GenGO: ACL Paper Explorer with Semantic Features

Sotaro Takeshita[1], Simone Paolo Ponzetto[1], Kai Eckert[2]
[1]Data and Web Science Group, University of Mannheim, Germany
[2]Mannheim University of Applied Sciences, Mannheim, Germany
{sotaro.takeshita, ponzetto}@uni-mannheim.de
k.eckert@hs-mannheim.de

Abstract

We present *GenGO*[1], a system for exploring papers published in ACL conferences. Paper data stored in our database is enriched with multi-aspect summaries, extracted named entities, a field of study label, and text embeddings by our data processing pipeline. These metadata are used in our web-based user interface to enable researchers to quickly find papers relevant to their interests, and grasp an overview of papers without reading full-text of papers. To keep *GenGO* available online as long as possible, we design *GenGO* to be simple and efficient to reduce maintenance and financial costs. In addition, the modularity of our data processing pipeline lets developers easily extend it to add new features. We make our code available to foster open development and transparency: https://gengo.sotaro.io/.[2]

1 Introduction

The rapidly growing number of scientific papers makes it difficult for researchers to keep up-to-date with the state of the literature (Bornmann and Mutz, 2015). Scholarly document processing (SDP) aims to support researchers with this challenge by building tools to process knowledge stored in research papers (Chandrasekaran et al., 2020) and has been increasingly drawing attention from academic and industrial communities. Together with the recent developments in natural language processing (NLP), a number of powerful technologies are now available in SDP. For instance, automatic scientific paper summarization systems provide short summaries encapsulating the essential points in a paper so that researchers can grasp the overview without reading its abstract or even full-text (Cachola et al., 2020; Yasunaga et al., 2019). A series of works on paper representation learning aims to obtain numerical representations of papers that can be used for information retrieval or recommendation (Ostendorff et al., 2022; Singh et al., 2023). Automatic information extraction systems allow repositories of papers to be organized in a structured manner (Jain et al., 2020; Viswanathan et al., 2021). There are also system demonstrations that implement user interfaces to the SDP technologies to make the aforementioned models available to scientists. Erera et al. (2019) introduce a system that lets users consume papers with automatically generated summaries. Hongwimol et al. (2021) is a web-based tool where users can explore different complex concepts and their relations by exploiting scientific knowledge graphs.

As the NLP community also faces a large number of publications (Bollmann et al., 2023) with dynamically changing trends (Schopf et al., 2023), there are systems which explicitly target this domain. The ACL Anthology (Bollmann et al., 2023) serves as an essential resource for the community by providing a repository of papers published in ACL conferences. Schäfer et al. (2011) introduce the ACL Anthology Searchbench that implements ACL Anthology a more fine-grained structured search. Ding et al. (2020) device a model-based semantic search in addition to the lexical search to provide a powerful literature search infrastructure. However, currently, there are no works which integrate different types of SDP technologies (e.g., summarization, information retrieval and extraction) in one system for NLP papers.

In this paper, we describe our system, dubbed *GenGO*, equipped with various semantic features to help researchers consume a large number of papers efficiently. Specifically, *GenGO* combines three types of SDP technologies. (1) **Multi-aspect summarization**: one paper in *GenGO* is coupled with four one-sentence summaries that summarize the Overview, Challenge, Approach, and Outcome of a paper. This enables researchers to quickly understand the overview of a paper from different

[1]gengo (言語) means 'language' in Japanese.
[2]Demo video: https://youtu.be/yYh9U5IVbIw

