Thamme Gowda

tgowdan@gmail.com | gowda.ai | (213) 536-3552 | @thammegowda
/thammegowda | /thammegowda | Google Scholar | he/him

Education

University of Southern California – Viterbi School of Engineering — *Los Angeles, CA, USA*
Ph.D. in Computer Science, 2022 (exp.) — Adviser: Prof. Jonathan May
Dissertation: (WIP) *Rare phenomena learning in machine translation*[1]
M.S. in Computer Science, 2017 — Adviser: Prof. Chris Mattmann

Visvesvaraya Technological University – SJC Institute of Technology — *Chikkaballapur, KA, India*
B.E. in Computer Science & Engineering, 2012

Employment History

2017/07 – Now	Research Programmer I, II	USC Information Sciences Institute	*Marina del Rey, CA, USA*
2016/06 – 2017/05	Data Scientist Intern	NASA Jet Propulsion Laboratory	*Pasadena, CA, USA*
2015/12 – 2016/05	Research Assistant	USC Information Retrieval & Data Science	*Los Angeles, CA, USA*
2014/01 – 2015/07	Co-founder	Datoin.com	*Bengaluru, KA, India*
2012/06 – 2014/12	Software Engineer	SimplyPhi Software Solutions Pvt Ltd	*Bengaluru, KA, India*

Technical Skills

Programming languages	: Python, Bash, JavaScript, Java, Scala, SQL
Machine learning	: PyTorch, Tensorboard, NumPy, Scikit-Learn
Data science	: Pandas, Matplotlib, MS Excel, Jupyter
Web tech	: HTML/CSS/JS, RESTful API (JAX-RS, Flask), RDBMS (MySQL, SQLite3)
Big data	: Apache Spark, Nutch, Hadoop, Solr, Tika
Office software	: Microsoft Office, LaTeX, AsciiDoctor

Research Experience

Neural Machine Translation (NMT) @ USC ISI — https://nlg.isi.edu

- Showed that NMT models suffer from class imbalance and Zipfian distribution of words is a curse
- Addressed word type imbalance at evaluation phase; justified the use of macro-averaged metric
- Curated a massive dataset having 500M parallel sentences, trained multilingual Transformer that can translate 500 languages to English. Translation service is available for free via DockerHub. *Demo:* http://rtg.isi.edu/many-eng
- Resolved the mystery of why some choices of byte-pair-encoding (BPE) subword vocabulary size hyper parameters are better than others; offering a convincing explanation and heuristic to select vocabulary size, which significantly reduced hyper parameter search space
- Actively participated in research programs and delivered translation models, often the best performing:
  - IARPA Machine Translation for English Retrieval of Info in Any Language (MATERIAL), SARAL team
  - DARPA Low Resource Languages for Emergent Incidents (LORELEI), ELISA team
  - DARPA Learning with Less Labels (LwLL), CORAL team

Memex, Mars Target Encyclopedia @ NASA JPL — https://memex.jpl.nasa.gov

- DARPA Memex: created scalable web crawlers, clustering tools, and information extraction tools
- Built machine learning based multi-modal (text and image) classifiers that detect unlawful acts on the web to facilitate law enforcement e.g., illegal weapon sales and human trafficking
- Mars Target Encyclopedia: created PDF parsers, named entity recognition models, and search engine backend for Mars location names, and rock and soil chemical composition
- Applied transfer learning techniques to adopt a generic image classification model (e.g., ImageNet) to Mars satellite imagery; created an easy to use image labeling tool

Updated on: 2022/02/06

[1] https://gowda.ai/files/2022-TG-diss-proposal.pdf

Publications

MT : Machine translation IR : Information retrieval CV : Computer vision
IE : Information extraction WD : Web search and data mining Bias : Bias analysis

