Effective Crowdsourcing for a New Type of Summarization Task

Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Walter S. Lasecki
Computer Science & Engineering
University of Michigan, Ann Arbor
{lyjiang,cfdollak,jkummerf,wlasecki}@umich.edu

Abstract

Most summarization research focuses on summarizing the entire given text, but in practice readers are often interested in only one aspect of the document or conversation. We propose "targeted summarization" as an umbrella category for summarization tasks that intentionally consider only parts of the input data. This covers query-based summarization, update summarization, and a new task we propose where the goal is to summarize a particular aspect of a document. However, collecting data for this new task is hard because directly asking annotators (e.g., crowd workers) to write summaries leads to data with low accuracy when there are a large number of facts to include. We introduce a novel crowdsourcing workflow, Pin-Refine, that allows us to collect high-quality summaries for our task, a necessary step for the development of automatic systems.

1 Introduction

Our lives are increasingly dependent on information, but so much is generated every day that manually processing it is overwhelming (Jones et al., 2004). For decades, research in NLP has focused on automatic summarization as a solution to this problem (Nenkova and McKeown, 2012). However, most of that research has focused on generic summarization, where the summary aims to produce a shorter form of a document. Variants of this task, query-based summarization and update summarization, consider summarization focusing on certain parts of the document, but neither covers the situation when a user wants the summary to capture a particular aspect of a document. For example, a legal case document can contain multiple types of information, such as facts, procedural history, and legal reasoning – but a lawyer may only want a summary of the facts stated in the document, while leaving out procedural history and legal reasoning.

This paper makes two contributions: First, we propose a new hierarchy of summarization task types, which provides a framework for understanding how tasks relate to one another and where gaps exist currently. We define a new concept, *targeted summarization*, that contrasts with generic summarization. We then define a new category of summarization task, *aspect-based summarization*, that covers cases like the law example above.

Second, we present and evaluate a new crowdsourced data collection workflow pattern, *Pin-Refine*, that splits the summarization task into two stages: *choosing what to summarize* and *writing the summary*. We apply this approach to a dialog dataset, where questions are expressed over multiple turns, to collect summaries that concisely express each question. Our results show that when more facts need to be summarized, the Pin-Refine workflow produces significantly more accurate summaries compared to a baseline approach in which crowd workers read text and write summaries in a single step. Our method enables efficient creation of datasets for this new task and may be beneficial for other summarization tasks.

2 Related Work

Our work on targeted summarization is related to previous work in automatic summarization and crowdsourced corpus generation.

2.1 Automatic Summarization

In the most common form of summarization, generic summarization, summaries cover all the content in the given text (Gong and Liu, 2001). Specific variants of the task exist for certain domains, such as narrative (Mani, 2004) and email-thread summarization (Rambow et al., 2004).

Proceedings of NAACL-HLT 2018, pages 628–633
New Orleans, Louisiana, June 1 - 6, 2018.

In contrast, summaries for query-based summarization only cover parts of the text that are about the topic specified by a query (Rahman and Borah, 2015). Another alternative is update summarization, in which the summary should cover content in one set of documents, but not in another set that the user has already read (Dang and Owczarzak, 2008). The specific form of summarization we are interested in does not fit within either query-based or update summarization. To clarify the relationships between all of these different summarization tasks, we propose a new term, aspect-based summarization, and present a hierarchy of tasks.

In data mining, recent work has explored summarizing different aspects of graph data given domain context (Jin and Koutra, 2017). In NLP, previous summarization tasks have explored summarization based on information types in individual domains, such as opinion summarization (Condori and Pardo, 2017) and task-focused email summarization (Corston-Oliver et al., 2004). Performance on these tasks is usually lower than traditional summarization tasks due to the difficulty of identifying relevant information in noisy text. We introduce a new crowdsourcing workflow, Pin-Refine, that improves the quality of data collection for specialized summarization tasks.