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 117–126
August 11-16, 2024

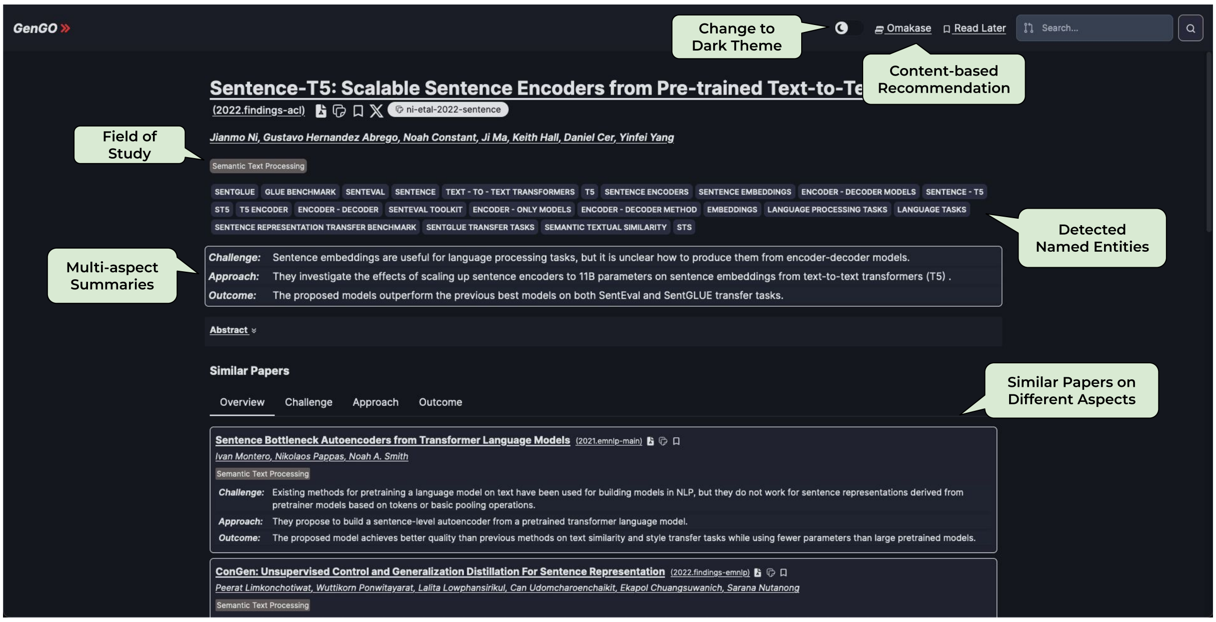

Figure 1: A screenshot exhibits how *GenGO* presents one paper. Each paper is complemented with multi-aspect summaries, field of study labels, named entities, and similar papers.

viewpoints. (2) **Semantic search**: our system can retrieve papers that are semantically relevant given user-provided queries. (3) **Information extraction**: *GenGO* automatically extracts technical terms and predicts the topic of a paper to enable fine-grained filtering and search. We design *GenGO* to be efficient and simple using cloud-based software so it requires low maintenance and financial cost, and currently achieves to index 30k+ papers.

2 Developments in SDP

In this section, we review existing papers and commercial products that target processing scholarly documents. The models and datasets used in the development of *GenGO* are marked with *.

2.1 Automatic text summarization

Text summarization systems for scientific papers have been an interest in NLP for decades. One of the early works, Paice (1980) proposes a system that aims to automatically produce abstract sections of papers. Recently, the performance of summarization models drastically improved with the emergence of neural network-based models (Rush et al., 2015; Gehrmann et al., 2018; Raffel et al., 2020a). While one straightforward approach for paper summarization would be to feed the texts of a paper to models to produce summaries, there are works that exploit unique characteristics in scholarly documents to improve their performance. Cohan and Goharian (2015) propose to use citation sentences to avoid inconsistency between summaries. Xiao and Carenini (2019) use global and local content in the paper to improve the summarization of long research papers. Additionally to the efforts on modelling, there are also various works that introduce language resources to develop and evaluate scientific paper summarization systems. Cohan et al. (2018) constructed a dataset by regarding the abstract sections in papers as summaries and the rest as inputs. Cachola et al. (2020)* collected the summaries written by authors and reviewers of a paper submitted to an open reviewing platform, which later extended into a cross-lingual variant by Takeshita et al. (2022, 2023). The most relevant summarization work to our system is the ACLSum dataset introduced by Takeshita et al. (2024a)*, in which 250 papers published in ACL-related conferences are annotated with abstractive and extractive summaries on three different aspects, namely Challenge, Approach, and Outcome.

