TL;DR PROGRESS: Multi-faceted Literature Exploration in Text Summarization

Shahbaz Syed [†] **Khalid Al-Khatib** [‡] **Martin Potthast** [†§]

[†]Leipzig University [‡]University of Groningen [§]ScaDS.AI

shahbaz.syed@uni-leipzig.de

Abstract

This paper presents TL;DR PROGRESS, a new tool for exploring the literature on neural text summarization. It organizes 514 papers based on a comprehensive annotation scheme for text summarization approaches and enables fine-grained, faceted search. Each paper was manually annotated to capture aspects such as evaluation metrics, quality dimensions, learning paradigms, challenges addressed, datasets, and document domains. In addition, a succinct indicative summary is provided for each paper, consisting of automatically extracted contextual factors, issues, and proposed solutions. The tool is available online at https://www.tldr-progress.de, a demo video at https://youtu.be/uCVRGFvXUj8.

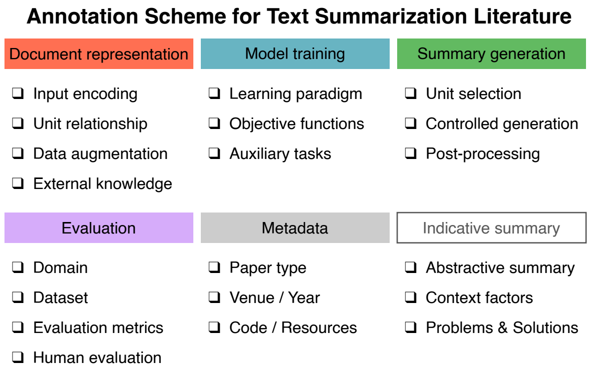


Figure 1: Our annotation scheme is based on a summarization literature analysis. Its four components and their respective facets enable a fine-grained unified analysis of relevant papers. The indicative summary is automatically generated.

1 Introduction

Research in the field of neural text summarization has evolved rapidly from the introduction of sequence-to-sequence (Sutskever et al., 2014; Rush et al., 2015) models to the era of transformers (Vaswani et al., 2017), greatly improving our ability to produce high-quality summaries in line with human preferences (Huang et al., 2020a; Goyal et al., 2022). As a result, the original focus of summarization research on the news domain has broadened to various other domains such as meetings, scientific papers, scripts and opinions.

To keep abreast of current advances, especially researchers new to the field must perform various tasks, including assimilating, organizing, annotating, and reviewing papers across multiple venues. Although search engines tailored to scholarly documents, such as Google Scholar, Semantic Scholar, DBLP, and the ACL anthology, provide access to a vast collection of articles, they merely support the discovery of relevant papers from within a multi-domain collection and do not (strongly) support an in-depth comparative paper analysis.

This paper introduces TL;DR PROGRESS, a literature explorer designed specifically for the text summarization literature. It contributes an intuitive annotation scheme designed to streamline fine-grained, facet-based systematic reviews (Figure 1 and Section 3). To demonstrate the capabilities of our tool, we manually analyze a collection of 514 summarization papers and cover various important aspects for an efficient literature search (Section 4). As part of its goal to organize summarization research, TL;DR PROGRESS provides an indicative summary for each paper, seamlessly integrating automatically extracted contextual information with manually annotated facets (Section 5). This includes identifying for what practical purpose a summarization approach is intended, the current challenges associated with summary generation, and a paper's contributions. Our tool also demonstrates the practical application of large language models (LLMs) in automatic terminology acquisition, involving the extraction of technical terms from papers, including glossary definitions for general concepts and acronym–expansion pairs to improve researchers' recall of specific papers (Section 6).[1]

TL;DR PROGRESS has a dual function: it provides insights into current research and serves as a basis for future automation. In particular, our literature explorer shows a way forward for future research on large-scale systematic reviews of the NLP literature by extensively leveraging LLMs.

[1]https://github.com/webis-de/eacl24-tldr-progress/

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics System Demonstrations, pages 195–206

2 Related Work

Paper aggregators such as Google Scholar, Semantic Scholar, DBLP, and the ACL Anthology provide access to a large number of papers from different disciplines and, most importantly, facilitate their discovery. However, these platforms lack solid support for in-depth comparative analysis. PersLEARN (Shi et al., 2023) introduces a perspective-based approach to exploring scientific literature. It empowers early career researchers to develop their viewpoints by interacting with a prompt-based model. The tool identifies evidence from relevant papers that relate to the given seed perspectives and summarizes them to make new connections. In contrast, TL;DR PROGRESS focuses on summarization and provides fine-grained facets for each paper to enhance understanding of their contributions and content, along with indicative summaries, a feature not present in PersLEARN.

