diverse set of teams that work in different areas. We also sought teams that communicate in different ways, including teams that are remote and predominantly rely on Slack as well as teams that physically sit together. Users were compensated $100 to participate in the study and have Tilda installed on their team Slack account for a week ($20 per day). We told teams that we would store and analyze metadata about chat conversations and Tilda markup over the time that Tilda was installed but no textual content related to the chat.

Team A is a 3-person academic research team that sits together but uses Slack to keep track of ongoing research projects. **Team B** is a 6-person software engineering start-up that is fully remote and conducts all communication via Slack. **Team C** is a 4-person software engineering team that is partially co-located and uses Slack to troubleshoot and share resources. **Team D** is a 3-person fully remote team behind an online news blog that uses Slack to coordinate writing and publishing.

5.3.1 Study Design. Before the study, for 3 out of 4 teams, the first author was invited into the team's Slack organization to install Tilda and instruct members on how to use Tilda. In the case of Team D, the first author trained one individual in the team who then installed Tilda and taught the rest of the team on his own, due to the team's preference to keep their chat private. The training sessions were overall quicker than in Study 1, taking under 15 minutes using the same training materials, due to people's expertise with using Slack.

Participants were each asked to make a minimum of three notes or tags per day using Tilda, or 15 Tilda items in total over the course of the study. We chose to set the required activity low so we could see voluntary usage. We also gave no further requirements or suggestions so that users would be free to decide how to use Tilda. At the end of the week, 13 out of the 16 total number of users filled out a survey about their experience. At that point, we let the teams uninstall Tilda on their own, and three out of four teams continued to use it voluntarily for one to three more days during a second work week before we eventually took it down. We collected metadata on users' activity while Tilda was installed, including the kinds of Tilda items that users added.

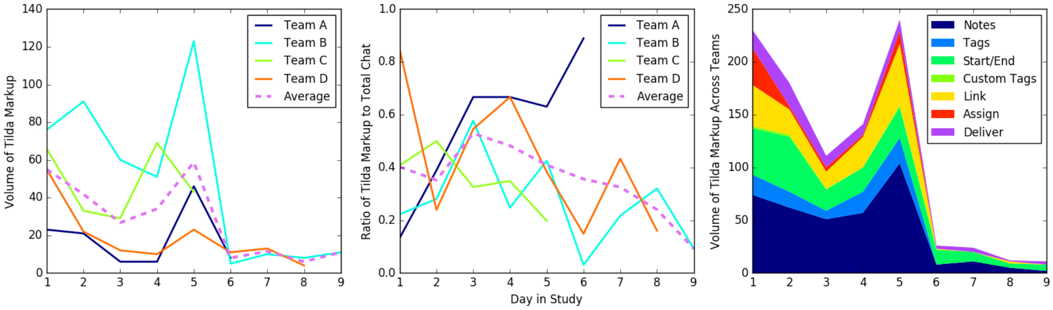


Fig. 5. We show activity over the course of the study. On the left is the total raw volume of markup added to chat using Tilda by each team each day, showing high variability between teams and across days. In the middle, the volume of markup is normalized by the total number of chat messages sent by the team for each day. Overall, we see sustained activity for a few days before a gradual tailing off. On the right is the raw volume of markup across all teams per day broken down by markup type. Notable is the preference for notes over tags and significant use of linking.

5.3.2 Results. **Teams were active in using Tilda to mark up their chat**. We report overall statistics in Table 3. As can be seen, there was variable usage of Tilda across as well as within the teams, that generally corresponded to how active they were in Slack as a whole. Almost all users went over the minimum number of items on a daily basis, and as mentioned, several teams used Tilda for longer than the required 5 days. The left side of Figure 5 shows the usage of Tilda over the course of the study, counting all possible markup that could be added to chat with Tilda. We remove days where there was no activity since some of the teams did not work on weekends. Different teams joined the study on different days, so the days of the week are not aligned. However, it was interesting that the peak activity was on different days for different teams. For team A and B, the peak was on the fifth day while it was the fourth day for Team C and first day for Team D. However, these fluctuations are perhaps a reflection of just overall activity in chat on those days. In the middle of Figure 5, we show the volume of Tilda markup normalized by the total number of chat messages posted in the team for each day. While these also fluctuate quite a bit, we can see that the average ratio for the four teams stays between 0.4 to 0.6 for around 6 days before decreasing. While we did not conduct a longitudinal study, the overall decrease in activity as the study concluded suggests that the tool will need to consider how to design for usage over longer periods of time. We present possible options further in the discussion.

