Overview of MSLR2022: A Shared Task on Multi-document Summarization for Literature Reviews

Lucy Lu Wang[1,2], **Jay DeYoung**[3], **Byron Wallace**[3]
[1]Information School, University of Washington, Seattle, WA, USA
[2]Allen Institute for AI, Seattle, WA, USA
[3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
lucylw@uw.edu; {deyoung.j, b.wallace}@northeastern.edu

Abstract

We provide an overview of the MSLR2022 shared task on multi-document summarization for literature reviews.[1] The shared task was hosted at the Third Scholarly Document Processing (SDP) Workshop at COLING 2022. For this task, we provided data consisting of gold summaries extracted from review papers along with the groups of input abstracts that were synthesized into these summaries, split into two summarization subtasks. In total, six teams participated, making 10 public submissions, 6 to the Cochrane subtask and 4 to the MSˆ2 subtask. The top scoring systems reported over 2 points ROUGE-L improvement on the Cochrane subtask, though performance improvements are not consistently reported across all automated evaluation metrics; qualitative examination of the results also suggests the inadequacy of current evaluation metrics for capturing factuality and consistency on this task. Significant work is needed to improve system performance, and more importantly, to develop better methods for automatically evaluating performance on this task.

1 Introduction

Systematic literature reviews aim to comprehensively summarize evidence from all available studies relevant to a research question. In medicine, such reviews constitute the highest quality evidence used to inform clinical care. Reviews are expensive to produce manually, taking teams of experts months to years to complete, and go out of date quickly (Shojania et al., 2007); (semi-)automation may facilitate faster evidence synthesis without sacrificing rigor. Toward this end, we initiated the MSLR2022 shared task to investigate challenges in multi-document summarization and synthesis for medical literature review. In addition to soliciting direct submissions towards the task, we encouraged work extending our task/datasets, e.g., proposing scaffolding tasks, methods for model interpretability, and improved automated evaluation methods.

We organized the task into two subtasks based on two datasets we provided: MSˆ2 (DeYoung et al., 2021) and Cochrane (Wallace et al., 2020). We received submissions and/or system reports from six participating groups. A selection of generated summaries from the final submissions will be sampled and subject to human annotation for quality and consistency against the gold summaries. The human annotations produced following the shared task will be released as a public dataset to encourage further work on this task and its associated automated evaluation metrics. In the rest of this overview, we provide descriptions of the shared task (Section 2), the baseline models (Section 3), submitted systems (Section 4), and a summary of insights and directions for future work (Section 5).

2 Task description

We give a brief description of the datasets, task, evaluation metrics, and submission protocol for the shared task.

Datasets We provided two datasets for model iteration and evaluation. The MSˆ2 dataset consists of 20k reviews (comprising 470K studies) from the literature to study the task of generating review summaries (DeYoung et al., 2021). Reviews and studies for MSˆ2 were collected from PubMed. Input studies were filtered from cited articles using keyword heuristics and a SciBERT-based suitability classifier trained on human annotations, and the target summary was extracted from the review abstract using a SciBERT-based sequential sentence classifier trained on manually-labeled sentences from over 200 abstracts (see DeYoung et al. (2021) for details). Target summaries in the test set were manually reviewed and corrected. In addition to the abstracts of input studies and summaries, MSˆ2 extracts a background section from each review as

[1]https://github.com/allenai/mslr-shared-task

Proceedings of the Third Workshop on Scholarly Document Processing, pages 175–180
October 12–17, 2022.

context for the research question.

The Cochrane dataset consists of 4.6K reviews from the Cochrane Library (Wallace et al., 2020).[2] The target summaries are the Authors' Conclusions sections of the review abstracts. The Cochrane dataset is smaller and more consistent than the MSˆ2 dataset since all Cochrane reviews follow a similar process. For more information on dataset construction, please refer to the original dataset papers (DeYoung et al., 2021; Wallace et al., 2020).

Task Given the abstracts of input studies pertaining to a research question (and in the case of MSˆ2, a background section describing that research question), the task is to produce a summary that synthesizes the information from the input studies. The synthesis of information typically results in an evidence "direction," e.g., the evidence overall suggests that the intervention studied *increases/decreases/does not change* the outcome measure for the studied population (DeYoung et al., 2020). The direction of the evidence indicated in a good generated summary should agree with that in the reference (gold) summary.

