Chat Disentanglement: Data for New Domains and Methods for More Accurate Annotation

Sai R. Gouravajhala[†], **Andrew M. Vernier**[†], **Yiming Shi**[†], **Zihan Li**[†]
Mark S. Ackerman[†], **Jonathan K. Kummerfeld**[†‡*]
[†]Computer Science & Engineering, University of Michigan, Ann Arbor
[‡]School of Computer Science, University of Sydney

Abstract

Conversation disentanglement is the task of taking a log of intertwined conversations from a shared channel and breaking the log into individual conversations. The standard datasets for disentanglement are in a single domain and were annotated by linguistics experts with careful training for the task. In this paper, we introduce the first multi-domain dataset and a study of annotation by people without linguistics expertise or extensive training. We experiment with several variations in interfaces, conducting user studies with domain experts and crowd workers. We also test a hypothesis from prior work that link-based annotation is more accurate, finding that it actually has comparable accuracy to set-based annotation. Our new dataset will support the development of more useful systems for this task, and our experimental findings suggest that users are capable of improving the usefulness of these systems by accurately annotating their own data.

1 Introduction

Rapid synchronous chat involving a large group often leads to overlapping conversations. The challenge of disentangling these conversations has been studied for over a decade, but the main datasets are expert annotated and based on discussion of Linux (Elsner and Charniak, 2008) and Ubuntu (Kummerfeld et al., 2019). Recent work has considered scripts from movies (Chang et al., 2023), but there is still the need for data from additional sources to measure the generalizability of methods.

A range of methods have been proposed to avoid expensive expert annotation in NLP, e.g., crowd work (Snow et al., 2008), games with a purpose (Jurgens and Navigli, 2014) and user feedback (Iyer et al., 2017). Various annotation methods have been used for disentanglement, but all focused on experts and only one study has compared annotation tools (Cerezo et al., 2021).

This work takes two key steps to expand this task to new domains: (1) we created a new, multi-domain, gold-standard dataset, and (2) we explored annotation methods to see if domain experts and crowd workers can do the task.

Our dataset includes several important variations not seen in existing datasets: (a) new types of conversations (e.g., meetings), (b) new types of user relationships (e.g., business-customer), and (c) a range of Internet Relay Chat (IRC) networks. We annotated 600 messages from each channel, which is enough to evaluate out-of-domain ccuracy.

It is impossible to collect expert labels for every domain. However, if we can develop the right tools, owners and users of channels may be able to improve models by annotating some of their own data. We conducted a user study with domain experts and crowd workers, exploring two types of variation in user interfaces: (1) whether annotators receive automatic guidance, and (2) what structure is annotated. Prior work has speculated that link annotation[1] is more accurate than set annotation[2] (Elsner and Charniak, 2010), but our work is the first controlled comparison.

We found that domain experts can effectively annotate data, and improve with automatic guidance. Crowd workers struggled with the task, doing worse than an automatic model. Set-based and link-based annotation are actually comparable in accuracy. We recommend link annotation as it provides the internal structure of conversations.

The dataset we release[3] will support the development of more generalizable models, and our findings show how to help domain experts annotate effectively. Together, these results will enable progress on this challenge in new domains, making conversations easier to follow for everyone online.

* jonathan.kummerfeld@sydney.edu.au

[1] Labeling reply-to relations between pairs of messages, then each connected graph of messages is a conversation.

[2] Putting messages into groups, where each group is a conversation.

[3] https://www.jkk.name/irc-disentanglement/

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 112–117
November 29 - December 1, 2023

Channel	Network	Purpose	Msg / Hr	Users / Hr	Tok / Msg	κ
Mediawiki	Wikimedia	Technical support regarding mediawiki software.	71	4.1	10	0.78
Rust	Mozilla	Help related to the Rust programming language.	33	8.0	12	0.80
Stripe	Freenode	Customer support for the payments processing service.	76	6.6	16	0.81
Ubuntu Meeting	Ubuntu	Developer meetings.	371	9.9	8	0.71
Ubuntu	Ubuntu	Technical support for users of the operating system.	395	32	10	0.72

Table 1: Expert annotator agreement (κ) and properties of the four channels we annotated and the Ubuntu channel used in Kummerfeld et al. (2019). The channels span multiple topics (programming languages, customer support, web applications) and conversation styles (question-answer, meetings).

2 Related Work

All prior annotation for conversation disentanglement has been done by trained experts, like many tasks in NLP (Ide and Pustejovsky, 2017). Early work on the task asked annotators to form sets of messages (Elsner and Charniak, 2008, 2010), but they speculated that annotators may be more consistent at annotating reply-to links. Subsequent work took the link approach (Riou et al., 2015; Mehri and Carenini, 2017; Kummerfeld et al., 2019; Cerezo et al., 2021). This work is the first controlled comparison of the two. Cerezo et al. (2021) compared a command-line UI and GUI, finding that annotators preferred the GUI, but accuracy was the same, and using the GUI was slower. Our study complements theirs by considering: (1) variation in who annotators are, (2) variation in the form of annotation, and (3) guidance.