2.2 Crowdsourced Corpus Generation

Large corpora are critical for training robust natural language processing systems, but traditional expert-driven data collection methods are both costly and time-consuming (Hovy et al., 2006). During the last decade, crowdsourcing has been broadly applied to collect natural language data at large scale with reasonable costs (Snow et al., 2008), including for translation (Zaidan and Callison-Burch, 2011), paraphrasing (Burrows et al., 2013; Jiang et al., 2017), dialog generation (Lasecki et al., 2013b,a), and annotation of corpora in tasks like sentiment classification (Hsueh et al., 2009).

Since individual workers' outputs are usually error-prone, aggregation mechanisms such as majority voting (Raykar et al., 2010) and quality verification tasks (Callison-Burch, 2009) have been developed to improve consistency. However, the results receiving the most votes may still miss information that should be included. To address this issue, crowdsourced iterative methods have been developed to divide a complicated task into a series of micro-stages, each with a different focus (Little et al., 2010; Merritt et al., 2017). For example, Ouyang et al. (2017) developed a dataset of aligned extractive and abstractive summaries by creating separate tasks for summarization, alignment, and classification of changes. However, maintaining accuracy when the complexity of the given text increases has remained an open question. In our Pin-Refine workflow, workers first identified all text relevant to the given information type, which was aggregated across workers with a threshold, then wrote the summary using that information. This aggregation and priming helps maintain accuracy as text grows more complex.

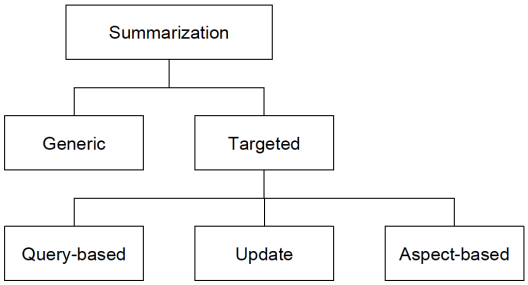

Figure 1: Proposed hierarchy of summarization tasks.

3 Targeted and Aspect-Based Summarization

Traditionally, the NLP community has divided summarization tasks into generic summarization, which covers the entire text, query-based summarization, which covers only topics related to a query provided by the user (Nenkova and McKeown, 2012), and update summarization, which covers only topics that were not addressed in documents already presented to the user (Dang and Owczarzak, 2008). In query-based and update summarization, what should and should not be summarized depends on a topic (defined either by the query, or the already-read documents). This view omits cases where the user describes an information need using something other than topics.

We therefore re-categorize summarization tasks (Figure 1) as generic or *targeted*. We define the latter as the task of generating a summary that captures the part of a document relevant to the user's information request. It includes (1) query-based summarization, where the information request is a query indicating the desired topic, (2) update summarization, where the request is whether information is new, and (3) *aspect-based summarization* where the request is the desired information type,

Person A: I am a *CS major* and need to *schedule classes for next semester*.
Person B: It looks like you have most of your prerequisites out of the way and you can start taking some more EECS classes.
Person A: Cool, I'm very *interested in Software Infrastructure Applications and web app*.

Figure 2: Example conversation for summarization. Targeted information units (TIUs) are in italics.

which can partially cover content of one or more topics. The information type targeted by aspect-based summarization varies based on users' needs. For example, meeting attendees might want a summary of a meeting transcript only including action items, while a supervisor evaluating an employee might need a summary of status updates the employee provided at the meeting.

This hierarchy shows the relationship among existing summarization tasks and provides a framework to understand how the aspect-based task and potential future tasks relate to each other.

4 Experimental Design

When developing datasets for summarization, we are concerned with two key properties: accuracy and fluency. We conducted experiments to investigate design options for crowdsourced aspect-based summarization, aiming to optimize both.

4.1 Conversation Generation

In this study, we focus on summarizing student questions regarding course selection from advising conversations. For such a question to be correctly expressed, the summary must include all relevant facts about the student's background and preferences that appear in the conversation, as shown in Figure 2. We call these facts *targeted information units* (TIU), because they are the pieces of information that must be part of the summary for the given information type.

