USC Viterbi
School of Engineering

The Inevitable Problem of Rare Phenomena Learning in Machine Translation

Dissertation Proposal

by

Thamme Gowda

Jan 25, 2022

Committee

Jonathan May (advisor)
Chris Mattmann
Xuezhe Ma
Aiichiro Nakano
Shri Narayanan
Xiang Ren

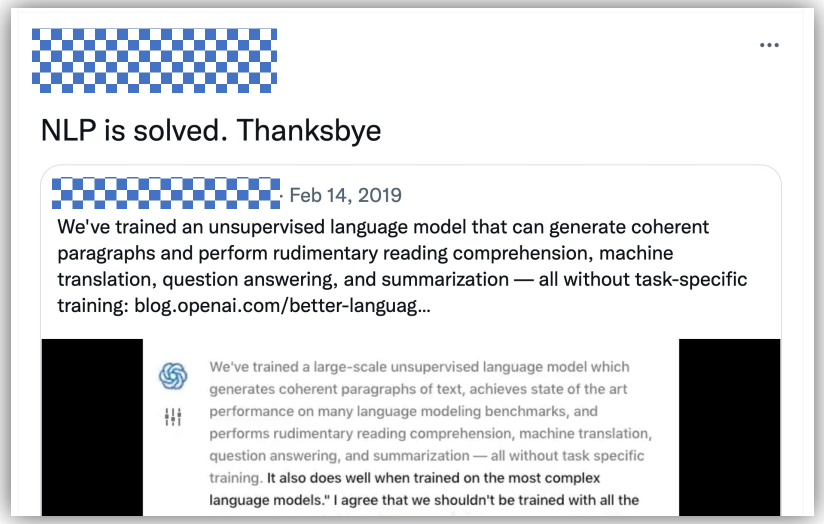
75 Years Ago ….

"A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. …

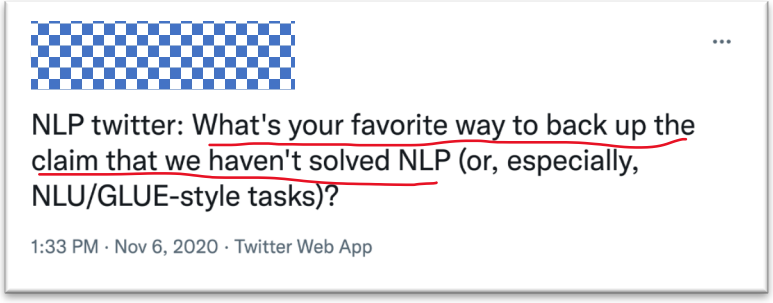
I have wondered if it were unthinkable to design a computer which would translate." – Warren Weaver, 1947[1]

[1] https://aclanthology.org/www.mt-archive.info/Weaver-1949.pdf

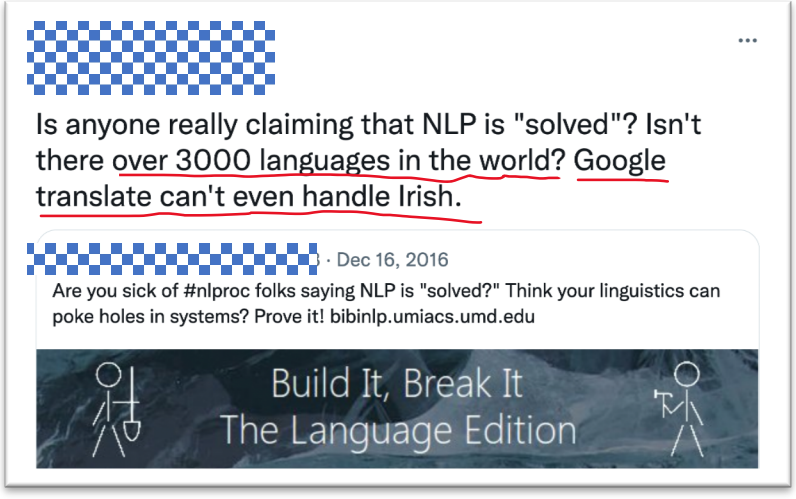
Is NLP/MT Solved Now?



1.