Crowd work can be cheaper and more scalable than expert annotation (Snow et al., 2008). Effective crowd annotation user interfaces and workflows have been developed for a range of tasks (e.g., Dumitrache et al., 2018; Finin et al., 2010; Larson et al., 2020), but there has been no prior work for disentanglement.

Guiding annotators using an automatic system has improved speed for other tasks (Marcus et al., 1993; Chiou et al., 2001). Recent work has applied similar ideas to crowd work (Gormley et al., 2010; Ramírez et al., 2019). We apply this idea to conversation disentanglement for the first time.

3 Data in New Domains

When multiple synchronous conversations are happening in the same channel they can be difficult to understand.[4] Conversation disentanglement is the task of identifying separate conversations, to make them understandable and useful.

Channel	F	1-1
Mediawiki	46	90
Rust	60	91
Stripe	83	94
Ubuntu Meeting	22	73
Ubuntu	43	82

Table 2: Model accuracy on conversations for each of the channels.

There are hundreds of active Internet Relay Chat (IRC)[5] channels, but only two have disentanglement annotations: #Ubuntu (Kummerfeld et al., 2019) and #Linux (Elsner and Charniak, 2008, 2010). To create a realistic out-of-domain setting, we annotated data from four diverse channels, described in Table 1. We chose channels that: (1) have public logs, (2) have various topics and conversation styles, and (3) are from different IRC networks, which may exhibit different conventions.

For each channel, we used three random samples, each 1,200 messages long (200 to annotate, 1,000 for context). This leads to a total of 2,400 annotated messages and a further 12,000 context messages. Our data is in the same format as Kummerfeld et al. (2019) to enable easy evaluation. This is the first work to annotate multi-domain data, enabling out-of-domain evaluation.

Expert Annotation To make a gold-standard reference, two of the authors labeled each file, then one of the authors adjudicated disagreements. To match the annotations of Kummerfeld et al. (2019) as closely as possible, we labeled reply-to links using their tool, SLATE (Kummerfeld, 2019), and the same annotation guidelines. Conversations are the connected components in the reply-to graph.

Table 1 shows agreement scores for reply-to

[4] Some services, e.g., Slack, WebEx, and Microsoft Teams, have the ability to split a conversation starting at a message, but that only solves the problem if the split is created as soon as a new topic is started and the conversation remains on topic.

[5] IRC is a protocol for synchronous chat in use since 1988.

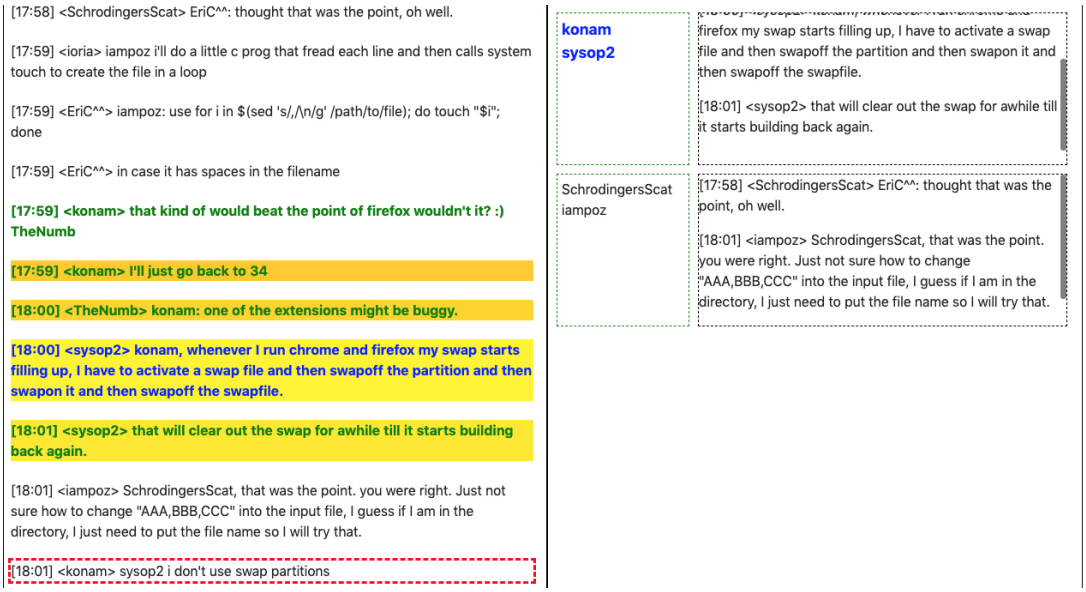

Figure 1: Part of the user interface for link annotation with guidance. The left side is the log of messages and the right side is the set of annotated conversations. The red box is the message to be annotated. Yellow / orange highlights are four of the predictions from the out-of-domain model. Blue and green text are explained in Section 4.1. The annotator needs to select the earlier message that the red message is replying to. The red message will then be added to the same conversation (on the right) as the message it is replying to. If the red message is the start of a new conversation the annotator will press a special button (not shown here).