As for annotating papers, Autodive (Du et al., 2023) automates the in-place annotation of entities and relationships in PDFs, using external domain-specific NER models. In contrast, our approach includes a domain-specific annotation scheme and manual annotation for quality assurance. In addition, our tool facilitates unsupervised automatic terminology acquisition using LLMs. SciLit (Gu and Hahnloser, 2023) recommends relevant articles based on keywords entered by the user and generates citation sets with extracted highlights. While our tool supports keyword-based lexical search, it is less reliant on user-defined keywords due to its facet-based retrieval system.

Paperswithcode[2] is a platform that links papers with their code implementations. It provides an overview of the state-of-the-art in various NLP tasks. TL;DR PROGRESS complements Paperswithcode (for the summarization task) by providing an interactive dashboard presenting relevant statistics (Section 7) for a comprehensive understanding of the state-of-the-art.

3 Annotation Scheme

To create a comprehensive annotation scheme for summarization papers, we performed an in-depth analysis of the recent relevant literature. As shown in Figure 1, this scheme encapsulates the basic components of a neural summarization architecture, laying the foundation for a fine-grained annotation tailored to individual contributions. Its underlying principle is to categorize contributions according to their main focus, as papers often address one or more components within the summarization pipeline. The annotation scheme distinguishes four components:

1. **Document representation.** Conversion of a source document into a vector representation to model relationships between document units (words, sentences, paragraphs). This may include input data enrichment with user or style-specific information and model augmentation using external knowledge bases.

2. **Model training.** Training of a model with suitable data under a user-defined objective (or reward) function. This may include using pre-trained models for tasks, such as missing text prediction, paraphrasing, and detecting textual entailment.

3. **Summary generation.** Generation of a summary based on the representation of its source document using a trained model. This may include selecting explicit units for inclusion, restricting the summary to a particular style, conditioning the generation process on certain aspects of the source document, and post-processing steps such as length normalization and redundancy removal.

4. **Evaluation.** Evaluation setup such as the document domains and datasets used for testing, automatic evaluation metrics reported, and the human evaluation criteria for qualitative assessment of the generated summaries.

These components encompass different facets. For example, document representation includes "input encoding", "unit relationship", "data augmentation", and "external knowledge". Definitions for each facet are given in Table 1. These facets are not mutually exclusive, i.e., a paper can contribute to several facets simultaneously. For example, a paper may present a novel input encoding scheme that explicitly models unit relationships in the source document. In such cases, we annotate the paper with multiple facets. The annotation scheme also includes metadata for each paper. Overall, our scheme enables a fine-grained retrieval of relevant summarization papers, a feature that is currently not available in other paper aggregators.

[2] https://paperswithcode.com/

Facet	Description / Examples
Document representation	
Input encoding	The paper presents methods to improve the encoding of source documents (e.g., hierarchical/graphical attention, inclusion of discourse structure, etc.)
Unit relationship	The paper investigates methods that explicitly model the relationship between units in the source document, such as words, sentences, or passages.
Data augmentation	The paper introduces methods that use data augmentation techniques, e.g., to extract aspects, to create contrasting examples of robustness, or to overcome data scarcity in low-resource domains.
External knowledge	The paper investigates methods for integrating external knowledge using resources such as knowledge graphs, domain-specific vocabularies, or information from pre-trained language models.
Model training	
Learning Paradigm	Supervised, unsupervised, or reinforcement learning.
Objective Function	The paper introduces methods that incorporate new objective functions that emphasize diversity, faithfulness, or custom objectives appropriate to the task of summarization.
Auxiliary Tasks	The paper explores methods such as multi-task learning or pre-training on related tasks (e.g., textual entailment, paraphrasing, gap sentence prediction) to improve the summarization task.
Summary generation	
Unit Selection	The paper presents methods that explicitly select relevant units, such as words, sentences, or passages, for summarization, addressing the information loss associated with generating fixed-length summaries through techniques such as copying or pointing.
Controlled Generation	The paper presents methods that encourage the model to generate summaries with certain attributes (e.g., style, length, tone), for example, by providing additional textual guidance or limiting the model's vocabulary to a specific domain.
Post Processing	The paper explores methods for post-processing generated summaries to improve their quality. This includes re-ranking, re-writing or swapping certain text spans to achieve the desired goals.
Evaluation	
Domain	The domain of the source documents (e.g., opinions, screenplays, papers, etc.)
Dataset	The datasets used for training/evaluation (e.g., CNN/DailyMail, XSum, etc.)
Evaluation metric	The metrics used for automatic evaluation (e.g., ROUGE, BLEU, etc.)
Human evaluation	The summary quality criteria that were evaluated manually (e.g., informativeness, fluency, etc.)
Metadata	
Paper type	A new method, analysis (evaluation), metric, dataset, or theory.
Venue / Year	Venue and year in which the work was published.
Code / Resources	Artifacts relevant to reproduce the paper's contribution.