2.2 Information retrieval

In recent NLP conferences, one proceeding can have more than 1k papers (EMNLP'23 main track has 1,047 papers), making efficient means of finding relevant papers essential. Bhagavatula et al. (2018) propose a method which takes a paper draft and finds relevant papers using a text embedding model. The aspect-based similarity model presented by Ostendorff et al. (2020) enables researchers to find papers by queries on different aspects. Cohan et al. (2020) introduce SPECTER, a scientific language model pre-trained using citation graphs, it produces high-performance paper embeddings. While it does not solely target scientific documents, all-MiniLM-L6-v2 (Reimers and Gurevych, 2019)* is a lightweight text embedding model which is trained on a number of sentence pair datasets including scientific texts using a contrastive objective function.

2.3 Text classification

Scientific documents often do not come with rich metadata that can be used to form a structured data repository. Jurgens et al. (2018) introduce a citation intent classification dataset for the NLP domain. This enables richer citation graphs of scientific papers. Schopf et al. (2023)* present a semi-automatically and manually annotated dataset for the field of study classification, respectively for training and evaluation purposes. The classification model trained on the dataset enables automatic analysis of how the research trend in NLP changes over time.

2.4 Information extraction

The ScienceIE task (Augenstein et al., 2017) aims to develop methods that extract keyphrases from scientific papers, which can be further used in retrieval systems. Jain et al. (2020)* presents, the SciREX dataset, a document-level information extraction dataset based on machine learning papers. Viswanathan et al. (2021) propose to use citation graph to perform information extraction from scientific documents.

2.5 Applications

While the aforementioned works present models or language resources, they require a user interface to be delivered to researchers. There are a number of academic system demonstrations and commercial services that are designed to fill this gap. The IBM Science Summarizer presents research papers coupled with automatically produced extractive summaries (Erera et al., 2019). Gökçe et al. (2020) present an online editor where users can explore the existing papers while writing a manuscript. Hongwimol et al. (2021) introduce a web-based interactive tool which visualizes explanations and relationships between concepts using graph-structured data. A system introduced by Gu and Hahnloser (2023) enables users to find relevant papers with generated or extracted summaries given user-provided context and keywords. In addition to the system demonstrations which are often done in academic institutions such as universities, there are also software developed more intensively including commercial products. Semantic Scholar (Kinney et al., 2023) is a search engine that also provides paper summarization and recommendations. Zeta Alpha (Fadaee et al., 2020) and Elicit[3] provide search features and also chat-based interfaces based on recent large language models.

To the best of our knowledge, there are no system demonstrations that combine automatic summarization, information retrieval and extraction methods in NLP papers. While web-based applications developed by private organizations mentioned such as Semantic Scholar provide similar functionalities, the closed nature of these software hinders transparency of how and which models are being used.

3 System requirements

One pragmatic challenge for a system demonstration project is to keep the system up and running, especially when it is maintained by a few developers with a limited budget. Indeed, 10 out of 27 web-based system demonstrations presented at ACL 2023 (held in July) are offline after less than one year (at the time of writing, March 2024). We speculate that this is due to (1) maintenance effort and (2) financial costs as mentioned by Bollmann et al. (2023). To achieve a long lifespan of our system, we design *GenGO* to minimize the aforementioned two factors. (1) Maintenance effort: we minimize the number of servers to be taken care of. Specifically, our system does not rely on custom servers that often demand intensive system maintenance. Instead, we opt for a server-less architecture by making use of managed cloud-based services. To reduce (2) Financial costs: we designed our pipeline not to make any online inferences that require GPUs. Our data processing pipeline can run fully offline, process each document once, and store the results in a database. The only online inference *GenGO* makes is when to compute a text embedding of a user-provided query. Instead of hosting a server to compute embeddings for each user request, we use a lightweight text embedding model and perform the inference on the user's device. Because of these design decisions, *GenGO* is currently running without any financial cost from