Evaluation We perform automated evaluation using ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and the *evidence inference* (Lehman et al., 2019) divergence metric defined in Wallace et al. (2020) and modified by DeYoung et al. (2021). For ROUGE, we report ROUGE-1, ROUGE-2, and ROUGE-L. For the evidence inference-based metric, we report the average divergence (ΔEI Avg) and the Macro-F1 (ΔEI F1) computed using a model trained on the dataset provided by DeYoung et al. (2020).

For human evaluation, we developed and iterated on an annotation protocol based on the analysis conducted by Otmakhova et al. (2022b). For each annotation task, annotators are shown a gold summary and a generated summary and asked to assess the latter for (i) fluency and (ii) agreement with the gold summary in terms of the "PICO" element alignment,[3] evidence inference directional agreement, and alignment regarding the strength of the claims made in summaries. We will provide further details on human annotation results following the shared task meeting.

[2]Cochrane is an international non-profit dedicated to using evidence to inform decision-making.

[3]A framework describing question important to evidence-based medicine. PICO stands for Population/Problem, Intervention, Comparator, and Outcome.

Submissions Leaderboards for submissions are provided for the two subtasks: MSˆ2[4] and Cochrane.[5] Submissions to the leaderboard are judged against the gold summaries in the test splits using the automated metrics described previously.

3 Baselines

We provide several baseline models for comparison. Baseline models from DeYoung et al. (2021) are based on the BART (Lewis et al., 2020) and Longformer (Beltagy et al., 2020) architectures. For both subtasks, we report results of the two baseline models finetuned on the subtask dataset and evaluated on the corresponding subtask test set, as well as on the opposing test set (e.g. trained on MSˆ2 and tested on Cochrane and vice versa).

For MSˆ2, we also evaluate the condition of simply providing the background section as the generated summary. This baseline performs relatively well, indicating potential limitations of the chosen automated evaluated metrics as alluded to in Otmakhova et al. (2022b).

4 Participating systems

We provide brief descriptions of all participating systems. System performance as assessed using automated evaluation metrics are given in Table 1.

ITTC (Otmakhova et al., 2022a) The team adapted PRIMERA (Xiao et al., 2022), a model based on Longformer Encoder-Decoder (Beltagy et al., 2020) that has been designed for multi-document summarization, resulting in strong performance on the MSLR Cochrane subtask. In addition to fine-tuning on the entire training sets of the MSLR shared task, the team also experimented with zero- and few- shot learning scenarios. The authors found that ROUGE did not adequately capture the performance drops observed in the zero- and 10-shot settings, where factuality of the generated summaries was poor. The team also experiment with using global attention to highlight PICO elements in the input and target texts. Though ROUGE did not vary significantly between these two settings, the authors found that when PICO spans are given global attention, the resulting summaries tended to be more abstractive.

LongT5-Pubmed (Yu, 2022) The author attempted to finetune a LongT5 model (Guo et al.,

[4]https://leaderboard.allenai.org/mslr-ms2/

[5]https://leaderboard.allenai.org/mslr-cochrane/

Submitted system (Cochrane)	R-1↑	R-2↑	R-L↑	BERTScore↑	ΔEI-Avg↓	ΔEI-F1↓
SciSpace (Shinde et al., 2022)	**0.262**	0.057	**0.197**	0.859	0.223	0.301
ITTC-2 (Otmakhova et al., 2022a)	0.246	**0.069**	0.184	**0.876**	**0.220**	0.309
LED-base-16k (Giorgi et al., 2022)	0.257	0.066	0.180	0.871	0.275	0.399
ITTC-1 (Otmakhova et al., 2022a)	0.241	0.064	0.179	0.873	0.288	0.338
PuneICT (Tangsali et al., 2022)	0.247	0.055	0.173	0.859	0.271	0.379
LongT5-Pubmed (Yu, 2022)	0.113	0.015	0.090	0.786	0.467	**0.287**
Baselines						
BART-Cochrane	0.240	0.067	0.176	0.863	0.208	0.335
Longformer-Cochrane	0.239	0.066	0.176	0.864	0.235	0.332
Longformer-MSˆ2	0.224	0.054	0.162	0.857	0.375	0.375
BART-MSˆ2	0.230	0.054	0.161	0.854	0.436	0.364

