Kapil Thadani

+1 · 917 · 573 · 8173 • kapil@cs.columbia.edu • www.cs.columbia.edu/∼kapil • 770 Broadway, New York, NY 10003

EXPERIENCE

Yahoo Research — New York, NY

- *Principal Research Scientist* — *Dec 2024—present*
- *Senior Research Scientist* — *Sep 2018—Nov 2024*
- *Research Scientist* — *Dec 2014—Aug 2018*

Large language models for content summarization, continuous tagging and classifier bootstrapping. News recommendation through storyline and question generation across related articles. Scalable detection of salient entities in news articles. Transfer learning techniques for taxonomic and few-shot multi-label text classification. Exploration strategies for unbiased news recommendation. Deep reinforcement learning for online ranking of articles and comments. Generation of videos from news articles with automated summarization, compression and coreference resolution. Deep neural network architectures and user interest models for personalized stream ranking. Statistical models for summarization of news articles and search results.

Columbia University School of Engineering and Applied Science — New York, NY

- *Adjunct Professor*

  Deep Learning for Computer Vision, Speech and Language — *Spring 2017, Fall 2018*

- *Teaching Assistant*

  Statistical Methods for Natural Language Processing — *Spring 2010*
  Artificial Intelligence — *Summer 2006, Fall 2007*
  Search Engine Technology — *Spring 2007*

Google Inc — New York, NY

- *Software Engineering Intern (Research)* — *May 2011—Aug 2011*

  Automatic closed captioning of YouTube videos. Developed supervised & unsupervised strategies to improve transcription accuracy by adapting an ensemble of topic-specific language models to each video.

- *Software Engineering Intern (Search Quality)* — *May 2009—Aug 2009*

  Geographic location disambiguation within a named entity pipeline. Developed a bootstrapping approach to improve identification and labeling of location mentions using the surrounding context.

Columbia University Natural Language Processing Group — New York, NY

- *Graduate Research Assistant* — *Jul 2011—Oct 2014*

  Part of Columbia's team in the IARPA-sponsored Foresight & Understanding from Scientific Exposition (FUSE) project. Working on identification, definition and summarization of scientific concepts.

- *Graduate Research Assistant* — *Jan 2011—Jul 2011*

  Part of IBM's team for phase 5 of the DARPA-sponsored Global Autonomous Language Exploitation (GALE) program. Developed question-answering systems to describe news events and reactions.

- *Graduate Research Assistant* — *Jan 2006—Jun 2009*

  Part of SRI International's distillation effort for GALE phases 1 through 3. Created systems for automated redundancy removal and semantic role labeling for disfluent text.

Columbia University Center for Computational Learning Systems New York, NY

· *Graduate Research Assistant* *Sept 2006—Dec 2006*

Analysis of the relationship between patterns of service requests and major electrical outages in New York City in collaboration with Con Edison of New York.

EDUCATION

Columbia University · Graduate School of Arts & Sciences New York, NY
PhD in Computer Science (Natural Language Processing) *May 2015*
Thesis: *Multi-structured Models for Transforming and Aligning Text*

Columbia University · School of Engineering & Applied Science New York, NY
MS in Computer Science (Machine Learning / Thesis track) *May 2007*
GPA 4.17; Thesis: *Decreasing Textual Redundancy*

University of Mumbai · Thadomal Shahani Engineering College Mumbai, India
Bachelor of Engineering (Information Technology) *May 2005*

PUBLICATIONS

· Bhavin Jawade, Joao Soares, Kapil Thadani, Deen Dayal Mohan, Erfan Eshratifar, Jack Culpepper, Paloma de Juan, Srirangaraj Setlur and Venu Govindaraju. SCOT: Self-supervised Contrastive Pretraining for Zero-Shot Compositional Image Retrieval. In *Proceedings of the 25th IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025).*

· E. Ye, X. Bai, N. O'Hare, E. Asgarieh, K. Thadani, F. Perez-Sorrosal, S. Adiga. Multilingual Taxonomic Web Page Classification through Ensemble Knowledge Distillation. In *IEEE Transactions on Knowledge and Data Engineering (TKDE 2024).*