[3] https://elicit.com/

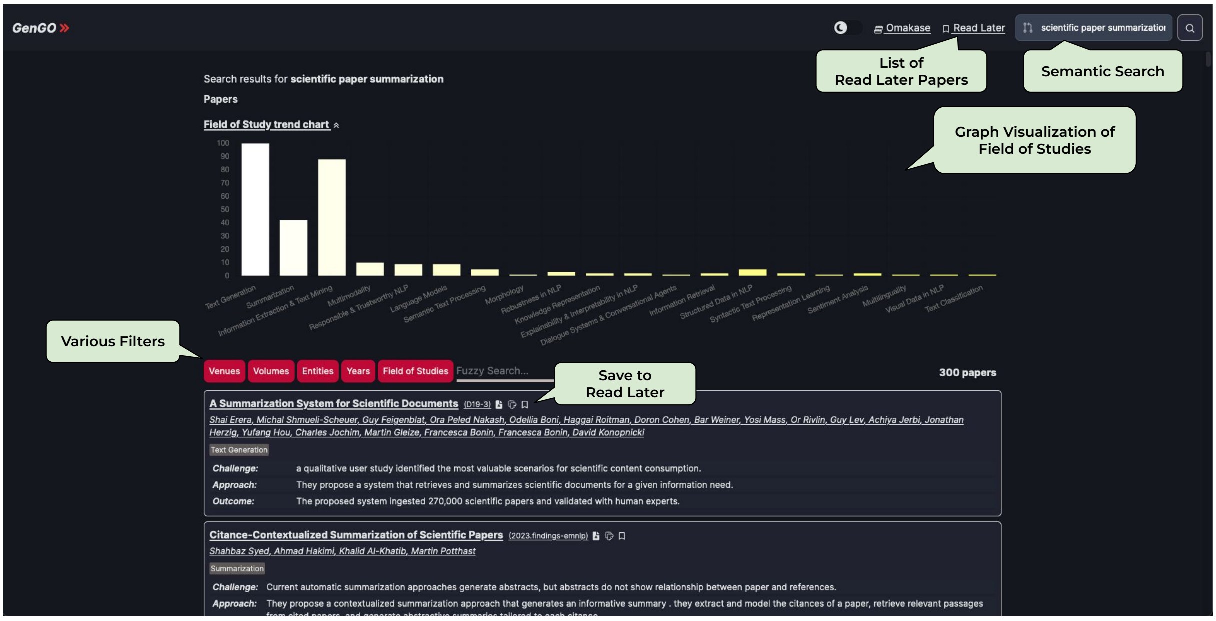

Figure 2: A screenshot showing how *GenGO* presents a list of papers retrieved by its semantic search functionality. Users can apply various filters to the list (e.g., venue, named entities, field of studies, etc...) to efficiently find relevant papers.

our disposal while indexing more than 30k papers. By using the cost estimator provided by Qdrant[4], we estimate that when the number of indexed papers increases by ten times, it would require around 10 Euro per month.

4 *GenGO*

In this section, we describe the main features of *GenGO* (§4.1) and a system overview (§4.2). Table 5 in Appendix lists all the models and datasets with external links.

4.1 Main features

Multi-aspect summarization. We complement each paper with automatically generated one-sentence summaries on four different aspects, namely Overview (Generic overall summary), Challenge (The current situation faced by the researcher), Approach (How they intend to carry out investigation), and Outcome (Overall conclusion). Refer Fisas et al. (2015) for the detailed definitions. This enables users to quickly understand these aspects of a paper without reading its full-text or even the abstract.

We use two summarization datasets and two transformer-based models. To generate Overview summaries, we use a BART-large model (Lewis et al., 2020) fine-tuned on the SciTLDR dataset

Dataset	R-1	R-2	R-L	R-K
SciTLDR - Overview	43.9	22.3	36.6	41.36
ACLSum - Challenge	18.9	2.5	13.6	50.27
ACLSum - Approach	44.8	22.4	38.4	55.85
ACLSum - Outcome	42.3	21.7	35.0	37.14

Table 1: Performance of summarization models.