Submitted system (MSˆ2)	R-1↑	R-2↑	R-L↑	BERTScore↑	ΔEI-Avg↓	ΔEI-F1↓
LED-base-16k (Giorgi et al., 2022)	**0.275**	**0.092**	**0.206**	**0.869**	**0.487**	0.424
PuneICT (Tangsali et al., 2022)	0.206	0.035	0.144	0.848	0.532	0.356
LongT5-Pubmed (Yu, 2022)	0.120	0.013	0.096	0.828	0.528	**0.343**
Baselines						
Longformer-MSˆ2	0.264	0.080	0.196	0.867	0.462	0.412
BART-MSˆ2	0.263	0.077	0.195	0.864	0.451	0.414
Copying background section	0.268	0.085	0.181	0.854	0.502	0.395
BART-Cochrane	0.242	0.061	0.170	0.857	0.460	0.331
Longformer-Cochrane	0.221	0.042	0.153	0.850	0.441	0.277

Table 1: System performance for the Cochrane (above) and MSˆ2 (below) subtasks. For baseline systems, the suffix '-MSˆ2' means the model is trained on the MSˆ2 training data, while '-Cochrane' means the model is trained on the Cochrane training data. Top scores among submitting systems are **bolded**; systems are ordered by ROUGE-L.

2022) on the MSLR datasets but found that training was cost and resource prohibitive. The final model submitted to the leaderboards is a LongT5 model pretrained on the Pubmed corpus but which had not been finetuned to the MSLR datasets.

Extract+BART-base (Obonyo et al., 2022) The team explored how input selection strategies can improve the performance of a BART-base mode. The authors fined BART-base on the summarization dataset introduced by Cohan et al. (2018). They considered several extractive techniques to reduce the size of the input sequence, comparing TextRank, LexRank, and models for results extraction to select salient sentences from input documents. Their results suggest that input sampling strategies are promising, though performance gains are inconsistent across the two MSLR subtasks.

PuneICT (Tangsali et al., 2022) The team experimented with finetuning BART-large, DistillBART, and T5-base for both the MSˆ2 and Cochrane subtasks. On the MSˆ2 subtask, finetuned BART-large had the highest performance of the three models based on ROUGE score; on the Cochrane subtask, DistillBART performed best.

SciSpace (Shinde et al., 2022) The team combined a BERT-based extractive method with a BigBird PEGASUS-based abstractive summarization model (Zaheer et al., 2020), leading to strong performance on the MSLR Cochrane subtask. For the extractive step, the authors use a Lecture Summarizer model to identify the most important sentences from the input documents; this method encodes input sentences using BERT, then clusters the contextual representations and selects the sentences closest to the cluster centroids. The resulting sentences are used as input into a BigBird PEGASUS model pretrained on Pubmed, which is finetuned on the MSLR training data. In analysis, the

authors observed that a common error is duplication of statements in the generated summary. The model submitted by the team to the Cochrane subtask leaderboard performs best among submissions based on ROUGE-L, though the authors report that the same training strategy does not lead to good performance on the MSˆ2 subtask due to the much longer input sequences in MSˆ2.

LED-base-16k (Giorgi et al., 2022) The team fine-tuned Longformer Encoder-Decoder following a similar protocol to PRIMERA (Xiao et al., 2022), improving performance over baselines in both subtasks. Their input sequence included the titles and abstracts of up to 25 studies, separated by special tokens. No system description was submitted.

5 Insights & future directions

Though we observe modest overall improvements to task performance based on automated summarization evaluation metrics such as ROUGE and BERTScore, results are inconsistent across evaluation metrics. This is especially the case when considering the evidence inference divergence metrics introduced to measure and bolster inference direction alignments between generated and gold summaries. Further, several participant groups discovered problems with factuality, consistency, duplication, and more with generated summaries upon qualitative examination of their results (Otmakhova et al., 2022a; Shinde et al., 2022). Based on the observations of submitting teams, we summarize two key directions for future research.

Multidocument representation strategies Several submissions explored methods for input extraction and filtering to reduce the size of the input sequence and increase the saliency of the input texts. For both subtasks, a large portion of input instances extend beyond even the token limits of long-sequence transformer language models, and this is especially the case for MSˆ2 (the median number of input documents for MSˆ2 is 17, nearly twice the number for the Cochrane dataset). Obonyo et al. (2022) explored several strategies for sentence selection, including results extraction models, and found promising but inconsistent performance gains over a base model. Shinde et al. (2022) employed a sentence embedding clustering and selection approach, which led to top performance on the Cochrane subtask when combined with a powerful long-sequence trained summarization model. However, Shinde et al. (2022) noted that their methods did not extend well to MSˆ2 due to the larger number of input documents.