· A. Eshratifar, J. Soares, K. Thadani, S. Mishra, M. Kuznetsov, Y. Ku and P. de Juan. Salient Object-Aware Background Generation using Text-Guided Diffusion Models. In *Proceedings of the Workshop on Generative Models for Computer Vision at CVPR 2024.*

· Y. Zhang, S. Herdade, K. Thadani, E. Dodds, J. Culpepper, Y. Ku. Unifying Margin-Based Softmax Losses in Face Recognition. In *Proceedings of the 23rd IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023).*

· E. Ye, X. Bai, N. O'Hare, E. Asgarieh, K. Thadani, F. Perez-Sorrosal, S. Adiga. Multilingual Taxonomic Web Page Classification for Contextual Targeting at Yahoo. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022).*

· M. Verma, K. Thadani, S. Mishra. Powering COVID-19 Community Q&A with Curated Side Information. In *Proceedings of the Workshop on Knowledge Injection in Neural Networks at CIKM 2021.*

· A. Gupta, K. Thadani, N. O'Hare. Effective Few-Shot Classification with Transfer Learning. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).*

· S. Mishra, M. Verma, Y. Zhou, K. Thadani, W. Wang. Learning to Create Better Ads: Generation and Ranking Approaches for Ad Creative Refinement. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020).*

· M. Yoder, Q. Shen, Y. Wang, A. Coda, Y. Jang, Y. Song, K. Thadani, C. Rosé. Phans, Stans and Cishets: Self-Presentation Effects on Content Propagation in Tumblr. In *Proceedings of the 12th International ACM Web Science Conference (WebSci 2020).*

· N. Zalmout, A. Pappu, K. Thadani. Unsupervised Neologism Normalization using Embedding Space Mapping. In *Proceedings of the Workshop on Noisy User-Generated Text at EMNLP 2019.*

· A. Pappu, R. Blanco, Y. Mehdad, A. Stent and K. Thadani. Lightweight Multilingual Entity Extraction and Linking. In *Proceedings of the 10th ACM International Conference on Web search and Data Mining (WSDM 2017).*

· J. Li, K. Thadani and A. Stent. The Role of Discourse Units in Near-Extractive Summarization. In *Proceedings of the 17th Annual SIGdial Meeting on Dialogue and Discourse (SIGDIAL 2016).*

· Y. Mehdad, K. Thadani, D. Radev, A. Stent, Y. Billawala and K. Buchner. Extractive Summarization under Strict Length Constraints. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).*

· K. McKeown, H. Daumé III, S. Chaturvedi, J. Paparrizos, K. Thadani, P. Barrio, O. Biran, S. Bothe, M. Collins, K. Fleischmann, L. Gravano, R. Jha, B. King, K. McInerney, T. Moon, D .O'Seaghdha, D. Radev, C. Templeton and S. Teufel. Predicting the Impact of Scientific Concepts using Full-Text Features. In *Journal of the American Society for Information Science and Technology (JASIST), 67(11), 2016.*

· K. Thadani. Approximation Strategies for Multi-Structure Sentence Compression. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014).*

· K. Thadani and K. McKeown. Supervised Sentence Fusion with Single-Stage Inference. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013).*

· Y. Petinot, K. McKeown and K. Thadani. Cluster-based Web Summarization. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013).*

· K. Thadani and K. McKeown. Sentence Compression with Joint Structural Inference. In *Proceedings of The 17th SIGNLL Conference on Computational Natural Language Learning (CoNLL 2013).*

· K. Thadani, S. Martin and M. White. A Joint Phrasal and Dependency Model for Paraphrase Alignment. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012).*

· K. Thadani, F. Biadsy and D. Bikel. On-The-Fly Topic Adaptation for YouTube Video Transcription. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012).*

· W. Wang, K. Thadani and K. McKeown. Identifying Event Descriptions using Co-training with Online News Summaries. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2013).*

· K. Thadani and K. McKeown. Optimal and Syntactically Informed Decoding for Monolingual Phrase-Based Alignment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011).*