In this paper, we tested our workflows on a course advising conversation dataset produced by undergraduates role-playing as students and advisors. The goal of the conversations was to determine what courses the students should take based on their needs, as shown in the example conversation in Figure 2. Each "advisor" received a list of course profiles, and each "student" received a made-up student profile, including courses they had taken. Participants were instructed to use the profiles they received while letting the conversation proceed as smoothly as possible.

Rewrite Questions

Please read each conversation below and rewrite ALL parts of Person A's question, so that Person B can answer it without seeing the conversation. For example:

Example Conversation

Person A: I wanted to talk about my classes for next semester.
Person B: Okay, great. How many credits are you planning to take?
Person A: I was hoping to have a relaxed semester, so I'm hoping to take 12 credits.

You may write: 'What classes can I take next semester if I want 12 credits?'

If one sentence is not enough, use multiple sentences.

Figure 3: Baseline task instructions and examples.

4.2 Conversation Selection

We selected 30 conversations from the dataset mentioned above. Each conversation focuses on answering one question, and the number of TIUs per conversation varies evenly between 1 and 6 among the 30 conversations. Three of the authors—two native English speakers and one fluent speaker—read each conversation and summarized each user question. The lead author then compared these summaries and chose one per conversation as the ground truth summary.

4.3 Conditions

Baseline We recruited crowd workers via LegionTools (Lasecki et al., 2014; Gordon et al., 2015) from Amazon Mechanical Turk, presenting them with instructions and an example as shown in Figure 3. Workers were shown 5 conversations, one at a time, and asked to write the question being asked, including all the details that need to be known in order to correctly answer the question. Each worker was paid 10 cents per conversation.

Highlight In this condition, workers were first asked to highlight all details in each conversation that must be known to correctly answer the question, then write the question being asked, including the details they highlighted. We hypothesize that workers were primed by the process of highlighting TIUs in conversations before writing the actual summaries. In this condition, workers were paid 15 cents per conversation[1].

[1]Payments in the Highlight and Pin-Refine conditions were higher than in the baseline, because those two conditions required workers to also do priming tasks.

Targeted Units	Time (s)			Correct Intent (%)			Targeted Information Units Captured									Fluency (%)		
							Recall			Precision			F_1					
	b	h	p	b	h	p	b	h	p	b	h	p	b	h	p	b	h	p
1	26.5	**45**	**107**†	100	94	96	100	96	92	94	96	88	96	96	89	84	90	86
2	35	**46**	**131.5**†	98	96	100	86	87	92	93	99	95	88	91	93	78	78	94
3	48	**63**	**155.5**†	94	84	96	73	78	86	89	93	91	79	83	87	66	72	78
4	60	**82**	**161.5**†	86	84	96	69	76	**85**	91	87	91	76	79	**87**	76	62	82
5	76.5	**94.5**	**192**†	78	88	94	68	71	**88**†	88	90	92	75	78	**90**†	80	80	76
6	92.5	**116.5**	**230**†	84	94	94	70	69	**81**†	87	88	90	77	76	**85**†	72	68	64

Table 1: Performance for a range of metrics (defined in § 4.4) as the number of targeted information units and the condition vary (b: Baseline, h: Highlight, p: Pin-Refine). Bold indicates a statistically significant difference compared to the baseline at the 0.05 level, and a † indicates significance compared to the highlight condition at the 0.05 level, both after applying the Holm-Bonferroni method across each row (Holm, 1979).

Pin-Refine This condition had two separate steps: pin and refine. In the pin step, workers selected sections of the text as in the highlight case, and were paid 5 cents per conversation. Highlights from multiple workers for each conversation were automatically aggregated by keeping highlights if the percentage of workers who assigned them was above a threshold. In the refine step, a different worker was shown the conversation with highlights and asked to write a justification of each highlight, *then* write a summary. Each worker was paid 15 cents per conversation for the refine step.