links before adjudication. Agreement is as good or better than prior work. Based on our experience doing annotation, the Ubuntu Meeting channel was harder to annotate because the discussion was rapid and interleaved. The model struggles in this domain, with by far the lowest performance, as shown in Table 2.

4 Improving Annotation

To go beyond expert-annotated resources, we need effective annotation methods for either users (e.g., domain experts who run a channel and are willing to annotate data for their own use) or crowd workers (who can be recruited at larger scale). We perform the first experiments in annotation with both of these groups, exploring several variations in tool support for them.

4.1 Annotation Tools

Figure 1 shows a screenshot of part of our tool. We considered two forms of variation: (1) the type of annotation and (2) whether guidance is provided. In all cases, there was an interactive tutorial that explained the interface and annotation conventions.

Annotation Type Conversations can be annotated in two ways: forming sets of messages, where each set is a conversation (set-based); or creating a graph of reply-to links between messages, in which case each connected component in the graph is a conversation (link-based). This is the first systematic comparison of these two types of annotation.

Guidance We implemented guidance to help annotators. We used the feedforward neural network model from Kummerfeld et al. (2019), trained on their Ubuntu data, to predict reply-to links. For details of the model architecture, training, and in-domain accuracy, see Kummerfeld et al. (2019). Our data is out-of-domain for the model, and it is not perfect even in-domain, so we showed the top five predictions, with darker shades of yellow indicating more likely options. On this data, the top five predictions have an average recall of 92%. In the link annotation case, we highlighted individual messages, as shown in Figure 1. In the set annotation case, we highlighted the conversations those messages belong to.[6]

We also changed the colour of messages to indicate likely interactions: (a) messages written by the current user and any message that addresses[7] the current user were green, (b) if the current message addresses someone, then we made messages from

[6] If multiple predicted messages were in the same conversation then the shade of yellow is based on the max probability of the options.

[7] This is when one user mentions another user in a message.

that user green as well, and (c) messages where both (a) and (b) were true were blue. Figure 1 shows examples of these variations.

4.2 Participants

Participants were randomly split into the four task conditions. They completed an interactive tutorial, then annotated a 34 message sample from each channel. Following Kummerfeld et al. (2019), we provided 1,000 prior messages as context. To mitigate learning and task fatigue effects, we varied the order of the channels across participants.[8]

Domain Experts We recruited seventeen fluent English speakers who were PhD students in Computer Science at the University of Michigan, but not doing research in NLP. They have knowledge of the subject area, but no prior experience with disentanglement. Each participant received an Amazon gift card valued at $25 for assisting in the study. We have excluded one participant, who misunderstood the task, performed extremely poorly, and expressed confusion.

Crowd We recruited 128 workers via Amazon Mechanical Turk, requiring that workers had a 98% HIT approval level and be U.S.-based. Each HIT was worth $3.75, an effective rate of $15 per hour when counting time spent reading instructions and doing the tutorial as well as the task.

4.3 Metrics

We considered three measures of agreement between our participants and the experts: κ, the standard metric applied to reply-to links; *Conv-F*, an F-Score calculated based on how many conversations match exactly; and *1-1*, a conversation-matching metric from Elsner and Charniak (2008). We also measured the time taken. Note that κ can only be calculated for cases where the type of annotation is reply-to links (Kummerfeld et al., 2019).

We also include the accuracy of the model that provided guidance. This provides a baseline that annotators must exceed for their work to be helpful.

We do significance testing with one-tailed unpaired t-tests. To control for family-wise errors, we apply the Holm-Bonferroni Method (Holm, 1979). Results of tests are described where relevant in the text below.

[8]The orders were: SRUM, RMSU, USMR, MURS (S = Stripe, R = Rust, U = Ubuntu Meeting, and M = Mediawiki).

Anno. Type	Guid-ance	Accuracy κ	Conv-F	1-1	Time (min)
Computer Science PhD Students					
Conv	No	-	51	80	6
Conv	Yes	-	58	87	7
Link	No	0.68	43	80	6
Link	Yes	**0.79**	**69**	**92**	10
Crowd workers					
Conv	No	-	33	74	5
Conv	Yes	-	39	70	6
Link	No	0.52	19	64	8
Link	Yes	0.55	37	69	9
Automatic		0.68	53	78	-

Table 3: Accuracy and time for each condition. Metrics are defined in Section 4.3. Domain experts provide high quality annotations, particularly with guidance.