Table 1: Description of the annotation scheme shown in Figure 1. Pipeline components correspond to the three major components of the scheme, Document Representation, Model Training, and Summary Generation.

4 Webis Summarization Papers Corpus

To create TL;DR PROGRESS, we compiled a corpus of research papers on neural text summarization, annotated it according to our scheme, and analyzed the distribution of the papers across different dimensions.

4.1 Corpus Construction

To collect summarization papers, we conducted a keyword search ("summ") in the proceedings of the most important venues, including AAAI, AACL, ACL, CHIIR, CIKM, COLING, CONLL, EACL, ECIR, EMNLP, ICLR, IJCAI, IJCNLP, NAACL, NEURIPS, SIGIR, and TACL. The initial collection of 801 papers was refined through a careful review of titles and abstracts to identify papers that were directly relevant to single-document summarization of English texts. These included papers that evaluated or analyzed existing approaches and proposed new metrics, human assessment methodologies, meta-evaluations, datasets, and new model architectures. To extract textual content from the PDFs, we used Science Parse.[3] Papers that could not be automatically extracted or were duplicates were excluded, so that we ended up processing 514 papers. For each of the 514 papers, we performed a thorough manual annotation, focusing on the facets of our annotation scheme. The annotation was performed by one of the authors of this paper. The annotation process was iterative, with the annotator revisiting the previously annotated sections to ensure consistency. Another author reviewed the annotations to ensure their quality.

[3]https://github.com/allenai/science-parse

Venue	Count	Venue	Count	Venue	Count
EMNLP	184	EACL	13	IJCNLP	4
ACL	115	TACL	12	ICLR	2
NAACL	60	CIKM	12	ECIR	2
COLING	34	AACL	11	NEURIPS	2
AAAI	29	IJCAI	9		
SIGIR	17	CONLL	8		

Table 2: Number of papers published per venue. Unsurprisingly, EMNLP and ACL are the most popular venues for summarization research.

Challenges in Text Summarization
Controlled and Tailored Summarization
Efficient Encoding of Long Documents
Exploiting the Structure of Long Documents
Hallucinations in the Generated Summaries
Identifying Important Contents from the Document
Information Loss / Incoherence in Extractive Summarization
Lack of Suitable Training Data
Pretraining and Sample Efficiency
Robust Evaluation Methods

Table 3: Manually annotated labels for problem statement clusters extracted from all papers, highlighting the prevalent challenges in text summarization.

4.2 Corpus Statistics

Table 2 shows the distribution of papers across venues, with EMNLP and ACL emerging as the top venues for summarization research. Among the 514 papers, we observed the following distribution of paper types: 353 dealt with methods, 79 with analysis (including meta-evaluation and quality/model analysis), 73 were corpus-related, 61 focused on metrics and one on theory. The majority of the proposed models were trained using supervised learning (73%), compared to unsupervised (17%) and reinforcement learning (10%). The different paper types were not mutually exclusive, so there were cases where a paper proposed a new dataset and applied methods to it at the same time. In terms of automatic evaluation, the ROUGE metric was used in 71.6% of papers, highlighting its widespread use for evaluating the quality of generated summaries in the field of single-document summarization of English texts. Only 39.5% of the papers included some form of manual evaluation. In terms of reproducibility, we found that 58% of the papers published their code, indicating a slow but growing trend of code availability in this area compared to previous years.

5 Indicative Summaries of Papers

In contrast to informative summaries that aim to replace the entire paper, our tool provides indicative summaries that help users quickly decide if a paper is relevant to their information need. Our indicative summaries are unique in that they encompass an abstractive summary of the paper as well as multiple facets such as datasets, domains, evaluation metrics alongside other information.

5.1 Beyond Abstract as a Summary

Traditionally, the paper abstract serves the purpose of an informative summary (Luhn, 1958) or an ultra-short abstractive summary (Cachola et al., 2020) that outlines the major contributions. Yet, when dealing with a large collection of documents, these summaries fall short, as they do not enable fine-grained retrieval of relevant papers. Moreover, studies have shown that abstracts can introduce bias and may not offer a comprehensive representation of the paper's contents (Elkiss et al., 2008).

