Anubhav **Jangra**

Computer Science Ph.D. student, Columbia University

🌐 Homepage   @ anubhav@cs.columbia.edu   Github   🎓 Google Scholar

Education

Present
Aug 2023 | **Columbia University** — **Current GPA: 3.93/4.0**
CS PhD (Supervisor: Smaranda Muresan)

Jun 2021
Jul 2017 | **Indian Institute of Technology, Patna** — **GPA: 8.82/10, Major GPA: 9.12/10**
Bachelor of Technology in Computer Science and Engineering

Research Experience

Aug 2024 – May 2024 | **Microsoft Research & Office of Applied Research** — **Redmond, USA**

JEM (Joint E+D & MSR) Research Intern | Advisors: Sujay Jauhar, Bahar Sarrafzadeh, Adrian de Wynter

Contributed to the development of style personalization for MS Word CoPilot, enhancing draft and rewrite features to reflect users' authentic voices.

Devised a human-centric evaluation process, developed an automatic evaluation mechanism for low-resource style evaluation and launched the feature for internal dogfooding.

Jul 2023 – Jul 2021 | **Google Research | Advertising Sciences Team** [🌐] — **Bangalore, India**

Pre-Doctoral Researcher | Advisor: Aravindan Raghuveer

Explored NLG techniques for creative advertisement generation. Investigating several research areas like text style transfer, data-to-text generation, automatic code generation, semantic representations etc.

Aug 2020 – Jul 2020 | **GREYC Lab, ENSI-CAEN** [🌐] — **Remote / Caen, France**

Research Intern | Advisor: Gaël Dias

Extended patch-based lexical semantic identification frameworks to a multi-modal setting. Developed the dataset and conducted the pilot studies of the project. [*ACM MM'22*]

Jun 2019 – May 2019 | **Graduate School of Informatics, Kyoto University** [🌐] — **Kyoto, Japan**

Research Intern | Advisor: Adam Jatowt

Explored various unsupervised optimization techniques to develop multi-modal summarization systems that generate text-image-audio-video summaries.

Publications

US=under submission, P=Preprints, C=Conference, B=Book, SP=Short Paper, J=Journal

Selected Works

[P] **Navigating the Landscape of Hint Generation Research: From the Past to the Future** [🔗]
Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, Smaranda Muresan
ArXiV 2404.04728 **[ArXiV, 2024]**

[C] **Large Scale Multi-modal Multi-lingual Summarization Dataset** [🔗]
Yash Verma*, Anubhav Jangra*, Raghvendra Verma, Sriparna Saha
The 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia **[EACL'23]**

[C] **T-STAR: Truthful Style Transfer using AMR Graph as Intermediate Representation** [🔗]
Anubhav Jangra*, Preksha Nema*, Aravindan Raghuveer
The 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, UAE **[EMNLP'22]**

[J] **A Survey on Multi-modal Summarization** [🔗]
Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, Mohammed Hasanuzzaman,
ACM Computing Surveys **[ACM CSUR'23]**

[C] **WIDAR - Weighted Input Document Augmented ROUGE** [🔗]
Raghav Jain*, Vaibhav Mavi*, Anubhav Jangra*, Sriparna Saha
44th European Conference on Information Retrieval, Stavanger, Norway **[ECIR'22]**

[C] **Multi-modal Supplementary Complementary Summarization using Multi-Objective Optimization** [🔗]
Anubhav Jangra, Sriparna Saha, Adam Jatowt, Mohammed Hasanuzzaman
44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual **[SIGIR'21]**

[C] **Semantic Extractor Paraphraser based Abstractive Summarization** [🔗]
Anubhav Jangra*, Raghav Jain*, Vaibhav Mavi*, Sriparna Saha, Pushpak Bhattacharyya,
17th International Conference on Natural Language Processing, Patna, India **[ICON'20]**

[SP] **Multi-Modal Summary Generation using Multi-objective Optimization** [🔗]
Anubhav Jangra, Sriparna Saha, Adam Jatowt, Mohammed Hasanuzzman,
43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China **[SIGIR'20]**

[SP] **Text-Image-Video Summary Generation using Joint Integer Linear Programming** [🔗]
Anubhav Jangra, Adam Jatowt, Mohammed Hasanuzzman, Sriparna Saha,
42nd European Conference on Information Retrieval, Lisbon, Portugal **[ECIR'20]**

Other Works

[B] **Multi-hop Question Answering** [🔗]
Vaibhav Mavi, Anubhav Jangra, Adam Jatowt
Foundations and Trends® in Information Retrieval Vol. 17 Issue 5 **[FnTs, 2024]**

[C] **TriviaHG: A Dataset for Automatic Hint Generation for Factoid Questions**
Jamshid Mozafari, Anubhav Jangra, Adam Jatowt
The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Wash. D.C., USA **[SIGIR'24]**