Extension of such methods would be a promising future direction. Beyond salient sentence selection, a strategy based on PICO alignment and results extraction may be more pertinent for the specific task. For example, one may only want to include the results sentence from an input document if it studies the same population and research question described in the review. Compression-based methods yielding less computationally intensive representations may also allow for full information retention, enabling salience determinations at the model-level, depending on other input studies and the review question at hand.

Evaluation metrics that better capture summary quality Unsurprisingly, our defined automated evaluation metrics are lacking, in many cases failing to capture summary quality issues identified during qualitative analysis (Otmakhova et al., 2022a; Shinde et al., 2022). Both of our task datasets are highly compressive, e.g. the average compression ratio for the Cochrane dev set is around 33 while that of the MSˆ2 dev set is over 100! Yet, a baseline such as copying the background section of MSˆ2 leads to fairly good performance when assessed using (fuzzy-)token overlap metrics such as ROUGE and BERTScore. This indicates that the task is perhaps less about summarizing and more about *synthesizing* relevant results, and hence, n-gram and token similarity-based metrics would be insufficient for capturing content similarity. These are similar concerns to those raised in single-document summarization evaluation (Fabbri et al., 2021; Deutsch et al., 2022).

We included evidence inference metrics in evaluation to offer a counterpoint to more traditional metrics, yet they bring their own challenges. The values of these metrics are not particularly comparable between the two subtask datasets, nor are the numbers easy to interpret, e.g., how much worse is a model that scores 0.4 to 0.3 ΔEI-F1 at a system level? Additionally, we currently perform evidence inference scoring for all possible PICO tuples, regardless of whether a relationship occurs between members of each tuple, which can lead to degradation in performance (where most tuples are classified as "no effect," washing out actual differences between documents; see discussion in DeYoung et al. 2021). Improvements on PICO tuple

detection and alignment between documents could dramatically improve the value of evidence inference for MSLR evaluation. In addition to evidence inference-based metrics, we anticipate investigating how entailment or question-answering-based evaluation metrics for single-document summarization evaluation (Pagnoni et al., 2021) could be extended into the multi-document space for this task (and how well existing approaches fare on this specialized data and task).

Further data is needed to iterate upon model-based evaluation metrics. Towards this, we intend to collect and release a dataset of human annotations of summary quality for a sample of generations submitted to this shared task, as described in Section 2: Evaluation. Initial results will be presented at the SDP 2022 workshop.

6 Conclusion

The MSLR2022 shared task initiated further investigation into the challenging task of automatically synthesizing study results into a literature review summary. The task received submissions from six teams, leading to modest improvements on task performance and significant insights into the remaining challenges for this task. A primary challenge involves the insufficiency of automated evaluation metrics for assessing performance improvements on this task, towards which we intend to provide new datasets and methods to support and incentivize further research on this problem.

Acknowledgements

We thank John Giorgi for help in data preparation and reviewing for this shared task. This work was funded in part by the National Science Foundation (NSF) awards 2211954 and 1750978, and the National Institutes of Health (NIH) / National Library of Medicine (NLM) award R01-LM012086.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS2: Multi-document summarization of medical studies. In *EMNLP*.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

John Giorgi et al. 2022. MSLR leaderboard: led-base-16384-ms2. https://leaderboard.allenai.org/mslr-ms2/submission/ccfknkbml1mljnftf7d0. Accessed: 2022-09-15.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *NAACL-HLT*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. *arXiv preprint arXiv:1904.01606*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ishmael Obonyo, Silvia Casola, and Horacio Saggion. 2022. Exploring the limits of a base BART for multi-document summarization in the medical domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yulia Otmakhova, Hung Thinh Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022a. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022b. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. 2022. An extractive-abstractive approach for multi-document summarization of scientific articles for literature review. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233.

Rahul Tangsali, Aditya Vyawahare, Aditya Mandke, Onkar Litake, and Dipali Kadam. 2022. Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. In *AMIA Annual Symposium*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Benjamin Yu. 2022. Evaluating pre-trained language models on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. *ArXiv*, abs/1904.09675.