In contrast to informative summaries, which essentially substitute the source, indicative summaries serve as a roadmap for the contents of the source document (Mani, 2001). They aid readers in deciding whether they want to explore the source document in greater detail. Particularly in the context of literature reviews, indicative summaries provide an exploratory overview of papers, allowing researchers to quickly navigate and comprehend their contributions. TL;DR PROGRESS introduces a novel indicative summary that integrates manually annotated facets with automatically extracted contextual information. Motivated by the significance of considering contextual factors in summarization (Jones et al., 1999), we extract information related to: (1) the purpose of the generated summaries, (2) the target audience for the summaries, (3) the downstream application of the generated summaries, and (4) the problems and corresponding solutions presented in the paper. Figure 2 (Appendix) exemplifies an indicative summary generated by our tool. This summary distinctly outlines all the pertinent information that a reader would need to determine whether they wish to delve into the paper in more detail.

5.2 Contextual Information Extraction

We demonstrate the utilization of LLMs for the task of indicative summarization by extracting the contextual information described above through

Context Factors Prompt (GPT3.5)
You are a helpful assistant that can read and analyze scientific papers. You are given the following paper: {*Introduction*} Answer the following three questions: (1) Why are the authors generating the summaries of the documents? (2) Who are they for? (3) How will they be used? You must not include the proposed approach by the authors for generating the summaries. You will output a list of the question-answer pairs where each question is prefixed by the token QUESTION: and each answer is prefixed by the ANSWER: token. Each pair is separated by two lines.
Problems and Solutions Prompt (GPT3.5)
You are a helpful assistant that can read and analyze scientific papers. You are given the following paper: {*Introduction*} Can you give me a list of the main problems tackled by the authors and their proposed solutions? In this list, each problem is described followed by a solution proposed by the authors. Each problem starts with the token PROBLEM and each solution starts with the token SOLUTION. Here is the list:

Table 4: Prompts for extracting contextual information from the introduction of a paper. This information is used to compose indicative summaries of papers. The specific instructions for controlling output format may not be required with newer models.

Glossary Prompt (GPT3.5)
You are a scientist who can read and summarize scientific papers. You are given the following paper: {*Introduction*}. Your task is to extract a list of key concepts along with correct definitions like a glossary of the paper. Follow the format [*Concept*: *Definition*].
Acronyms Prompt (GPT3.5)
You are a scientist who can read and summarize scientific papers. You are given the following paper: {*Introduction*}. Your task is to extract a list of acronyms that the authors use along with correct expansions from the paper. For example (1) EDU: Elementary Discourse Unit, (2) SEHY: Simple Yet Effective Hybrid Model, (3) PLM: Pretrained Language Model. Exclude acronyms for which no expansion is explicitly provided by the authors. Follow the format [*Acronym*: *Expansion*].

Table 5: Prompts for automatic terminology acquisition from the introduction of a paper. We extract glossary as well as acronym-expansion pairs. For the latter, we provide examples of the expected output format.

generative question-answering. To extract this information, we input the introduction section of the paper into the prompt. We devised two prompts corresponding to the *context factors* and *problems and solutions* (see Table 4). Each prompt poses specific questions related to the context, necessitating the generation of answers in a specific format using the relevant content from the paper. We employed GPT-3.5 for our experiments.[4]

We conducted additional analysis of this contextual information to identify the frequently addressed challenges in text summarization. In particular, we employed a soft clustering approach (HDBSCAN (Campello et al., 2013)) on the set of problem statements.[5] This process yielded 9 clusters, which we manually labeled with their respective challenges, as illustrated in Table 3.

6 Automatic Terminology Acquisition

Scientific terminology plays a vital role in research, requiring researchers to recall papers related to specific concepts or acronyms representing models/metrics. Moreover, previously defined terminology might be directly referenced in subsequent papers without detailed explanation (Ball et al., 2002) or even inaccurately paraphrased, compelling researchers to trace back through multiple papers to find the original definitions. The task of automatic terminology acquisition (Judea et al., 2014) aims to tackle this issue by extracting various concepts defined in a paper along with their definitions. In our exploration of this task, we opted for LLMs instead of supervised methods that necessitate labeled data.

We utilized prompt engineering, leveraging GPT-3.5, with the introduction section of the paper as input for automatic terminology acquisition. We formulated two prompts specifically for extracting *glossary definitions* and *acronym-expansion pairs*. Examples of extracted glossary terms and acronym-expansion pairs are provided in Table 6. The prompts are shown in Table 5.