As seen on the right side of Figure 5, notes overall saw higher usage than tags, thanks to Teams B and D, who favored notes almost exclusively, while Teams A and C were evenly split between the two. Across the board, custom tags were rarely used. One reason for this may be because the use of channels in Slack is already a decent separator of topics. Finally, there was surprisingly considerable usage of the linking feature as well as usage of the assignment feature earlier on in the study. From the post-study surveys, we asked users about their favorite feature, with several users mentioning the linking feature (3), Question and Answer items (3), Action items (2), and the automatic summary logging (3). In terms of missing or faulty features in Tilda, many users complained about how Tilda would take up too much real estate in the channel by posting (6), while 1 user wanted the ability to resolve Action items, 1 user wanted to link to multiple items, and 2 users wanted the ability to export the summaries to a document or their Trello board.

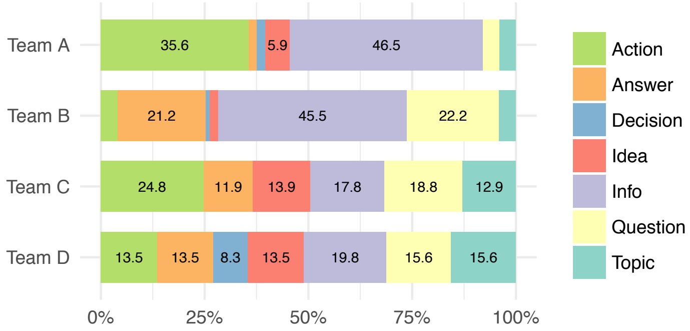


Fig. 6. Percentage of Tilda items that were of each discourse act type for each team in the field study.

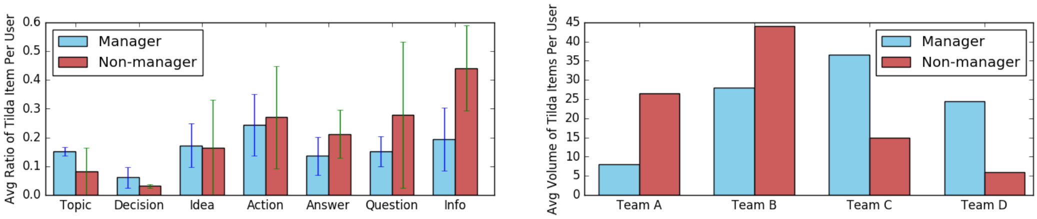


Fig. 7. On the left, the ratio of Tilda items that were of each discourse act type for each user, averaged and grouped by managers and non-managers, from the field study. On the right, the volume of Tilda items added per user in each team, averaged and grouped by managers and non-managers.

Teams and individuals personalized their use of Tilda to suit their needs. On average, except for Team D, users added around 5 items to each summary of a conversation, though this had high variance. Looking at the breakdown of items into their types, we can see in Figure 6 that Info was used frequently across all the groups, while Questions and Answers were used heavily by both software engineering teams. Other types were used more infrequently, especially Decision, possibly because users found the Action type more apt for the conclusion of conversations. We also asked users to rate each category for their usefulness in the survey; results were similar to Figure 6, with Action and Question rated the highest on average (4.3/5) and Decision and Topic rated the lowest (3.5/5). Overall, many users stated that they found the provided discourse acts expressive for all their notetaking needs. One participant said "*The clear variety of different add actions was very useful; I didn't feel limited, like I had to shoehorn my types of choices into one box.*"

In Figure 7, we break down user behavior by managers versus non-managers, with management role self-reported from surveys. On the left-side figure, we break down the average ratio for different Tilda discourse act types for managers and non-managers. As can be seen, managers had a higher ratio of agenda-setting types such as Topic and Decision, while non-managers had a higher ratio for types more relevant to implementation details such as Info or Question and Answer. We also look at average volume of Tilda items left by managers versus non-managers, finding Team A and B had more non-manager participation, while C and D had less.

