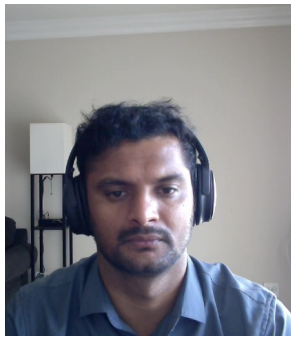
Many-to-English Machine Translation Tools, Data, and Pretrained Models

ACL 2021 System Demonstrations (Virtual)



Thamme Gowda
Information Sciences Institute
and Dept. of Computer Science
University of Southern California
tg@isi.edu

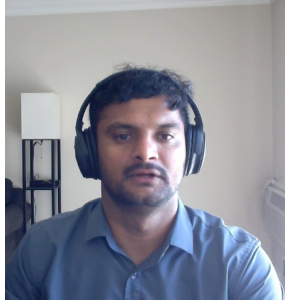
Zhao Zhang
Texas Advanced Computing Center
University of Texas
and NASA Jet Propulsion Lab

Chris Mattmann
Dept. of Computer Science
University of Southern California
and NASA Jet Propulsion Lab

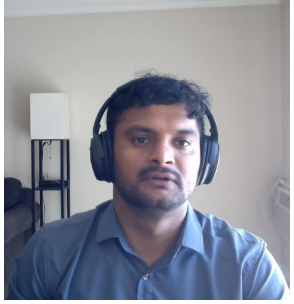
Jonathan May
Information Sciences Institute
and Dept. of Computer Science
University of Southern California

http://rtg.isi.edu/many-eng/

Overview



- Three tools for machine translation: MTData, NLCodec, RTG
- Task: 500-to-English translation in an opensource way
  - Collect a massive (bitext) dataset
  - Train a massively multilingual NMT model
- Applications:
  - Ready to use translation service; available via docker
  - Parent model for transfer learning



Tools

Focus: reproducibility and scalability

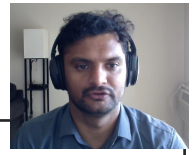
1. MTData: parallel dataset catalog and downloader
   - As of June 2021, 120K+ datasets, hundreds of languages; ISO 639-3
   - Publicly listed datasets: OPUS, Statmt.org, Paracrawl, ...
2. NLCodec: Vocabulary manager; and database layer
   - PySpark backend for large datasets
   - NLDb: Efficient storage and retrieval layer; parallelizable
3. Reader Translator Generator (RTG): NMT toolkit based on Pytorch
   - Reproducible experiments; a `conf.yml` per experiment
   - All the necessary ingredients for NMT research → production