7 Dashboard and Figure Browser

TL;DR PROGRESS includes an interactive dashboard that provides real-time visualizations of key statistics gathered from the annotated documents. The dashboard displays: (1) the number of papers annotated per year, (2) the distribution of publicly released code and resources per year, (3) the popular datasets and document domains for training/evaluation, (4) the commonly emphasized quality criteria of summary (5) the dominant components targeted from the annotation scheme, and (6) the distribution of addressed challenges.

This extensive dashboard delivers a quantitative overview of the text summarization landscape, in line with the detailed facets and additional metadata

[4] https://platform.openai.com/docs/models/gpt-3-5

[5] We clustered the contextual embeddings (Reimers and Gurevych, 2019) combined with dimensionality reduction using UMAP (McInnes and Healy, 2018).

in our annotation scheme. Key findings from the dashboard include:

1. Authors consistently practice releasing code for reproducibility and adoption.
2. News (54.2%) and scholarly documents (13.3%) dominate as the most studied domains, calling for more diverse investigations.
3. The top three evaluated dimensions for summary quality are informativeness (17%), fluency (10%), and coherence (8.1%).
4. The majority of papers propose new objective functions and input encoding approaches.
5. Predominant challenges include controlled summarization, comprehensive evaluation, insufficient datasets, and risks of hallucinations.

The tool also incorporates a dedicated figure browser (Appendix, Figure 3) hosting **1524** figures and tables (with captions) linked to their sources. This resource streamlines navigation and serves as a handy reference for researchers exploring standard illustrations depicting model architectures or layouts for presenting evaluation results.[6]

Term	**Definition / Expansion**
Glossary	
Co-Decoding	An algorithm that takes two review sets as input to compare and contrast the token probability distributions of the models to generate more distinctive summaries (Iso et al., 2021).
Concept-Pruning	An approach to reduce the number of concepts in a model to find optimal solutions efficiently (Boudin et al., 2015).
Drop-Prompt Mechanism	An approach to drop out hallucinated entities from a predicted content plan and to prompt the decoder with the modified plan to generate faithful summaries (Narayan et al., 2021).
Facet Bias Problem	The problem of centrality-based models tending to select sentences from one facet of a document, rather than important sentences from different facets (Liang et al., 2021).
Indegree Centrality	A measure of centrality that assumes a word receiving more relevance score from others is more likely to be important (Xu et al., 2020).
Acronyms	
ADAQSUM	Adapter-based query-focused abstractive summarization (Brazinskas et al., 2022).
COLO	Contrastive learning based re-ranking framework for one-stage summarization (An et al., 2022).
PLATE	Pseudo-labeling with larger attention temperature (Zhang et al., 2022).
ASGARD	Abstractive summarization with graph-augmentation and semantic-driven reward (Huang et al., 2020b).
ASAS	Answer selection and abstractive summarization (Deng et al., 2020).

Table 6: Examples of automatically extracted glossary and acronym–expansion pairs from the papers.

8 Evaluation

We conducted an empirical evaluation of the tool's efficacy in supporting systematic literature reviews for text summarization. The study involved presenting targeted inquiries relevant to beginners in the field and instructing participants to leverage both TL;DR PROGRESS and Semantic Scholar for retrieving relevant papers. Additionally, we systematically gathered feedback on the tool's usability and utility for understanding the effectiveness of its features.

8.1 Purpose-driven User Study

We conducted a study with five participants (3 PhDs, 2 PostDocs) specializing in natural language processing or information retrieval, but unfamiliar with text summarization research. Their task was to find up to five relevant papers for each of the ten research questions, covering various aspects of summarization research using both TL;DR PROGRESS and Semantic Scholar.[7] The following research questions were crafted in reference to the NewSumm Workshop's Call for Papers.[8]

1. How do neural text summarization models address hallucination challenges in abstractive summarization?
2. What are the efficient encoding strategies for handling long documents in neural text summarization?
3. How do neural text summarization models control/tailor the generated summaries to user preferences/aspects/facets?
4. How can pretrained language models be leveraged for improving text summarization?
5. How can additional sources of external knowledge be integrated into the text summarization pipeline?
6. What are the annotation strategies for evaluating hallucination, faithfulness, and factuality in summarization?

[6] We used `PDFFigures 2.0` (Clark and Divvala, 2016).
[7] https://www.semanticscholar.org/
[8] https://newsumm.github.io/2023/

7. List at least five corpora that can be used to train scientific document summarization models?
8. List at least five diverse text domains studied in text summarization?
9. What reward functions are proposed to improve summarization via reinforcement learning?
10. What are the various summary quality criteria evaluated via human assessment?

The participants were instructed to evaluate paper relevance using the summaries from the tools, rating them on a scale from 1 (least relevant) to 5 (most relevant). Alongside this, they were requested to share feedback on the usability, the utility of certain features of TL;DR PROGRESS, and its strengths and limitations. This evaluation provides both a comparative analysis and a qualitative understanding of the tool's practicality.