(Cachola et al., 2020). To generate multi-aspect summaries, we fine-tune one T5-base (Raffel et al., 2020b) on each aspect using the ACLSum dataset (Takeshita et al., 2024a). Table 1 shows the performance of each model evaluated on corresponding datasets by ROUGE F1 (Lin, 2004) and its keyword-focused extension, ROUGE-K (Takeshita et al., 2024b). In *GenGO*, we follow the corresponding papers and use the abstract for the overview summarization, and the abstract, introduction and conclusion sections for multi-aspect summarization as inputs.

Semantic search and recommendation. *GenGO* can retrieve semantically relevant documents given a user-provided query, and provide two types of paper recommendations. (1) Content-based recommendation: where each paper is presented with its similar papers in four different aspects (Overview, Challenge, Approach, and Outcome). (2) History-based recommendation: papers relevant to the user's reading history.

[4] https://cloud.qdrant.io/calculator

Dataset	NDCG@10	MAP@10
SciFact	0.645	0.596
SciDocs	0.216	0.129

Table 2: Performance of retrieval model

Dataset	Precision	Recall	F1
NLP Taxonomy	92.46	93.99	92.21

Table 3: Performance of field of Study classification models.

We use the all-MiniLM-L6-v2 model[5] as our underlying text encoder model. To perform the semantic search, we compute cosine similarities between an embedding of a user-provided query and embeddings of all of the indexed papers and return the 300 most similar papers. To compute paper embeddings, we use paper titles and abstracts as inputs to the text encoder. For content-based recommendation on the Overview aspect, we use an embedding of the opened paper as a query vector. For the other three aspects, we use the generated summaries (refer to Section 4.1) on the corresponding aspect as a query text. To provide reading history-based recommendations, we use an average vector from the papers in the user's reading history as a query vector. Table 2 shows the performances of the all-MiniLM-L6-v2 model on two retrieval benchmark datasets from the scientific domain, SciDocs (Cohan et al., 2020) and SciFACT (Wadden et al., 2020). To respect the user's privacy, we store the reading history information using the localStorage property of the user's web browser.

Field of study filtering. We predict the field of study labels (g.g., Text Generation) for each paper to enable users to apply filtering by the topics of their interests.

We use the model published by Schopf et al. (2023). Its underlying model architecture is BERT-base (Devlin et al., 2019) and parameters are initialized using SPECTER model (Cohan et al., 2020) and the authors fine-tune the model using their classification dataset. We present the performance of the model reported in the original paper in Table 3. In *GenGO*, we follow the strategy from the original paper and use titles and abstracts as inputs.

Field of study visualizer. Every time a user sees a list of papers in *GenGO*, it is complemented by a bar-graph visualization of the field of study labels. Users can grasp topic popularity in a conference proceeding or a publication list of a researcher. Figure 2 shows an example over a list of search results.

[5]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Type	Precision	Recall	F1
Method	72.18	70.43	71.29
Task	65.58	53.41	58.87
Dataset	50.00	53.26	51.58
Metric	72.17	60.27	65.68

Table 4: Performance of named entity recognition model.

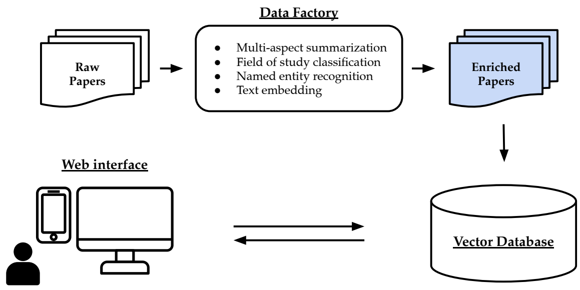

Figure 3: A system overview.

Named entity filtering. We attach named entities extracted from the paper content as metadata to achieve fine-grained filtering.