One interesting aspect of Tilda was how the norms of the team adjusted around the introduction of the tool. For instance, one participant described how adding Tilda improved the nature of conversation: "*I also noticed that Tilda changed our conversation flow (for the better). Since we were working with a finite set of tags...[the] messages...served a specific purpose that fell into one of the tags...questions were less likely to be lost...the action tag was a perfect way to remind us to take what* *we were chatting about and turn it into a tangible takeaway.*" While we did not collect empirical data on this as we did not have access to prior activity in the teams, future work could analyze how the additional structure that Tilda allows might encourage certain types of discourse. Norms may also need to be set around the use of Tilda. One manager of Team B complained that the team left *too* many notes, leaving him to review unimportant information: "*I believe the summaries were useful and made it easier for me to review the notes as a manager. However, it hinged on the team being disciplined to only include important notes and this wasn't always the case...Overall I think if we got into the habit of using it effectively, it seems like Tilda would be big help to our workflow.*" Perhaps some of the drop-off in proportional usage of Tilda by Team B starting from Day 3 was a result of decisions made, either implicitly or explicitly, to mark fewer items.

Tilda was effective for catching up and looking back. We were not able to capture data on reading of summaries due to limitations of the Slack API and so instead asked participants to self-report. Eight users said they used the summaries to catch up on missed conversations and rated the experience of catching up an average of 4.4 out of 5. One user said about catching up, "*Before Tilda I would try to scroll...This was very tedious...With Tilda this process was much smoother. I would usually check our Tilda responses channel and skim through the summaries to see what I missed. If a topic seemed interesting, I would expand it all and read through everything. If I was uninterested in the topic I would just move on.*" Participants that were in the discussion also mentioned their motivation of marking up chat to keep absent team members up-to-date. One person who was on the partially remote Team C said "*I work with a remote user a lot, and it was helpful to document what he needed to work on and clarify things he didn't understand.*"

Eleven people used the summaries to look back at old conversations they were in and rated their experience an average of 4.2 out of 5. Participants mentioned that a motivation for marking up chat was for themselves in order to keep track of things they needed to do or remember. One user remarked on using Tilda to look back through old conversations: "*Without Tilda - Scroll through or search for a keyword and try to find the message I think I remembered. If I can't remember or misremember something it can be frustrating trying to find it. With Tilda - Mark it and simply find the Summary either in the channel itself or the channel we had our responses in. Much less frustrating.*"

Ten users said they chose to set up a team-wide channel dedicated to summaries from the other channels, while 2 users chose to follow personalized summaries via their direct message with Tilda.

Tilda was used for structuring conversation and tracking important information. Some of the teams already had some mechanism for tracking longer term information and tasks, such as a Trello board or various Google Docs files. One person described how they liked having information tracked in one place, saying "*Tilda gives us a somewhat better way to track information. It's useful to have everything all in one place...instead spread out like in Trello or Google Doc. Trello can get pretty messy easily...And I find our Google Drive directory hard to navigate...*" However, some team members were used to the existing workflow they had with other systems and wished there was a way to sync them. Another participant thought of a separate site where summaries could be archived and searchable: "*...I really think that summaries should be exported/exportable to a different interface...for example to send to people off of Slack or to archive as a piece of important info...summary search...could be implemented on this page...For example, it would be nice if all action items could be pulled out to a running todo list organized by the topic of the conversation they came from (and linked of course).*"

For the teams that did not have mechanisms for keeping everyone on board and relied only on Slack, some members were excited about the additional structure that Tilda encouraged: "*We really didn't have a good system...Tilda made it muuuuch easier for us to fill someone in on something that happened...Overall I think Tilda greatly improved team communication over the week we used it. Conversations had better structure, team members were better kept up to date, and we actually had a way to save...results of our conversations for future use.*"

6 DISCUSSION

Tilda markup adds structure to group chat conversations that can be beneficial to chat participants. First, in contrast to traditional notetaking tools like documents, Tilda's light-weight markup allows notetaking without forcing users to leave the conversation. As was suggested in Study 1, this approach offers a promising design pattern for making collaborative notetaking easier compared to alternatives. Study 1 also provides evidence that Tilda does introduce some mental load to users, but this could be a worthy trade-off for the organizational benefits it provides when it comes to discussing complicated things. Such benefits were echoed by participants in the field study who used chat extensively for work. In Study 1 and the field study, we also noticed high variability in how different users take notes, both in terms of note volume and their manner of notetaking. Similarly, we saw variability in Study 1 in how groups used the Google Doc to take notes, with different quality of outcomes. Tilda is more structured of a tool but still leaves room for variation, such as the number of items that make up a single summary. These observations suggest that, like good notetaking practices for documents, there may be some strategies to encourage better notetaking in Tilda. For instance, future iterations of Tilda could suggest closing a summary and starting a new one if many notes have been added, or asking users to pick the most important notes from a summary to create a higher level summary, in a recursive fashion [90].