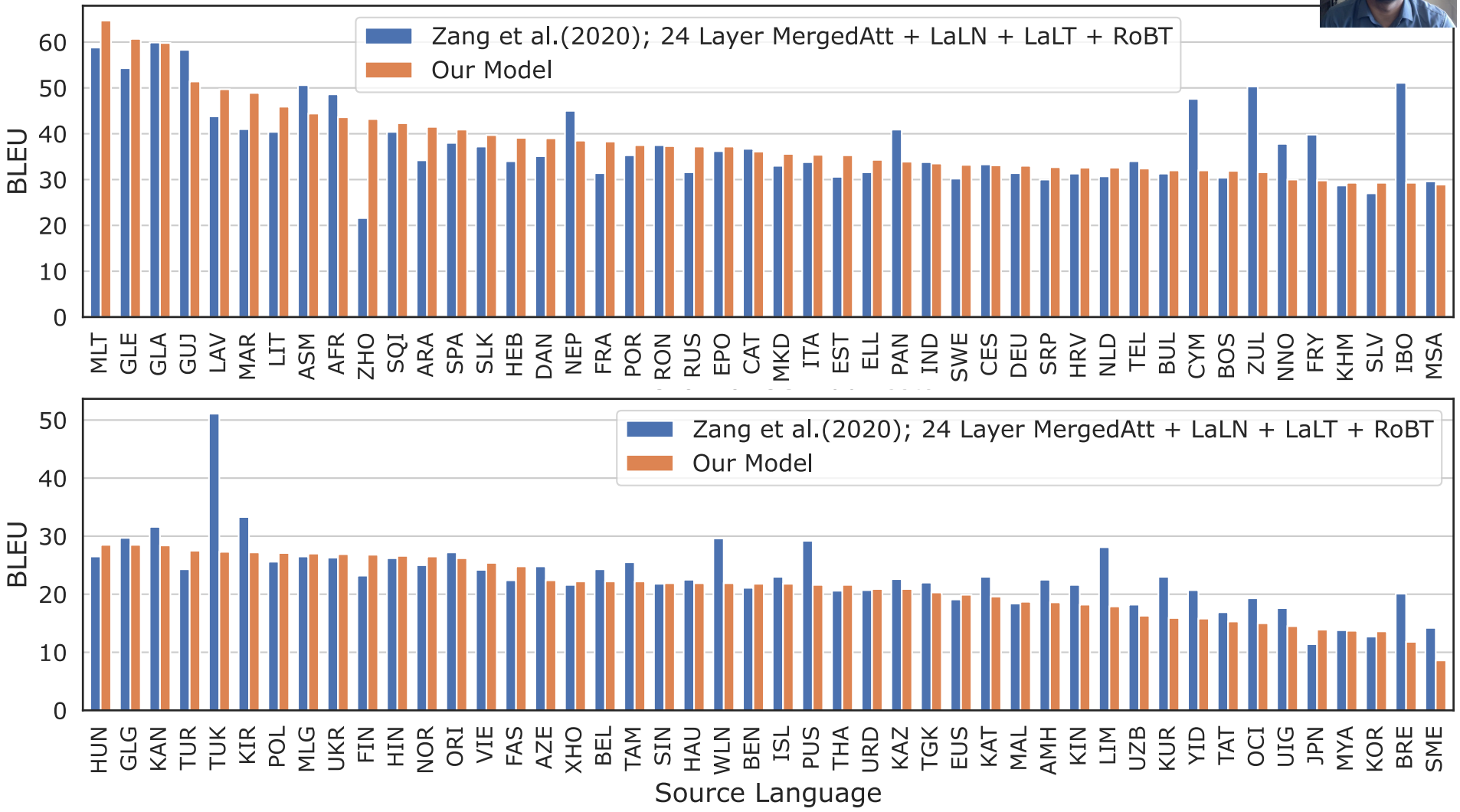
- `pip install mtdata nlcodec rtg`

500→English Translation

- Dataset: 500+ languages
  - Dedupe, cleaning, etc …
  - Excluding the known test sets e.g. NewsTest, OPUS-100, …
  - ➔ ~474 million sentence pairs; 9 billion tokens on each side
- Model: Transformer: 768d, 9 encoder, 6 decoder,
  - Separate BPE vocabularies: 512k source and 64k target embeddings
  - Large batches: ~720k toks per step, 200K steps
  - Gradient accumulation (5x), Float-16 ops, and distributed training on 8x A100 GPUs

Translation Service: BLEU on OPUS-100 Test Set

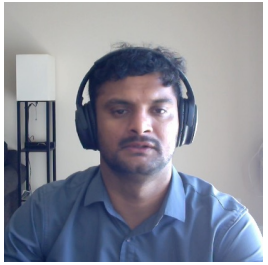




Transfer Learning: Fine Tuning

- E.g., two low-resource langs
  BRE: 1.2M ENG toks
  SME: 100K ENG toks
- Huge improvements in BLEU!

Model	BRE-ENG	SME-ENG
Baseline	12.7	10.7
500-eng parent	11.8	8.6
Finetuned	**22.8**	**19.1**



Take Away

- Home page: http://rtg.isi.edu/many-eng/
  - Demo service, data, models, tutorials …
  - Dataset: http://rtg.isi.edu/many-eng/data-v1.html
    https://opus.nlpl.eu/MT560.php [Thanks, Jörg Tiedemann]
  - Models: http://rtg.isi.edu/many-eng/models/
- Docker: `IMAGE=tgowda/rtg-model:500toEng-v1`
  `docker run --gpus '"device=0"' --rm -i -p 6060:6060 $IMAGE`
- Integrated to Apache Tika
  - Parses html, pdf, epub, docx, ppt, … runs OCR on images

Last Slide: Thanks 🙏

- Help bring more languages and datasets into MTdata
  - Issues and Pull requests are welcome

- Future work:
  MTData currently uses ISO 639-3, which has limitations
  - Language ID: (lang, script, region) e.g., BCP-47
    - ISO 639 for language names
    - ISO 15924 for script names
    - ISO 3166-1 for region names
  - Script and region can be optional, i.e., assume default values