4.4 Ethics

The use of public IRC logs was approved by the University of Michigan's IRB, as was the annotation study with human participants (Study IDs HUM00176661 and HUM00172084). To protect the identities of crowd workers, their Amazon IDs will not be released. Details of compensation are provided above, with values chosen to ensure fair payment without being so high as to be coercive. Our results are limited by the range of participants we had in the task and so may not be representative of all domains. This work does not introduce any significant new risks that we are aware of.

5 Results

Table 3 shows results for each of the conditions, which allow us to answer several questions.

Domain experts can annotate accurately. Comparing the top half of the table to the automatic results (bottom row), our participants provide annotations that are more accurate than the model, but only when given guidance (this difference is statistically significant).

Guidance helps domain experts. The conditions with guidance have higher accuracy (significant at the 0.05 level), though at the cost of more time (also significant). This is the reverse of the pattern seen in annotations for tasks such as POS tagging and NER, where guidance improves speed of annotation while keeping accuracy the same. One possible explanation is that the guidance is prompting annotators to read additional options, which helps them find an option they may have otherwise missed, but also leads them to read more,

which takes time. In contrast, guidance in classification tasks such as POS tagging and NER does not reveal additional options (there is a fixed, known tag set) and does not lead to more reading.

Further work is needed to support crowd workers. Crowd workers are worse than the out-of-domain model in every condition. This indicates that further research is needed to help crowd workers succeed. It also shows that the needs of crowd workers and domain experts are different, as the domain experts were effective and improved with guidance, while crowd workers did not (the variations are not statistically significant). However, a few workers did have high accuracy. In a survey, we found that some of our workers had substantial technical knowledge, for example "My Unix experience goes back to SVR4 days (mostly IRIX & Solaris - ugh), and I still code on Linux occasionally". This suggests that domain experts exist in the crowd workforce and if they can be identified, e.g., by pre-screening, they may be as accurate as the students in our study.

Link-based and set-based annotation are comparable in accuracy. When comparing conditions that are equivalent except for the type of annotation, there is no statistically significant difference. This result answers the question from Elsner and Charniak (2010). We advise future work to annotate reply-to links as it provides additional information about the internal structure of conversations.

How should future work annotate disentanglement? Use domain experts, provide them with guidance, and ask them to annotate links. This led to our best results and provides internal structure.

6 Limitations

There are three main limitations of this work. First, the study participants are an approximation of domain experts, rather than being actual users of the IRC channels we consider. We believe Computer Science students are a reasonable proxy, given their knowledge of the subjects discussed in these channels, but it is possible that they are unaware of community-specific conventions or jargon.

Second, we only considered online communities writing in English. It is possible that communities writing in other languages use significantly different conventions that make this task easier or harder.

Third, our sample size is only large enough to make strong claims about some of the variations in results. It's possible that other variations in Table 3 would also be significant if we had a larger set of participants.

7 Conclusion

This work makes two key contributions. First, the new dataset we are releasing expands the scope of multi-domain evaluation of conversation disentanglement models. Second, our user study of variations in annotation tools shows that domain experts can effectively annotate, particularly when given automatic guidance. Together, these contributions show how better models and systems can be created that give domain-expert users the ability to improve systems. That will enable the use of this technology in a wide variety of new domains.

Acknowledgements

This material is based in part on work supported by DARPA (grant #D19AP00079), and the ARC (DECRA grant).

References

Jhonny Cerezo, Felipe Bravo-Marquez, and Alexandre Henri Bergel. 2021. Tools impact on the quality of annotations for chat untangling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 215–220.

Kent Chang, Danica Chen, and David Bamman. 2023. Dramatic conversation disentanglement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4020–4046, Toronto, Canada.

Fu-Dong Chiou, David Chiang, and Martha Palmer. 2001. Facilitating treebank annotation using a statistical parser. In *Proceedings of the First International Conference on Human Language Technology Research*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of The Sixth AAAI Conference on Human Computation and Crowdsourcing*, pages 12–20.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 834–842.

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

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88.

Matthew R. Gormley, Adam Gerber, Mary Harper, and Mark Dredze. 2010. Non-expert correction of automatically generated relation annotations. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 204–207.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Nancy Ide and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Springer Netherlands.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

Jonathan K. Kummerfeld. 2019. Slate: A super-lightweight annotation tool for experts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S. Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623.

Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 144–152.

Matthieu Riou, Soufian Salim, and Nicolas Hernandez. 2015. Using discursive information to disentangle French language chat. In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC 2015) / Social Media at GSCL Conference 2015*, pages 23–27.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.