To find and validate the correct threshold in the pin step, we repeated the data collection and aggregation of the pin step twice on the same set of conversations. In both attempts, all of the TIUs were covered by aggregated highlights at 40% agreement, and no completely irrelevant information was covered. While we used 40% agreement as our threshold, we also observed that coverage was robust to variation in this value. When very high agreement was required (70%) we still found on average 90% of correct phrases were covered (recall remains high), and when very low agreement was required (20%) only 7% of highlighted phrases were irrelevant (precision remains high).

4.4 Metrics

We evaluate question summaries on three metrics: *time* was measured directly; *accuracy* and *fluency* were independently rated by three of the authors. We used Fleiss' Kappa to measure the inter-annotator agreement between the three annotators before discussing each case of disagreement for consensus judgment. The kappa scores were .95 for intent accuracy, .86 for TIU accuracy (both near-perfect agreement), and .62 for fluency (substantial agreement) (Altman, 1990).

Accuracy An accurate question summary must ask for the information sought by the student (intent accuracy) and include all the information needed to define the question (TIU accuracy). Three authors rated the question intent in each summary and counted how many of the gold TIUs were present, as well as how many information units not in the gold appeared in the summary. To measure TIU accuracy, we calculated recall, precision, and F_1 score. We used Fisher's exact tests and Mann Whitney U tests to measure significance of intent and TIU accuracy (respectively) between each pair of conditions in our study.

Fluency A fluent summary is grammatically correct or correct but for minor errors of punctuation. Run-on sentences, sentences with grammatical errors that obscure their meanings, sentences missing words, and so on, are not fluent. We used a χ^2 test to measure significance.

Time To estimate time-to-completion and ensure fair payment, we measured and calculated the average time between when a worker submitted one summary and the next. Time spent on the first summary was excluded because it typically includes time spent reading the instructions and understanding the task, which would skew the data. We report the median time to avoid skewing due to outliers, such as a value of five minutes when a worker took a break, and used a Moods Median test to measure significance.

5 Results

We spent $153.50, including initial testing, to collect 900 summaries: 10 summaries for each of the 30 conversations in all 3 conditions. We have released this dataset as an attachment to this paper.

Table 1 shows the results accross all of our metrics. We find there was relatively little variation in correctness and fluency of summaries across conditions. For the baseline, accuracy and fluency of summaries decreases as the number of TIUs per conversation increases.

Aggregation and priming had a major impact on recall and F_1 when the number of TIUs was greater than three. After that point, the Pin-Refine condition achieved significant improvements in recall and F_1 compared to the baseline. There was no significant difference between accuracy of the baseline and highlight conditions, likely because workers were primed by their own mistakes and chose not to highlight information they believed was not important. Precision remains relatively high with no significant difference across all conditions, implying that workers' ability to effectively exclude information is not related to the targeted information type.

The time workers spent summarizing one conversation increases as the number of TIUs per conversation increases. The significant time increase between the baseline and the other two conditions was caused by the additional work involved in highlighting and writing justifications.

On average, workers spent significantly longer on the justification task in the Pin-Refine condition (65.37s) than the highlighting task in the highlight condition (36.68s). Workers' justifications include single words like "timestamp," short phrases like "why they want a specific course," and long sentences like "This shows what grade they're in, what related class they've taken, what their interest is, and what kind of help they need." One possibility is that simply encouraging workers to spend more time writing their summaries improved performance, but fitting a linear model we find the correlation coefficient between time and F_1 is -0.06, indicating no linear correlation between time and accuracy across conditions. Therefore, we believe that the significant accuracy improvement observed in the Pin-Refine condition is the result of active priming with aggregated TIUs.

6 Conclusion

In this paper, we have identified a previously unexplored summarization problem that targets specific information in a document instead of aiming to extract all key elements: aspect-based summarization. Then, to address the corresponding gap in techniques for data collection for this new problem, we proposed the Pin-Refine crowdsourcing workflow, which leverages input aggregation and worker priming effects. This approach leads to significantly higher summarization accuracy when the number of targeted information units (TIUs) is large. Our work provides methods and task design guidance for future data generation efforts, which are crucial for the development of robust summarization systems.

