Thamme Gowda

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

Education

University of Southern California – Viterbi School of Engineering — *Los Angeles, CA, USA*
2018/08 – Now **Ph.D.** in Computer Science
2015/08 – 2017/05 **M.S.** in Computer Science
Visvesvaraya Technological University – SJC Institute of Technology — *Belgaum, KA, India*
2008/08 – 2012/05 **B.E** in Computer Science and Engineering

Employment History

2017/07 – Now	Research Programmer	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	University of Southern California	*Los Angeles, CA, USA*
2014/01 – 2015/07	Technical Co-founder	Datoin.com	*Bengaluru, KA, India*
2012/06 – 2013/12	Software Engineer	SimplyPhi Software Solutions Pvt Ltd	*Bengaluru, KA, India*

Skills

Programming languages : Python, Bash, JavaScript, Java, Scala, SQL,
Machine learning : PyTorch, Tensorboard, NumPy, Scikit-Learn
Data science : Pandas, Matplotlib, MS Excel charts, Jupyter
Web tech : HTML/CSS/JS, RESTful API (JAX-RS, Flask), RDBMS (MySQL, SQLite3)
Big data : Apache Spark, Nutch, Hadoop, Solr, Tika
Natural languages : Kannada (native), Telugu, Hindi, and English

Research Experience

Neural Machine Translation (NMT):

- Showed that NMT models also suffer from type (i.e. label) imbalance due to Zipfian distribution
- Argued for accounting the type imbalance during evaluation: justified the use of macro-average
- Resolved the mystery by offering a convincing explanation for why some choices of byte-pair-encoding (BPE) subword vocabulary size hyper parameters are better than others
- Curated a massive parallel dataset covering 500+ languages, trained a multilingual NMT model for 500-to-English translation, and released it for free via DockerHub. *Demo:* `http://rtg.isi.edu/many-eng`
- Actively participated in the following evaluations, delivered translation models (often the best):
  - SARAL team in IARPA Machine Translation for English Retrieval of Info in Any Language (MATERIAL)
  - ELISA team in DARPA Low Resource Languages for Emergent Incidents (LORELEI)
  - CORAL team in DARPA Learning with Less Labels (LwLL)

Mars Target Encyclopedia and Deep Mars

- Created tools for parsing research articles (PDF files) from planetary science literature
- Developed named entity recognition models for Mars location names, rock and soil chemical composition
- Trained image classification model for Mars imagery; applied transfer learning techniques

Memex

- In the spirit of DARPA Memex, created scalable web crawlers and clustering tools
- Created extractors for names, organizations, phone numbers, items on sale etc from web pages
- Text classifiers that detect unlawful act e.g., illegal weapon sales and human trafficking on web
- Image based forensic tools that detect illegal items from advertisement photos (for law enforcement)

Updated on: 2022/01/06

Publications

MT: machine translation | IR: information retrieval | CV: computer vision
IE: information retrieval | WD: web search and data mining | Bias: Bias analysis

- **Gowda, Thamme**, Mozhdeh Gheini, and Jonathan May. 2022. Improving robustness in multilingual machine translation via data augmentation. *Under Review*
- **Gowda, Thamme**, 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
- **Gowda, Thamme**, 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
- **Gowda, Thamme** 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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, **Gowda, Thamme**, 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
- **Gowda, Thamme**, 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. Association for Computing Machinery WD CV
- **Gowda, Thamme** 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, **Gowda, Thamme**, 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*** – 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 indexed
    Open sourced at github.com/thammegowda/mtdata; also available via `pip install mtdata`
  - ***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.
    Docs: isi-nlp.github.io/rtg; *code:* github.com/isi-nlp/rtg; also available via `pip install rtg`
  - ***NLCodec*** – a scalable vocabulary management with support for byte-pair-encoding subwords
    Scalable to large datasets using Apache PySpark
    Open sourced at github.com/isi-nlp/nlcodec and also available via `pip install nlcodec`
- **Sparkler** – a scalable web crawler on Apache Spark, with crawldb on Apache Lucene/Solr index
  Code: github.com/uscdatascience/sparkler
- **AutoExtractor** – web page clustering based on HTML structure and CSS style
  Code: github.com/USCDataScience/autoextractor
- **Tensorflow+DL4J+Spark** – image recognition at scale using Apache Spark backend
  Code: github.com/thammegowda/tika-dl4j-spark-imgrec
- **Parser-Indexer** – tools for parsing documents, extracting named entities and creating search index
  Code: Python: github.com/USCDataScience/parser-indexer-py, and Java: /parser-indexer
- **SupervisingUI** – a web UI for creating labels for image classification, used by a lot of researchers
  Code: github.com/USCDataScience/supervising-ui
- **Datoin.com** – 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

- PyTorch – at USC ISI Good Engineering; video: `https://www.youtube.com/watch?v=8u4QqvtbAIw`
- Python reproducibility – at USC GRIDS; slides: `https://bit.ly/3vTj2Mh`
- 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`
- 2016–*: Apache Tika Committer and PMC `https://tika.apache.org`
- 2016–*: Apache Joshua Committer and PMC `https://github.com/apache/joshua`
- 2017: Google Summer of Code mentor
  `https://summerofcode.withgoogle.com/archive/2017/projects/4859682480979968`
- 2016–17: Apache Nutch Committer and PMC `https://nutch.apache.org`