When it comes to the output side, the field study echoed the results from the needfinding interviews, showing that catching up on or looking back over chat is a common task, and that it was improved with Tilda summaries. In Study 2 we found evidence that the links between conversations and notes were helpful for enabling newcomers to get up to speed more efficiently. As we observed in our field study, this structure was useful in providing additional context to conversations, allowing teams to organize and collaborate more successfully. Additionally, in Study 2, users felt less rushed using Tilda to catch up or look back over a separate document. This may be because the Tilda summaries are an alternative presentation or entry point for navigation into the original chat log and add no *new* content. In contrast, a document contributes new text and also leads to information spread out between two places. Given the use of Tilda summaries as a navigational tool, this suggests that an alternative presentation of Tilda summaries could have them overlaid or beside the original chat instead of posted to a separate channel or direct message.

6.1 Who Annotates and Why?

When it came to intrinsic motivation for users, we saw in the field study users mentioning that they added notes and tags in order to keep track of their own tasks and requests in the day, which then became helpful for other users. We also had examples of working with remote users in different time-zones where adding markup was helpful with asynchronous chatting. However, we did observe the importance of setting shared groups norms towards adoption of Tilda in our studies. A similar need for groups to get on the same page was expressed about group chat in general in our interviews, where some group chat users complained that inconsistent or non-reciprocal usage of certain features like threading sometimes led to even greater confusion. Even in the field study, we saw some people take many notes while others took only a few, though this could be because they did not use Slack or were not core members of the team. For Tilda to be successful, norms may need to be set by team leaders to motivate usage long-term.

In this work, our evaluations mainly focused on small groups of people conversing, and we did not explore how size of a team can alter the way Tilda is used. In a larger group, with hundreds or thousands of members, issues like social loafing, fear of participation, or contested norms [87] may be exacerbated. In such cases, an alternative design to Tilda's collaborative notetaking, reflected in earlier meeting bots like Debian MeetBot, could allow for the designation of an owner role for each meeting, who is in charge of adding notes, much like notetakers in live meetings. In some situations, such as in more ad hoc teams like Study 1 with no defined leader, this clear delineation of roles might be preferable. In future iterations of Tilda, the bot could also encourage participation by sending targeted proactive prompts to individuals to solicit notetaking.

Due to our decision to make Tilda a chatbot instead of an alternative chat system, we were constrained in the ways we could present summaries or messages to the group. This became an issue in the field study where the biggest complaint was about Tilda messages to the group taking up too much screen real estate. Due to these evident user experience issues, we chose not to pursue a longitudinal field study with the current implementation of Tilda. Additional deployments of Tilda could empirically examine alternative types of markup and summary presentations using short field studies or lab studies. In the future, a longer study on a new chat system where we have full control over presentation could allow us to further examine how norms and motivations around chat markup develop over time.

6.2 Towards Automatic Summarization of Chat

This work presents a first step towards a human-centered conceptualization of the goal of automatic chat summarization. In interviews, we collected empirical data around what kinds of summaries are desirable to chat readers, finding that structured summaries highlighting discourse acts were preferred over conventional presentations such as purely abstractive or extractive summaries. This result allows us to consider that the difficult problem of automatically summarizing chat conversation could potentially be tackled by breaking the problem down. Machine-learned models could augment the work that Tilda users do, such as by suggesting actions or simply performing some of them. Standard supervised machine learning techniques could be brought to bear on intermediate automatable problems include delineating separate conversations in a stream, labeling the discourse act of a message [88], finding messages that are candidates for tagging, linking messages to prior ones, and populating abstractive topic sentences or auto-tagging topics. These tasks have the benefit of reducing the learning curve and effort involved in using Tilda.

To build such models however, one must collect training data; luckily, Tilda too provides a path for fulfilling this role. More broadly, collecting rich training data can be a significant hurdle in developing models towards discussion summarization. In early pilots of our studies we conducted towards paid crowd annotation of public chat logs, we found that it was difficult for workers to make sense of a chat conversation they were not a part of. And as we saw in interviews, even if people are members of a group, it still takes effort to parse the back-and-forth when looking back over chat. Tilda manages this problem by making it possible to mark up chat conversations while taking part in them, when the conversational context is still fresh in their minds. In addition, we provide evidence that the Tilda system has value and direct benefits to users even in its current implementation as a primarily manual annotation tool.