8.2 Results

Our tool effectively narrowed down the large collection of papers to a set of relevant results. The multi-faceted search, in particular, facilitated quick paper filtering without keyword use. Three out of five participants favored our tool for literature review. However, Semantic Scholar offers a more "familiar" search experience and more recent results, albeit requiring extra effort for relevance filtering. Both tools received a score of 4 for the relevance of results. Users also rated the usefulness of features on a scale of 1 (least useful) to 5 (most useful). The advanced search (combining facets) was highly useful (mean score of 4.5), allowing users to easily adapt searches to the research question at hand. This underscores the utility of our annotation scheme (Table 1). Indicative summaries and the list of challenges were sufficiently useful (mean score of 3.6) for quickly skimming paper contents and finding papers addressing specific problems, respectively. Results are visualized in the Appendix, Figure 4.

8.3 Feedback

Users found the tool intuitive and easy to use, appreciating the multi-faceted search and indicative summaries. The dashboard was viewed as a useful resource for obtaining a quantitative overview of the text summarization field. Users offered constructive feedback, suggesting incorporating a more sophisticated search mechanism and integrating it with facet-based filtering. They pointed out that searching only by conceptual components was insufficient, as the resulting set of papers was still large and required further filtering. These insights will be considered for future improvements to the tool.

9 Conclusion

In summary, TL;DR PROGRESS offers an interactive platform for nuanced exploration of over 500 neural text summarization papers from top venues. Utilizing a tailored annotation scheme, the tool guides users through multifaceted retrieval, provides insightful indicative summaries, outlines challenges, and presents a quantitative overview, easing the entry for newcomers into the field.

Limitations

The tool leverages LLMs for automated summarization, extracting contextual factors like summary purpose, issues, solutions, and scientific terminology from papers. While we conducted a random accuracy check, a comprehensive assessment of hallucinations or faithfulness in the extracted information was not performed. We anticipate that with more advanced models, such as GPT-4, we can enhance the assurance of quality and structure the extracted content more effectively. Currently confined to summarization, the tool's annotation scheme can be readily extended to other domains, bootstrapped by experts accordingly. However, existing facets like datasets, domains, metrics, qualitative evaluation, and learning paradigms can be directly annotated for new domains. Additionally, a forthcoming feature is the tool's capability to incorporate new papers, automating the annotation process—a feature we plan to implement in future updates to the tool.

References

Chenxin An, Ming Zhong, Zhiyong Wu, Qin Zhu, Xuanjing Huang, and Xipeng Qiu. 2022. Colo: A contrastive learning based re-ranking framework for one-stage summarization. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5783–5793. International Committee on Computational Linguistics.

Philip Ball et al. 2002. Paper trail reveals references go unread by citing authors. *Nature*, 420(6916):594–594.

Florian Boudin, Hugo Mougard, and Benoît Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1914–1918. The Association for Computational Linguistics.

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot finetuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1509–1523. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. TLDR: extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4766–4777. Association for Computational Linguistics.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer.

Christopher Andreas Clark and Santosh Kumar Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016*, pages 143–152. ACM.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020. Joint learning of answer selection and answer summary generation in community question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7651–7658. AAAI Press.

Yi Du, Ludi Wang, Mengyi Huang, Dongze Song, Wenjuan Cui, and Yuanchun Zhou. 2023. Autodive: An integrated onsite scientific literature annotation tool. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 76–85. Association for Computational Linguistics.

Aaron Elkiss, Siwei Shen, Anthony Fader, Günes Erkan, David J. States, and Dragomir R. Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *J. Assoc. Inf. Sci. Technol.*, 59(1):51–62.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356.

Nianlong Gu and Richard H. R. Hahnloser. 2023. Scilit: A platform for joint scientific literature discovery, summarization and citation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 235–246. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020a. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 446–469. Association for Computational Linguistics.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020b. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5094–5107. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex aggregation for opinion summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3885–3903. Association for Computational Linguistics.

K Sparck Jones et al. 1999. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12.

Alex Judea, Hinrich Schütze, and Soeren Bruegmann. 2014. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 290–300. ACL.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1685–1697. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Reseach and Devopment*, 2(2):159–165.

Inderjeet Mani. 2001. Summarization evaluation: An overview.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan T. McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Trans. Assoc. Comput. Linguistics*, 9:1475–1492.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389. The Association for Computational Linguistics.