7 Acknowledgements

We would like to thank Sai Gouravajhala, Stephanie O'Keefe, and the anonymous reviewers for their helpful suggestions on this work, and all of our study participants for their work.

This material is based in part upon work supported by IBM under contract 4915012629. Any opinions, findings, conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of IBM.

References

Douglas G Altman. 1990. *Practical statistics for medical research*. CRC press.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4(3):43. https://dl.acm.org/citation.cfm?id=2483676.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. pages 286–295. https://www.aclweb.org/anthology/D/D09/D09-1030.pdf.

Roque Enrique López Condori and Thiago Alexandre Salgueiro Pardo. 2017. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications* 78:124–134.

Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. *ACL Workshop on Text Summarization Branches Out* http://aclweb.org/anthology/W/W04/W04-1008.pdf.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *Text Analysis Conference*. pages 10–23.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic

analysis. In *Research and Development in Information Retrieval*. ACM, pages 19–25. https://dl.acm.org/citation.cfm?id=383955.

Mitchell Gordon, Jeffrey P Bigham, and Walter S Lasecki. 2015. Legiontools: a toolkit+ ui for recruiting and routing crowds to synchronous real-time tasks. In *User Interface Software and Technology*. ACM, pages 81–82. https://dl.acm.org/citation.cfm?id=2815729.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70. http://www.jstor.org/stable/4615733.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. pages 57–60. https://dl.acm.org/citation.cfm?id=1614064.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *NAACL HLT 2009 workshop on active learning for natural language processing*. pages 27–35. https://dl.acm.org/citation.cfm?id=1564137.

Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding task design trade-offs in crowdsourced paraphrase collection. In *Association for Computational Linguistics (Volume 2: Short Papers)*. pages 103–109. http://aclweb.org/anthology/P17-2017.

Di Jin and Danai Koutra. 2017. Exploratory analysis of graph data by leveraging domain knowledge. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, pages 187–196.

Quentin Jones, Gilad Ravid, and Sheizaf Rafaeli. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research* 15(2):194–210.

Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *User Interface Software and Technology*. ACM, pages 551–562. https://dl.acm.org/citation.cfm?id=2647367.

Walter S Lasecki, Ece Kamar, and Dan Bohus. 2013a. Conversations in the crowd: Collecting data for task-oriented dialog learning. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013b. Chorus: a crowd-powered conversational assistant. In *User Interface Software and Technology*. ACM, pages 151–162. https://dl.acm.org/citation.cfm?id=2502057.

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Turkit: human computation algorithms on mechanical turk. In *User Interface Software and Technology*. ACM, pages 57–66. https://dl.acm.org/citation.cfm?id=1866040.

Inderjeet Mani. 2004. Narrative summarization. *Traitement Automatique des Langues* 45(1).

David Merritt, Jasmine Jones, Mark S Ackerman, and Walter S Lasecki. 2017. Kurator: Using the crowd to help families with personal curation tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. pages 1835–1849. https://dl.acm.org/citation.cfm?id=2998358.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data* pages 43–76.

Jessica Ouyang, Serina Chang, and Kathy McKeown. 2017. Crowd-sourced iterative annotation for narrative summarization corpora. In *European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pages 46–51. http://aclweb.org/anthology/E/E17/E17-2008.pdf.

Nazreena Rahman and Bhogeswar Borah. 2015. A survey on existing extractive techniques for query-based text summarization. In *Advanced Computing and Communication (ISACC), 2015 International Symposium on*. IEEE, pages 98–102.

Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*. pages 105–108. http://aclweb.org/anthology/N/N04/N04-4027.pdf.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11(Apr):1297–1322.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Empirical Methods in Natural Language Processing*. pages 254–263. http://aclweb.org/anthology/D/D08/D08-1027.pdf.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pages 1220–1229. http://www.aclweb.org/anthology/P11-1122.