6.3 Integration with Knowledge Management Tools and Email

Integration with outside knowledge management tools, such as wikis or documents, came up as feedback in both Study 1 and our field study. One could imagine Tilda chatting with existing bots or integrating with APIs to post to task management tools like Trello, Q&A sites like Stack Overflow, calendars [28], and code repositories. Likewise, one could imagine a website where additional organization of the summaries themselves could happen. Such an interface could be useful for newcomers looking to quickly make sense of the prior discussion in the team. Additionally, several interviewees described issues with triaging conversations that spill into both email and chat. Summaries could be inserted as embedded items in platforms such as email or forums that are more asynchronous. In all these cases, automatic links back to the original discussion in chat as well as automatic updating of content across links could manage the issue of information lost within multiple potential locations.

While Tilda bridges synchronous and asynchronous *access* of conversation, there are still questions about how to facilitate *partaking* in conversation for those who missed out. For instance, one person in our field study wanted a way to reopen a conversation that they had missed. This could be done by posting the summary to the relevant channel to remind users of the context and then writing a comment underneath. Any ensuing notes from the new conversation could get added to the original summary.

7 FUTURE WORK AND LIMITATIONS

We have released Tilda as a public tool[3] and open-sourced the code[4], and aim to collect training data using Tilda towards automatic summarization tasks. Another area where we believe Tilda would be useful is for notetaking and summarization of video, audio, and in-person meetings, with the help of speech-to-text technology for transcription. Such a system could even work in concert with systems for crowdsourcing real-time transcriptions [49]. For instance, participants could collaboratively fix issues with transcription and highlight, tag, or vote on aspects of the discussion while conversing. While our work focuses on catching up and gaining an overview of a large chat log, we also uncovered issues that interviewees had with searching for particular items within chat. Future work could consider whether scrolling and other forms of orienteering behavior while searching [75] could be aided by signals left by Tilda. Currently, Tilda is a Slack-only tool; however because it was implemented using Microsoft's Bot Framework, it could be extended to other chat platforms that support bot integration with minimal additional development. The Slack features that we use, including emoji reactions, slash commands, and inline emojis, have uneven but growing support across other major chat platforms. For instance, emoji reactions are now supported in Facebook Messenger. Additionally, almost all platforms now support inline emojis, while slash commands could be simulated using hashtags.

8 CONCLUSION

In this work, we studied how users of group chat make sense of chat conversations when they need to catch up or look back, and we investigated how marking up chat messages to provide additional structure could help. From presenting 15 interviewees with different representations of chat information, we determined the importance of structure and discourse acts towards quickly understanding the gist and relevance of a chat conversation. From these findings, we developed Tilda, a tool for participants to mark up an ongoing chat conversation with various signals. The tool allows users to catch up on chat conversations through structured and linked summaries automatically created from users' notes and tags. From lab studies and field studies, we find that Tilda is effective for both taking notes and catching up on conversation.

ACKNOWLEDGMENTS

Thank you to our interviewees and all the study participants for their time and feedback, members of Microsoft FUSE Labs for their feedback, and developers of Microsoft Bot Framework for their assistance. Thank you to Andrés Monroy-Hernández for feedback.

REFERENCES

[1] [n. d.]. Focus on the important things with Highlights in Slack. *Slack*, 14 June 2017. Available: https://slackhq.com/focus-on-the-important-things-with-highlights-in-slack-5e30024502cd [Last accessed: 2017-09-15]. ([n. d.]).

[3] tildachat.com

[4] https://github.com/Microsoft/tilda

[2] [n. d.]. Google Hangouts Chat. https://chat.google.com. ([n. d.]).

[3] [n. d.]. HipChat. https://www.hipchat.com/. ([n. d.]).

[4] [n. d.]. MeetBot. *Debian*, 8 Janary 2017. Available: https://wiki.debian.org/MeetBot [Last accessed: 2017-09-08]. ([n. d.]).

[5] [n. d.]. Microsoft Teams. https://teams.microsoft.com. ([n. d.]).

[6] [n. d.]. Slack. https://slack.com/. ([n. d.]).

[7] [n. d.]. Zakim. *W3C*, 7 September 2015. Available: https://www.w3.org/2001/12/zakim-irc-bot.html [Last accessed: 2018-04-18]. ([n. d.]).