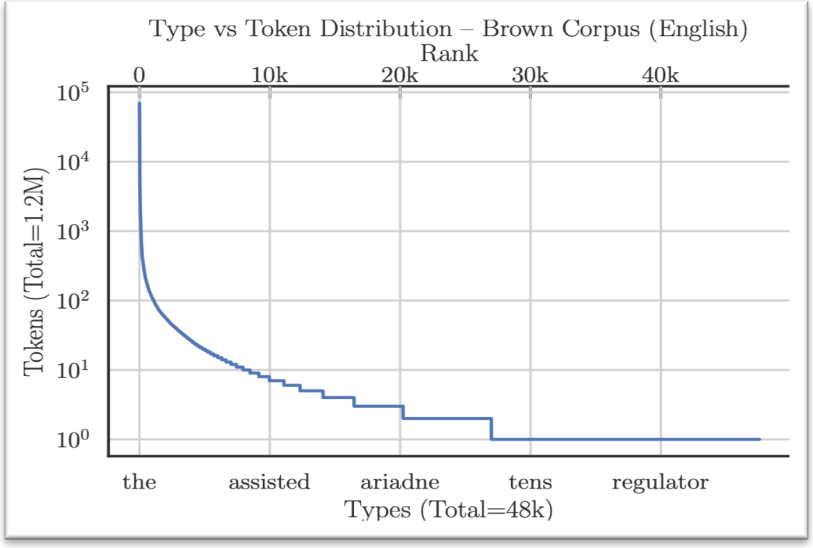


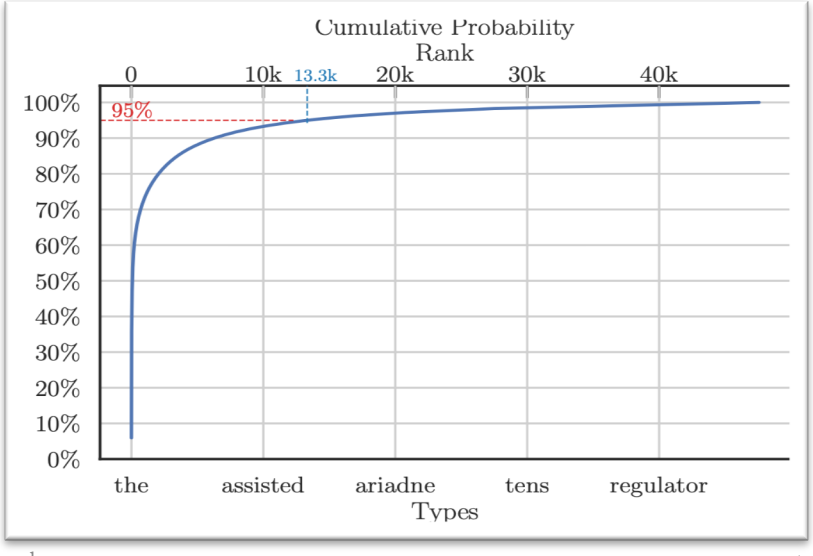
2.



Source: https://twitter.com/search?q="NLP is solved"

"Zipf's law says that most of the variance in language behavior can be captured by a small part of the system. ... Zipf's law, also says that most of the information about the language system as a whole is in the Long Tail.
... the machine learning techniques that we rely on are actually very bad at inducing systems for which the crucial information is in rare events. One day, either because of the demise of Moore's law, or simply because we have done all the easy stuff, the Long Tail will come back to haunt us." – Mark Steedman, 2008[1]