· Y. Petinot, K. McKeown and K. Thadani. A Hierarchical Model for Web Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011).*

· K. Thadani and K. McKeown. Towards Strict Sentence Intersection: Decoding and Evaluation Strategies. In *Proceedings of the Workshop on Monolingual Text-to-Text Generation at ACL HLT 2011.*

- K. McKeown, S. Rosenthal, K. Thadani and C. Moore. Time-Efficient Creation of an Accurate Sentence Fusion Corpus. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2010).*
- M. Jha, J. Andreas, K. Thadani, S. Rosenthal and K. McKeown. Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment. In *Proceedings of the Workshop for Creating Speech and Text Language Data using Amazon's Mechanical Turk at NAACL HLT 2010.*
- S. Rosenthal, W. J. Lipovsky, K. McKeown, K. Thadani and J. Andreas. Toward Semi-Automated Annotation of Prepositional Phrase Attachment. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010).*
- K. Thadani and K. McKeown. A Framework for Decreasing Textual Redundancy. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008).*
- T. Jebara, Y. Song and K. Thadani. Density Estimation under Independent Similarly Distributed Sampling Assumptions. In *Proceedings of the 20th Annual Conference on Advances in Neural Information Processing Systems (NIPS 2007).*
- T. Jebara, Y. Song and K. Thadani. Spectral Clustering and Embedding with Hidden Markov Models. In *Proceedings of the 18th European Conference on Machine Learning (ECML 2007).*

PATENTS

- B. Jawade, A. Eshratifar, K. Thadani, P. de Juan, J. Soares and B. Culpepper. Systems and Methods for Image Compositing via Machine Learning.
  US Patent application filed Mar 2024.
- K. Thadani, A. Bahadur, D. Rughwani, P. de Juan and J. Soares. Systems and Methods for Using AI to Facilitate Image Editing.
  US Patent application filed Mar 2024.
- E. Ye, X. Bai, N. O'Hare, E. Asgarieh, K. Thadani, F. Perez-Sorrosal and S. Adiga. Method and System for Webpage Classification and Content Delivery.
  US Patent application filed Aug 2023.
- P. de Juan, A. Shaw, E. Dodds, B. Culpepper, K. Thadani, L. Kesiraju, P. Mareedu, S. Shirwadkar, X. Zhou and Y. Ku. System and Method for Generating Video in Target Language.
  US Patent application filed Jul 2022.
- K. Thadani, P. Hairr, L. Oosterhof and X. Gao. Generation and Presentation of Summary List based upon Article.
  US Patent application filed Jun 2021.
- K. Thadani, A. Soni, P. Shah, T. Chevalier, S. Ramakrishnan, A. Nagao and Z. Qu. Ranking User Comments on Media Using Reinforcement Learning Optimizing for Session Dwell Time.
  US Patent application filed Jun 2019, granted Apr 2023.
- A. Balasubramanian, K. Thadani and A. Crews. Generating Presentations based on Articles.
  US Patent application filed Dec 2018, granted Jan 2022.
- A. Pappu, K. Thadani and N. Zalmout. Systems and Methods for Unsupervised Neologism Normalization of Electronic Content Using Embedding Space Mapping.
  US Patent application filed Oct 2018, granted Oct 2020.
- A. Pappu, R. Blanco, Y. Mehdad, A. Stent and K. Thadani. Entity Disambiguation.
  US Patent application filed Feb 2017, granted Feb 2024.

- Y. Billawala, Y. Mehdad, D. Radev, A. Stent and K. Thadani. Scalable and Effective Document Summarization Framework.
  US Patent application filed Feb 2016, granted Oct 2020.
- D. Bikel, K. Thadani, F. Pereira, M. Shugrina and F. Biadsy. Speech Recognition using Topic-Specific Language Models.
  US Patent application filed Dec 2012, granted Apr 2016.

TECHNICAL BACKGROUND

- Python, Java, C++, Matlab/Octave, Perl, Bash, Solidity, Common Lisp
- TensorFlow & PyTorch
- LaTeX, HTML & CSS