[C] **Can Multimodal Pointer Generator Transformers produce topically relevant summaries?** [🔗]
Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, Adam Jatowt,
2023 International Joint Conference on Neural Networks (IJCNN) **[IJCNN'23]**

[C] **Topic-aware Multimodal Summarization** [🔗]
Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, Adam Jatowt,
2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics **[Findings in AACL'22]**

[P] **A Survey on Medical Document Summarization** [🔗]
Raghav Jain, Anubhav Jangra, Adam Jatowt, Sriparna Saha
ArXiv 2212.01669 **[ArXiv, 2022]**

[C] **Combining Vision Language Representations for Patch-based Identification of Lexico-Semantic Relations** [🔗]
Prince Jha, Gaël Dias, Alexis Lechervy, José G Moreno, Anubhav Jangra, Sebastião Pais, Sriparna Saha
30th ACM International Conference on Multimedia, Lisbon, Portugal **[ACM MM'22]**

[C] **MAKED: Multi-lingual Automatic Keyword Extraction Dataset** [🔗]
Yash Verma, Anubhav Jangra, Sriparna Saha, Adam Jatowt, Dwaipayan Roy
13th Conference on Language Resources and Evaluation **[LREC'22]**

[J] **Identifying Complaints based on Semi-Supervised Mincuts** [🔗]
Apoorva Singh, Sriparna Saha, Mohammed Hasanuzzaman, Anubhav Jangra
Elsevier's Expert Systems with Applications, Volume 186, 2021 **[ESWA'21]**

[J] **Extractive Single Document Summarization using Multiobjective Optimization: Exploring Self-organized Differential Evolution, Grey Wolf Optimizer and Water Cycle Algorithm** [🔗]
Naveen Saini, Sriparna Saha, Anubhav Jangra, Pushpak Bhattacharyya,
Elsevier's Knowledge Based Systems, 2018 **[KBS'18]**

Selected Research Projects

Hint Generation Sept'23 - Present
Advisor: Smaranda Muresan

- Exploring the hint generation frameworks and human-centered evaluation strategies to improve the engagement and learnings to augment student learning experience. (ongoing)
- Wrote the first of it's kind interdisciplinary survey on automatic hint generation. (under submission)

Text Style Transfer Sept'21 - Nov'22
Advisors: Aravindan Raghuveer, Preksha Nema

- Developed an AMR graph based framework to improve content preservation in generation. [*EMNLP'22*]
- Proposed method significantly out- performs state-of-the-art techniques by achieving on an average **15.2% higher content preservation** with negligible loss ($\sim$3%) in style accuracy.
- Performed human evaluations to illustrate that T-STAR has **50% lesser hallucinations** compared to SoTA TST models.

Multi-modal summarization Jan'19 - Jul'21

Advisors: Sriparna Saha, Adam Jatowt, Mohammed Hasanuzzaman

- Developed and implemented various systems using optimization techniques like integer linear programming, differential evolution, grey wolf optimizer etc. to solve text, image, and video summary generation. [*ECIR'20, SIGIR'20, SIGIR'21*]
- Formally defined the complementary/supplementary enhanced multi-modal summaries, and achieved a new state-of-the-art on unsupervised MMS, surpassing the predecessor by almost **twice as better ROUGE-2 scores**. [*SIGIR'21*]
- Wrote the first ever literature survey on multi-modal summarization. [*ACM Computing Surveys'23*]
- First work towards topic-aware multi-modal news summarization. [*Findings AACL'22*]
- Curated large-scale multi-modal multi-lingual summarization corpus spanning over 20 languages. [*EACL'23*]

Automatic Text Summarization Jul'19 - Dec'21

Advisors: Sriparna Saha, Pushpak Bhattacharyya

- **Extractive Summarization.** Utilized nature-inspired algorithms like Differential Evolution, Grey Wolf Optimizer, Water Cycle Algorithm etc. in a multi-objective optimization framework to generate extractive summaries. [*KBS'18*]
- **Abstractive Summarization.** Proposed an RL-based 'extractor-abstractor' framework to outperform its predecessors by a margin of **0.5 ROUGE-1, 0.4 ROUGE-2, 1 METEOR, and 0.9 WMS scores**. A knowledge discovery that seq2seq networks like PGN model implicitly extract and paraphrases sentences was brought to light through this work. [*ICON'20*]
- **Evaluation Metrics.** Proposed WIDAR, a ROUGE-based evaluation metric that evaluates generated summary by taking into account both the reference summary and input document. WIDAR correlates better than ROUGE by **26%, 76%, 82%, and 15% in coherence, consistency, fluency, and relevance** on human judgement scores provided in the SummEval dataset. It was able to obtain comparable results with the SOTA while requiring $\sim\frac{1}{64}^{th}$ of computational time. [*ECIR'22*]

Other Experiences

Jul 2021
Jun 2021
IBM **Remote / Chennai, India**
Global Research Mentee | Advisor: Ganesan Narayanasamy
Developed the project framework for *Health Care App*, that uses knowledge graphs and named-entity recognition to help users self-diagnose themselves.