We train the transition-based parser model implemented in spaCy (Honnibal et al., 2020) on the SciRex dataset (Jain et al., 2020) to extract entities from titles and abstracts. Table 4 shows the results of the test split of SciRex.

Save to read later. While this does not require any NLP-based methodologies, since we intend our platform to be used to digest many papers quickly, we implemented this feature so that a user can "save" a paper while skimming over a long list of papers for later to study in detail.

We use the localStorage property of web browsers to save the paper information. This enables to keep the data locally on the user's device instead of external servers.

4.2 System description

GenGO stands on three components, *data factory*, *vector database*, and *web-based user interface* as depicted in Figure 3. In the remainder of this section, we describe each component in detail.

Data factory. This module prepares the published research articles from the ACL anthology to be indexed in our database. Concretely, it first downloads a paper in the PDF format from ACL Anthology, runs Grobid (Lope, 2008–2023) to extract the full-text, and segments the text into sentences using Spacy[6]. After having the structure data of a paper, we run models described in Section 4.1 for each paper. After this enrichment process, as the final step in the data factory, it uploads the resulting data to the vector database, described in the following paragraph. Currently, we opt for relatively lightweight models so that this pipeline can run even on a consumer laptop without GPUs. For instance, it takes 4 hours to process 700+ papers from ACL 2020 on a MacBook Air M2.

Vector database. Instead of traditional relational databases, we use a vector database to host our enriched paper data. Vector databases can store documents with metadata and use numerical vectors for indexing to achieve vector-based searching given a query vector. We can achieve semantic search by using semantic embedding vectors of papers' texts and user-provided queries. Among several available vector database implementations, we opt for Qdrant[7]. This is due to that Qdrant is an open-source software making the retrieval mechanism transparent, followed by more minor technical reasons such as the support of multiple vectors for one data point enabling *GenGO* to retrieve papers on different aspects, and fast search speed compared to the other implementations. We host our Qdrant instance using the managed solution provided by a company which leads the development of Qdrant. At the time of writing, *GenGO* indexes publications from nine major ACL conferences from 2018 to 2023, resulting in more than 30k papers and running free of charge. This database is the only component in our system that requires financial costs when scaling up the system. In the current pricing options, we estimate that the Qdrant cloud solution would require c.a. 10 Euro if we increase the number of indexed papers by 10 times.

Web-based user interface. Like most modern web applications, *GenGO* is developed using a JavaScript framework to achieve interactivity. Specifically, we use Svelte[8]. To achieve responsive design, we use tailwindcss[9] as our CSS framework, and host our frontend application using Vercel[10].

[6]https://spacy.io/
[7]https://qdrant.tech/
[8]https://svelte.dev/
[9]https://tailwindcss.com/
[10]https://vercel.com/

5 Limitation of current system

Because the metadata used to build the *GenGO* is from the ACL Anthology project, our system inherits its challenges such as the disambiguation of author names (Bollmann et al., 2023). In addition, there are several limitations unique to our system. (1) Speed and the number of indexed papers: Compared to similar services developed by large institutions such as Semantic Scholar, our system relies on limited computational resources. This currently results in slower response time especially when showing a long list of papers, and also to compensate for the limited size of the data storage, we only index approximately 30% of papers from the whole ACL Anthology repository. We are also hindered in applying the most powerful models available in our data factory due to the lack of sufficient computational resources. (2) Summarization quality: Hallucination in text generation remains an open challenge (Dong et al., 2022; Koh et al., 2022; Ma et al., 2023). During the development, we observed cases where the generated summaries do not convey the full information of a paper or contain information inconsistent with the corresponding paper. (3) Retrieval bias: While the semantic search approach used in our system can work with paraphrases or synonyms, due to the 'blackboxness' of model-based encoders, it remains unclear whether such models have a bias towards retrieving certain styles of texts (MacAvaney et al., 2022).