- **Thamme Gowda**, Mozhdeh Gheini, and Jonathan May. 2022. Improving robustness in multilingual machine translation via data augmentation. *Under Review* MT
- **Thamme Gowda**, Weiqiu You, Constantine Lignos, and Jonathan May. 2021a. Macro-average: Rare types are important too. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1138–1157, Online. Association for Computational Linguistics MT , IR
- **Thamme Gowda**, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021b. Many-to-English machine translation tools, data, and pretrained models. 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 306–316, Online. Association for Computational Linguistics MT
- **Thamme Gowda** and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics MT
- Ninareh Mehrabi, **Thamme Gowda**, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 231–232, New York, NY, USA. Association for Computing Machinery IE , Bias
- Elizabeth Boschee, Joel Barry, Jayadev Billa, Marjorie Freedman, **Thamme Gowda**, Constantine Lignos, Chester Palen-Michel, Michael Pust, Banriskhem Kayang Khonglah, Srikanth Madikeri, Jonathan May, and Scott Miller. 2019. SARAL: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 19–24, Florence, Italy. Association for Computational Linguistics MT , IR
- Leon Cheung, **Thamme Gowda**, Ulf Hermjakob, Nelson Liu, Jonathan May, Alexandra Mayn, Nima Pourdamghani, Michael Pust, Kevin Knight, Shrikanth Narayanan, David Chiang, Heng Ji, et al. 2017. ELISA system description for LoReHLT 2017 MT
- Xiaoman Pan, **Thamme Gowda**, Heng Ji, Jonathan May, and Scott Miller. 2019. Cross-lingual joint entity and word embedding to improve entity linking and parallel sentence mining. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 56–66, Hong Kong, China. Association for Computational Linguistics MT , IE
- Kyle Hundman, **Thamme Gowda**, Mayank Kejriwal, and Benedikt Boecking. 2018. Always lurking: Understanding and mitigating bias in online human trafficking detection. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 137–143, New York, NY, USA. Association for Computing Machinery Bias , WD
- Kiri Wagstaff, Raymond Francis, **Thamme Gowda**, You Lu, Ellen Riloff, Karanjeet Singh, and Nina Lanza. 2018a. Mars target encyclopedia: Rock and soil composition extracted from the literature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1) IE
- Kiri Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, **Thamme Gowda**, and Jordan Padams. 2018b. Deep Mars: CNN classification of Mars imagery for the PDS imaging atlas. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1) CV
- **Thamme Gowda**, Kyle Hundman, and Chris A. Mattmann. 2017. An approach for automatic and large scale image forensics. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*, MFSec '17, page 16–20, New York, NY, USA. ACM WD CV
- **Thamme Gowda** and Chris A. Mattmann. 2016. Clustering web pages based on structure and style similarity (application paper). In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 175–180 WD
- Chris A. Mattmann, Grace Hui Yang, Harshavardhan Manjunatha, **Thamme Gowda**, Andrew Jie Zhou, Jiyun Luo, and Lewis John McGibbney. 2016. Multimedia metadata-based forensics in human

trafficking web data. *The Search and Exploration of X-Rated Information (SEXI) workshop on the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016).*, pages 16–20 WD IE

Software Engineering Experience

- **Machine translation tools**
  - ***MTData*** https://github.com/thammegowda/mtdata | pip install mtdata
    A tool to locate, download, and extract parallel corpora for machine translation
    Handles various file formats for parallel corpus. Reduces network IO by using local cache
    Hundreds of languages and over hundred thousand datasets have been curated
  - ***RTG*** https://github.com/isi-nlp/rtg | pip install rtg
    Reader Translator Generator, a neural machine translation toolkit (NMT) based on PyTorch
    Features: YAML configuration for reproducible experiments, mixed float precision training, multi-node multi-GPU parallelism, gradient accumulation for large batches, flexible batching, flexible vocabulary management options, flexible learning rate schedules, embedding weight tying, fine-tuning and transfer-learning from parent to child, a web and REST API for model serving, etc
  - ***NLCodec*** https://github.com/isi-nlp/nlcodec/ | pip install nlcodec
    A scalable vocabulary management with support for chars, words, and byte-pair-encoding subword schemes; salable to large datasets using Apache PySpark
- **Sparkler** https://github.com/uscdatascience/sparkler
  A scalable web crawler on Apache Spark, with Apache Lucene/Solr backend and Banana dashboard
- **AutoExtractor** https://github.com/USCDataScience/autoextractor
  Scalable web page clustering toolkit, support clustering based on HTML DOM structure and CSS styles
- **Tensorflow+DL4J+Spark** https://github.com/thammegowda/tika-dl4j-spark-imgrec
  Image recognition at scale using Apache Spark; ported a Tensorflow model to Java using DL4J
- **Parser-Indexer** https://github.com/USCDataScience/parser-indexer-py
  Tools for parsing documents, extracting named entities and creating search index
- **SupervisingUI** https://github.com/USCDataScience/supervising-ui
  Web UI for creating labels for image classification, used by a many researchers
- **Datoin Platform** https://datoin.com/home/platform
  A software as a service platform for machine learning and big data applications
  In the capacity as technical co-founder, took an idea from whiteboard to minimal viable product demo, including the first set of machine learning applications to demonstrate its power. Wrote the first version of Datoin batch driver on Apache Hadoop, the second version on Apache Spark, glued various behind-the-scene synchronous services using REST APIs and asynchronous services using queues

Presentations

- Python reproducibility – at USC ISI and USC GRIDS; slides: https://bit.ly/3vTj2Mh
- PyTorch – at USC ISI Good Engineering; video: https://www.youtube.com/watch?v=8u4QqvtbAIw
- Sparkler – at Spark Summit 2017; video: https://www.youtube.com/watch?v=1fTomN1UMWI

Volunteering

- 2021–22: USC ISI NL Seminar Organizer https://nlg.isi.edu/nl-seminar
- 2017 : Google Summer of Code mentor for Apache Tika
  https://summerofcode.withgoogle.com/archive/2017/projects/4859682480979968
- 2016–18: Apache Tika Committer and PMC https://tika.apache.org
- 2016–17: Apache Joshua Committer and PMC https://github.com/apache/joshua
- 2016–17: Apache Nutch Committer and PMC https://nutch.apache.org

References

Available upon request