[8] Anton Bogdanovych, Helmut Berger, Simeon Simoff, and Carles Sierra. 2005. Narrowing the gap between humans and agents in e-commerce: 3D electronic institutions. In *EC-Web*, Vol. 5. Springer, 128–137.

[9] David Brumley. [n. d.]. Tracking hackers on IRC. *Usenix*, November 1999. Available: https://www.usenix.org/legacy/publications/login/1999-11/features/hackers.html [Last accessed: 2017-09-14]. ([n. d.]).

[10] Ann Frances Cameron and Jane Webster. 2005. Unintended consequences of emerging communication technologies: Instant messaging in the workplace. *Computers in Human behavior* 21, 1 (2005), 85–103.

[11] Joyce Chai, Veronika Horvath, Nicolas Nicolov, Margo Stys, Nanda Kambhatla, Wlodek Zadrozny, and Prem Melville. 2002. Natural language assistant: A dialog system for online product recommendation. *AI Magazine* 23, 2 (2002), 63.

[12] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* Sage.

[13] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1999–2008. http://dx.doi.org/10.1145/2470654.2466265

[14] Patrick Chiu, John Boreczky, Andreas Girgensohn, and Don Kimber. 2001. LiteMinutes: an Internet-based system for multimedia meeting minutes. In *Proceedings of the 10th international conference on World Wide Web.* ACM, 140–149.

[15] Patrick Chiu, Ashutosh Kapuskar, Sarah Reitmeier, and Lynn Wilcox. 1999. NoteLook: Taking notes in meetings with digital video and ink. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1).* ACM, 149–158.

[16] Mercia Coetzee, Annette Wilkinson, and Daleen Krige. 2016. Mapping the social media landscape: a profile of tools, applications and key features. (2016).

[17] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 467–474.

[18] Josh Constine. [n. d.]. Facebook Messenger hits 1.2 billion monthly users, up from 1B in July. *TechCrunch*, 12 April 2017. Available: https://techcrunch.com/2017/04/12/messenger/ [Last accessed: 2017-09-14]. ([n. d.]).

[19] Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. In *ACL-04 Workshop: Text Summarization Branches Out.* 43–50.

[20] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 2382–2393.

[21] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. 2000. Instant messaging and interruption: Influence of task type on performance. In *OZCHI 2000 conference proceedings*, Vol. 356. 361–367.

[22] Richard C Davis, James A Landay, Victor Chen, Jonathan Huang, Rebecca B Lee, Frances C Li, James Lin, Charles B Morrey III, Ben Schleimer, Morgan N Price, et al. 1999. NotePals: Lightweight note sharing by the group, for the group. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* ACM, 338–345.

[23] Judith Donath, Karrie Karahalios, and Fernanda Viegas. 1999. Visualizing conversation. *Journal of Computer-Mediated Communication* 4, 4 (1999), 0–0.

[24] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1175–1184.

[25] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael Bernstein. 2017. Iris: A Conversational Agent for Complex Tasks. *arXiv preprint arXiv:1707.05015* (2017).

[26] Daniel C Feldman. 1984. The development and enforcement of group norms. *Academy of management review* 9, 1 (1984), 47–53.

[27] Eric N Forsythand and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on.* IEEE, 19–26.

[28] Siwei Fu, Jian Zhao, Hao Fei Cheng, Haiyi Zhu, and Jennifer Marlow. 2018. T-Cal: Understanding Team Conversational Data with Calendar-based Visualization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 500.

[29] R Kelly Garrett and James N Danziger. 2007. IM= Interruption management? Instant messaging and disruption in the workplace. *Journal of Computer-Mediated Communication* 13, 1 (2007), 23–42.

[30] Werner Geyer, Heather Richter, and Gregory D Abowd. 2005. Towards a smarter meeting recordâĂŤcapture and access of meetings revisited. *Multimedia Tools and Applications* 27, 3 (2005), 393–410.

[31] Antonietta Grasso and Gregorio Convertino. 2012. Collective intelligence in organizations: Tools and studies. *Computer Supported Cooperative Work (CSCW)* 21, 4-5 (2012), 357–369.

[32] 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*. ACM, 1–10.

[33] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

[34] James Hartley. 1983. Note-taking research: Resetting the scoreboard. *Bulletin of the British Psychological Society* (1983).

[35] James D Herbsleb, David L Atkins, David G Boyer, Mark Handel, and Thomas A Finholt. 2002. Introducing instant messaging and chat in the workplace. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 171–178.