Yu-Zhe Shi, Shiqian Li, Xinyi Niu, Qiao Xu, Jiawen Liu, Yifan Xu, Shiyu Gu, Bingru He, Xinyang Li, Xinyu Zhao, Zijian Zhao, Yidong Lyu, Zhen Li, Sijia Liu, Lin Qiu, Jinhao Ji, Lecheng Ruan, Yuxi Ma, Wenjuan Han, and Yixin Zhu. 2023. Perslearn: Research training through the lens of perspective cultivation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 11–30. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

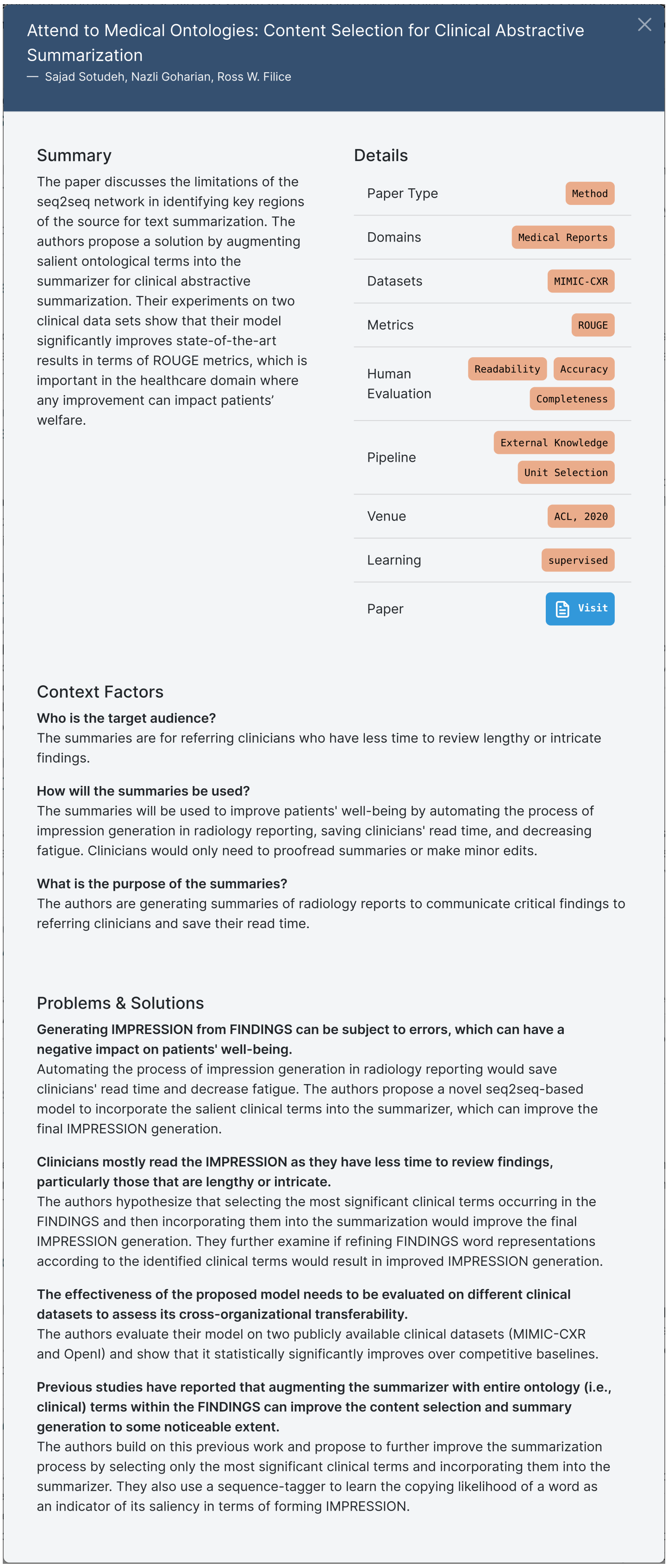
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1355–1362. Association for Computational Linguistics.

Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. 2022. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 127–141. Association for Computational Linguistics.

A Illustrations

This section shows the indicative summaries of a paper and the figure browser. For indicative summary, we combined automatically extracted contextual information using prompts.



Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization

— Sajad Sotudeh, Nazli Goharian, Ross W. Filice

Summary

The paper discusses the limitations of the seq2seq network in identifying key regions of the source for text summarization. The authors propose a solution by augmenting salient ontological terms into the summarizer for clinical abstractive summarization. Their experiments on two clinical data sets show that their model significantly improves state-of-the-art results in terms of ROUGE metrics, which is important in the healthcare domain where any improvement can impact patients' welfare.