6 Conclusion

In this paper, we described *GenGO*, a system to explore ACL papers with various types of semantically empowered functionalities. Our system enables NLP researchers to quickly find relevant papers using semantic search and various fine-grained filters, and grasp paper overviews by reading multi-aspect summaries. *GenGO* also provides utility features such as paper recommendations or 'read it later' to further enhance user experience.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–

555, Vancouver, Canada. Association for Computational Linguistics.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the ACL Anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore, Singapore. Empirical Methods in Natural Language Processing.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, 66(11):2215–2222. Publisher: Wiley.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Anita de Waard, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert M. Patton, and Michal Shmueli-Scheuer, editors. 2020. *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online.

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.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shane Ding, Edwin Zhang, and Jimmy Lin. 2020. Cydex: Neural search infrastructure for the scholarly literature. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 168–173, Online. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. 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): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.

Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp, and Jakub Zavrel. 2020. A new neural search and insights platform for navigating and organizing AI research. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 207–213, Online. Association for Computational Linguistics.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Onur Gökçe, Jonathan Prada, Nikola I. Nikolov, Nianlong Gu, and Richard H.R. Hahnloser. 2020. Embedding-based scientific literature discovery in a text editor application. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics: System Demonstrations, pages 320–326, Online. Association for Computational Linguistics.

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 (Volume 3: System Demonstrations)*, pages 235–246, Toronto, Canada. Association for Computational Linguistics.

Pollawat Hongwimol, Peeranuth Kehasukcharoen, Pasit Laohawarutchai, Piyawat Lertvittayakumjorn, Aik Beng Ng, Zhangsheng Lai, Timothy Liu, and Peerapon Vateekul. 2021. ESRA: Explainable scientific research assistant. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 114–121, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Patrice Lope. 2008–2023. Grobid. `https://github.com/kermitt2/grobid`.

Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A benchmark of unfaithful minimal pairs for meta-evaluation of faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12788–12812, Toronto, Canada. Association for Computational Linguistics.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, 10:224–239.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206, Barcelona, Spain (Online). International Committee on Computational Linguistics.

C. D. Paice. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR '80, page 172–191, GBR. Butterworth & Co.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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)*, pages 3982–3992, Hong Kong, China. 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*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 7–13, Portland, Oregon. Association for Computational Linguistics.

Tim Schopf, Karim Arabi, and Florian Matthes. 2023. Exploring the landscape of natural language processing research. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1034–1045, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-scitldr: cross-lingual extreme summarization of scholarly documents. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2023. Cross-lingual extreme summarization of scholarly documents. *International Journal on Digital Libraries*, pages 1–23.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024a. ACLSum: A New Dataset for Aspect-based Summarization of Scientific Publications. ArXiv:2403.05303, To appear at NAACL 2024.

Sotaro Takeshita, Simone Paolo Ponzetto, and Kai Eckert. 2024b. ROUGE-K: Do Your Summaries Have Keywords? ArXiv:2403.05186.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the citation graph for scientific information extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–731, Online. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. 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)*, pages 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

A Appendix

Name	Task	URL
BART$_{\text{LARGE}}$	Summarization (Overview)	https://huggingface.co/sobamchan/bart-large-scitldr
T5$_{\text{BASE}}$	Summarization (Challenge)	https://huggingface.co/sobamchan/t5-base-aclsum-challenge-nofilter
T5$_{\text{BASE}}$	Summarization (Approach)	https://huggingface.co/sobamchan/t5-base-aclsum-approach-nofilter
T5$_{\text{BASE}}$	Summarization (Outcome)	https://huggingface.co/sobamchan/t5-base-aclsum-outcome-nofilter
SciTLDR	Summarization (Overview)	https://github.com/allenai/scitldr
all-MiniLM-L6-v2	Retrieval	https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
Taxonomy classifier	Field of study classification	https://huggingface.co/TimSchopf/nlp_taxonomy_classifier
SciREX	NER	https://github.com/allenai/SciREX/

Table 5: A list of models and datasets with external URLs used in *GenGO*.