[36] Enamul Hoque and Giuseppe Carenini. 2014. Convis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library, 221–230.

[37] Yifeng Hu, Jacqueline Fowler Wood, Vivian Smith, and Nalova Westbrook. 2004. Friendships through IM: Examining the relationship between instant messaging and intimacy. *Journal of Computer-Mediated Communication* 10, 1 (2004), 00–00.

[38] Samuel Hulick. [n. d.]. I used to be obsessed with Slack but now I'm dropping it completely âĂŤ here's why. *Business Insider*, 1 March 2016. Available: http://www.businessinsider.com/i-used-to-be-obsessed-with-slack-but-now-im-dropping-it-completely-heres-why-2016-3 [Last accessed: 2017-09-09]. ([n. d.]).

[39] Shamsi T Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 677–686.

[40] Samuel P. Jacobs. [n. d.]. How E-Mail Killer Slack Will Change the Future of Work. *Time*, 28 October 2015. Available: http://time.com/4092354/how-e-mail-killer-slack-will-change-the-future-of-work/ [Last accessed: 2017-09-08]. ([n. d.]).

[41] Adrianne Jeffries. [n. d.]. We're Taking a Break from Slack. Here's Why. *Motherboard*, 16 May 2016. Available: https://motherboard.vice.com/en_us/article/aekk85/were-taking-a-break-from-slack-heres-why [Last accessed: 2017-09-09]. ([n. d.]).

[42] Matthew Kam, Jingtao Wang, Alastair Iles, Eric Tse, Jane Chiu, Daniel Glaser, Orna Tarshish, and John Canny. 2005. Livenotes: a system for cooperative and augmented note-taking in lectures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 531–540.

[43] Bernard Kerr. 2003. Thread arcs: An email thread visualization. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*. IEEE, 211–218.

[44] Kenneth A Kiewra. 1985. Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist* 20, 1 (1985), 23–32.

[45] Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 192–202.

[46] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[47] Steve Kovach. [n. d.]. I figured out a way to kill work email once and for all. *Venture Beat*, 19 February 2015. Available: https://venturebeat.com/2016/10/20/slack-passes-4-million-daily-users-and-1-25-million-paying-users/ [Last accessed: 2017-09-08]. ([n. d.]).

[48] James A. Landay and Richard C. Davis. 1999. Making sharing pervasive: Ubiquitous computing for shared note taking. *IBM Systems Journal* 38, 4 (1999), 531–550.

[49] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 23–34.

[50] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P. Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*. ACM, 151–162. http://dx.doi.org/10.1145/2501988.2502057

[51] Minha Lee, Lily Frank, Femke Beute, Yvonne de Kort, and Wijnand Ijsselsteijn. 2017. Bots Mind the Social-technical Gap. In *Proceedings of 15th European Conference on Computer-Supported Cooperative Work-Exploratory Papers*. European Society for Socially Embedded Technologies (EUSSET).

[52] Gary S Lynn, Richard R Reilly, and Ali E Akgun. 2000. Knowledge management in new product teams: practices and outcomes. *IEEE transactions on Engineering Management* 47, 2 (2000), 221–231.

[53] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective Sensemaking in Online Health Forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3217–3226. https://doi.org/10.1145/2702123.2702566

[54] Matthew K Miller, John C Tang, Gina Venolia, Gerard Wilkinson, and Kori Inkpen. 2017. Conversational Chat Circles: Being All Here Without Having to Hear It All. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2394–2404.

[55] Xiangming Mu. 2010. Towards effective video annotation: An approach to automatically link notes with video content. *Computers & Education* 55, 4 (2010), 1752–1763.

[56] Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 773–782.

[57] Kevin K Nam and Mark S Ackerman. 2007. Arkose: reusing informal information from online discussions. In *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM, 137–146.

[58] Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval* 5, 2–3 (2011), 103–233.

[59] Jarkko Oikarinen and Darren Reed. 1993. Internet Relay Chat Protocol. *IETF*, 1993. Available: https://www.ietf.org/rfc/rfc1459.txt [Last accessed: 2017-09-08]. (1993).

[60] Stephen T Peverly, Vivek Ramaswamy, Cindy Brown, James Sumowski, Moona Alidoost, and Joanna Garner. 2007. What predicts skill in lecture note taking? *Journal of Educational Psychology* 99, 1 (2007), 167.