Mar 2021
Dec 2020
Huawei Technologies Co., Ltd **Remote**
Project Member | Advisor: Sriparna Saha
Developed a Proof of Concept (POC) for the task of automatic tagline generation and product description using existing neural summarization systems for the upcoming collaborative project of IIT Patna and Huawei.

Jan 2020
Dec 2019
TCS Innovation Lab [🌐] **Kolkata, India**
Research Intern | Advisor: Arijit Ukil
Investigated generative modeling to tackle the insufficiency of data in time-series signal classification.

Jan 2019
Dec 2018
CFILT Lab, IIT Bombay [🌐] **Mumbai, India**
Research Intern | Advisor: Pushpak Bhattacharyya
Examined Unsupervised NMT for distant language pairs (Indo-Aryan languages) using attention based seq2seq models.

Jul 2021
Jul 2018
AI-NLP-ML Lab, IIT Patna [🌐] **Patna, India**
Undergraduate Research Scholar | Advisor: Sriparna Saha
Worked extensively in the area of summarization (*e.g.,* multi-modal summarization, extractive and abstractive text summarization), complaint mining and multi-label classification.

Academic Service

PC Member	Coling 2025, LREC-Coling 2024, Text2Story Workshop (ECIR 2023, 2024), IACT - International Workshop on Implicit Author Characterization from Texts for Search and Retrieval (SIGIR 2023)
Reviewer (Conference)	ACL ARR (*since Dec 2023*), CIKM 2023, ACL 2023, EMNLP 2022
Reviewer (Journals)	ACM Computing Surveys (*since Jan 2021*), ACM TALLIP (*since May 2020*), Applied Artificial Intelligence (*since Oct 2021*), and IEEE Transactions on Computational Social Systems (*since Jan 2022*), Expert Systems with Applications (*since Sept 2022*), Engineering Applications of Artificial Intelligence (*since Feb 2022*), IEEE Internet Computing (*since Feb 2022*)
Secondary Reviewer	AAAI 2020, EACL 2021, ACL 2021, EMNLP 2021, CIKM 2021, KDD 2022, and WebConf 2022

Mentor Mentored three undergraduate interns, two masters student researchers, and one undergraduate student researcher as part of the AI-NLP-ML lab, IIT Patna.

Volunteer Volunteered as a reviewer in Google's CS Research Mentorship program to help review applicants from historically marginalized communities for the mentorship program. Reviewer for PhD Pre-Application Review (PAR) program at Columbia university.

Community Service Creator and organizer of CARE program—a Community for AI Research and Education to guide students and early-stage researchers through their research-related queries.

Honours and Awards

Google Research AI summer school, 2020 [🌐] One of the 50 participants out of 1000+ applicants in the NLU track.

MSU-IITR-IISc course and workshop [🌐] Attended a short term course and workshop on "Pragmatic Optimization for Practical Problem Solving" conducted by Michigan State university, IIT Roorkee and IISc Bangalore, limited to 40 students.

IIT JEE Ranked in National Top 0.2% (amongst 1,400,000 candidates) in JEE Mains 2017 and Top 1.5% (amongst 2,00,000 candidates) in IIT-JEE Advanced 2017.

Talks

- **Abstract Meaning Representation (AMR) Graphs at work!** - AI-NLP-ML Lab, IIT Patna Oct 2022
- **Automatic Text Summarization** PyData Patna Conference Dec 2020

Teaching and Leadership Roles

Barnard College at Columbia University *Teaching Assistant* Sep 2024-Dec 2024

- Teaching Assistant for BC3997 (Natural Language Processing) course taught by Dr. Smaranda Muresan.

Google DSC IIT Patna, Patna, India *ML Department Lead* 2019-2020

- Supervised three projects and gave lectures on Machine Learning theory and its applications.

Univeristy of Innsbruck, Austria *Teaching Assistant* Jun 2020

- Part-time Teaching Assistant in the course 2021S703836 VU (Natural Language Processing). Prepared lectures on automatic summarization.

Miscellaneous

- **AnthroKrishi project at Google**: Conducted semi-structured interviews with farmers on understanding motivations and barriers for changing farming practices for carbon sequestration.
- Invited to the FODO.AI podcast to share my research journey.
- Outside of work, I love to create origami and write in calligraphy.
- I have gracefully failed at learning violin in the past.