Details

Paper Type	Method
Domains	Medical Reports
Datasets	MIMIC-CXR
Metrics	ROUGE
Human Evaluation	Readability, Accuracy, Completeness
Pipeline	External Knowledge, Unit Selection
Venue	ACL, 2020
Learning	supervised
Paper	Visit

Context Factors

Who is the target audience?
The summaries are for referring clinicians who have less time to review lengthy or intricate findings.

How will the summaries be used?
The summaries will be used to improve patients' well-being by automating the process of impression generation in radiology reporting, saving clinicians' read time, and decreasing fatigue. Clinicians would only need to proofread summaries or make minor edits.

What is the purpose of the summaries?
The authors are generating summaries of radiology reports to communicate critical findings to referring clinicians and save their read time.

Problems & Solutions

Generating IMPRESSION from FINDINGS can be subject to errors, which can have a negative impact on patients' well-being.
Automating the process of impression generation in radiology reporting would save clinicians' read time and decrease fatigue. The authors propose a novel seq2seq-based model to incorporate the salient clinical terms into the summarizer, which can improve the final IMPRESSION generation.

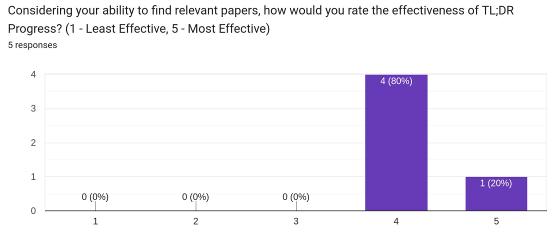
Clinicians mostly read the IMPRESSION as they have less time to review findings, particularly those that are lengthy or intricate.
The authors hypothesize that selecting the most significant clinical terms occurring in the FINDINGS and then incorporating them into the summarization would improve the final IMPRESSION generation. They further examine if refining FINDINGS word representations according to the identified clinical terms would result in improved IMPRESSION generation.

The effectiveness of the proposed model needs to be evaluated on different clinical datasets to assess its cross-organizational transferability.
The authors evaluate their model on two publicly available clinical datasets (MIMIC-CXR and OpenI) and show that it statistically significantly improves over competitive baselines.

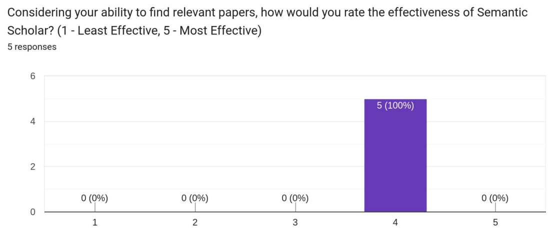
Previous studies have reported that augmenting the summarizer with entire ontology (i.e., clinical) terms within the FINDINGS can improve the content selection and summary generation to some noticeable extent.
The authors build on this previous work and propose to further improve the summarization process by selecting only the most significant clinical terms and incorporating them into the summarizer. They also use a sequence-tagger to learn the copying likelihood of a word as an indicator of its saliency in terms of forming IMPRESSION.

Figure 2: Indicative summary of a paper containing (1) an abstractive summary of the introduction, (2) manually annotated metadata attributes (details), (3) purpose of the summary encompassing the target audience, the downstream use, and the purpose, (4) claims and contributions of the paper.

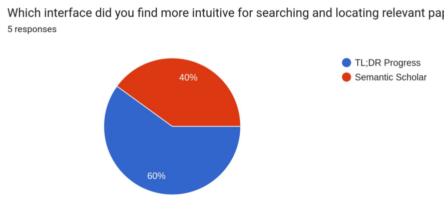
Figure 3: An overview of the figure browser which contains all the tables and figures pulled from the papers, accompanied by their captions.



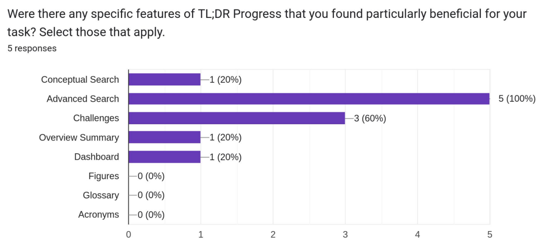
(a) Retrieval effectiveness of TL;DR PROGRESS.



(b) Retrieval effectiveness of Semantic Scholar.



(c) Preference of TL;DR PROGRESS over Semantic Scholar for literature review.



(d) Usefulness of features in TL;DR PROGRESS. Advanced search which allows for combining multiple facets for filtering papers is the most useful feature, followed by the enumerated list of challenges.

Figure 4: Evaluation results of the effectiveness and usefulness of TL;DR PROGRESS compared to Semantic Scholar.