[61] Annie Piolat, Thierry Olive, and Ronald T Kellogg. 2005. Cognitive effort during note taking. *Applied Cognitive Psychology* 19, 3 (2005), 291–312.

[62] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. 2–4.

[63] Anabel Quan-Haase, Joseph Cothrel, and Barry Wellman. 2005. Instant messaging for collaboration: A case study of a high-tech firm. *Journal of Computer-Mediated Communication* 10, 4 (2005), 00–00.

[64] Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, 105–108.

[65] Elizabeth Reid. 1991. Electropolis: Communication and community on internet relay chat. (1991).

[66] Evan F Risko, Tom Foulsham, Shane Dawson, and Alan Kingstone. 2013. The collaborative lecture annotation system (CLAS): A new TOOL for distributed learning. *IEEE Transactions on Learning Technologies* 6, 1 (2013), 4–13.

[67] Maya Sappelli, Gabriella Pasi, Suzan Verberne, Maaike de Boer, and Wessel Kraaij. 2016. Assessing e-mail intent and tasks in e-mail messages. *Information Sciences* 358 (2016), 1–17.

[68] John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.

[69] Bayan Abu Shawar and Eric Atwell. 2007. Chatbots: are they really useful?. In *LDV Forum*, Vol. 22. 29–49.

[70] Bayan Abu Shawar and Eric Steven Atwell. 2005. Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics* 10, 4 (2005), 489–516.

[71] Emad Shihab, Zhen Ming Jiang, and Ahmed E Hassan. 2009. On the use of Internet Relay Chat (IRC) meetings by developers of the GNOME GTK+ project. In *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*. IEEE, 107–110.

[72] Lokesh Shrestha and Kathleen McKeown. 2004. Detection of question-answer pairs in email conversations. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 889.

[73] Marc Smith, Jonathan J Cadiz, and Byron Burkhalter. 2000. Conversation trees and threaded chats. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 97–105.

[74] Jaime Teevan. 2016. The future of microwork. *XRDS: Crossroads* 23, 2 (2016), 26–29. https://doi.org/10.1145/3019600

[75] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 415–422.

[76] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2657–2668. https://doi.org/10.1145/2858036.2858108

[77] Jaime Teevan, Daniel J. Liebling, and Walter S. Lasecki. 2014. Selfsourcing personal tasks. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2527–2532. http://dx.doi.org/10.1145/2559206.2581181

[78] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-mediated Task Management. *Proceedings of the SIGCHI conference on Human factors in computing systems*.

[79] Fernanda B Viégas, Scott Golder, and Judith Donath. 2006. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 979–988.

[80] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[81] Raelene Wilding. 2006. âĂŸVirtualâĂŹintimacies? Families communicating across transnational contexts. *Global networks* 6, 2 (2006), 125–142.

[82] David R Woolley. 1994. PLATO: The emergence of online community. (1994).

[83] Lori Wright. [n. d.]. Expand your collaboration with guest access in Microsoft Teams. *Microsoft Office Blog*, 11 September 2017. Available: https://blogs.office.com/en-us/2017/09/11/expand-your-collaboration-with-guest-access-in-microsoft-teams/ [Last accessed: 2017-09-14]. ([n. d.]).

[84] Liu Yang, Susan T Dumais, Paul N Benne, and Ahmed Hassan Awadallah. 2017. Characterizing and Predicting Enterprise Email Reply Behavior. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2017, Tokyo, Japan.*

[85] Alexander Seeshing Yeung, Putai Jin, and John Sweller. 1998. Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary educational psychology* 23, 1 (1998), 1–21.

[86] David M Zajic, Bonnie J Dorr, and Jimmy Lin. 2008. Single-document and multi-document summarization techniques for email threads using sentence compression. *Information Processing & Management* 44, 4 (2008), 1600–1610.

[87] Amy X. Zhang, Mark S. Ackerman, and David R. Karger. 2015. Mailing Lists: Why Are They Still Here, What's Wrong With Them, and How Can We Fix Them?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4009–4018. https://doi.org/10.1145/2702123.2702194

[88] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. In *ICWSM*.

[89] Amy X Zhang, Michele Igo, Marc Facciotti, and David Karger. 2017. Using Student Annotated Hashtags and Emojis to Collect Nuanced Affective States. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM, 319–322.

[90] Amy X Zhang, Lea Verou, and David R Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization.. In *CSCW*. 2082–2096.

[91] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, 1883–1892.

Received April 2018; revised July 2018; accepted September 2018