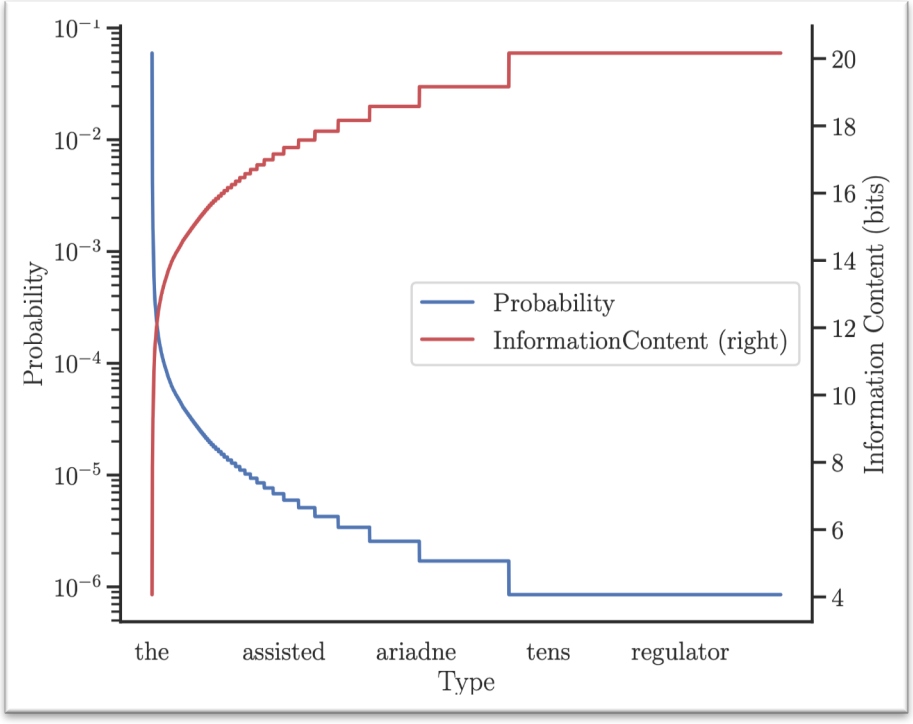


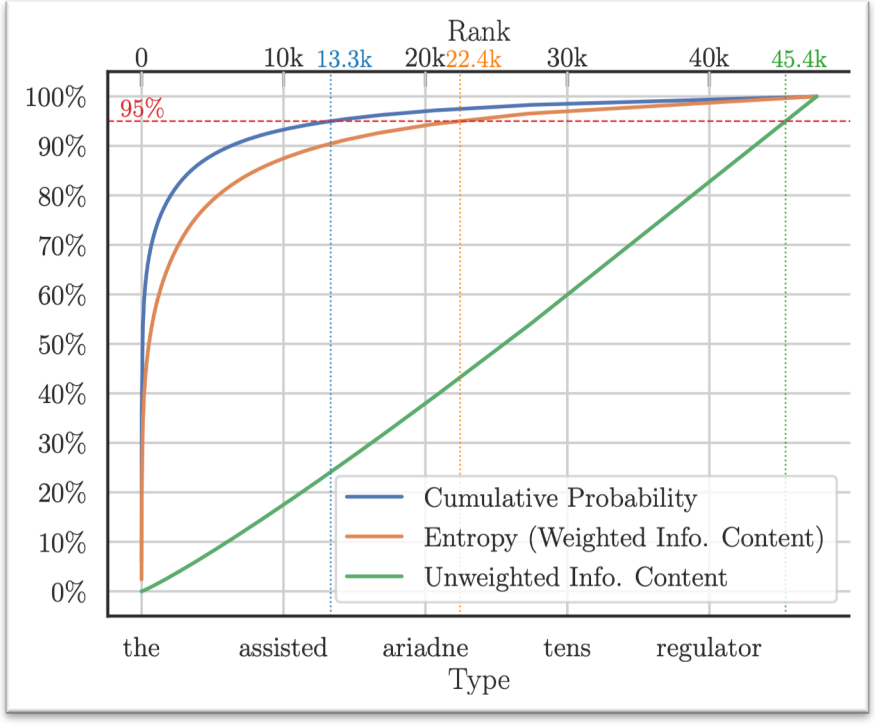


[1] https://aclanthology.org/J08-1008/

Long-tail Curse

- Much of *information of a language system* is in its rare types
- Information content of a type = $-\log_2 \mathrm{P}(x)$
- ML models stutter (and hallucinate) on data scarcity





MT is not solved until we address the long-tail problem, e.g., rare words

Language Coverage

- There are 7,000+ known living languages[1]
- But only about 100 languages are supported by popular MT
  - Google Translate[2]: 108; Microsoft Translate[3]: 103
- Much of research efforts are also limited to the top 100 languages
- MT models are unavailable for many languages

[1] Ethnologue: Languages of the World. http://www.ethnologue.com [2] https://blog.google/products/translate/five-new-languages/
[3] https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting/

TG's Thesis

- Rare phenomena learning is inevitable in machine translation (MT)
- It manifests in various forms

1. Rare words
   I. At training time → Mitigate the imbalance in word type distributions
   II. At evaluation time → Recognize the importance of rare words
2. Rare languages → Support low-resource languages
3. Rare phrases → Improve robustness

- Until we address these challenges, MT is far from being solved

Progress: Dealing with Rare Phenomena in MT

I – Rare words in Training

"Finding the Optimal Vocabulary Size for NMT" [EMNLP 2020 Findings]

II – Rare words in Evaluation

"Macro-Average: Rare Types are Important Too" [NAACL 2021]

III – Rare Languages (500+ languages)

"Many-to-English MT Tools, Data, and Pretrained Models" [ACL 2021 Demos]

IV – Robustness / Language Switching

"Improving Robustness in MT via Data Augmentation" [Under review/NAACL22]

Implications: Look Ahead

	Before/Now	After/End of the Presentation
NMT	NMT is generation	NMT is classification[1]
Vocabulary Size	Arbitrary hyperparameter	Well reasoned parameter; chosen using a heuristic [1]
Evaluation	Treat each 'token/instance' equally	Important tokens are treated more important[2]
Scaling NMT	To ~100 languages	To ~500 languages; Bunch of useful tools[3]
Multilingual NMT robustness	Not robust	Robust to language switching [4]

[1] **Gowda** and May, *'Finding the optimal vocabulary size for NMT'*, EMNLP 2020 Findings
[2] **Gowda** et al, *'Macro-average: Rare types are important too'*, NAACL 2021
[3] **Gowda** et al, *'Many-to-English tools, data, and pretrained models'*, ACL 2021 Demos
[4] **Gowda** et al, *'Improving multilingual MT robustness via data augmentation,* [under review/NAACL2022]

Co-Authors

- Jonathan May, USC ISI (+advisor+committee)
- Chris Mattamann, USC & JPL (+committee)
- Mozhdeh Gheini, USC ISI
- Zhao Zhang, UT & JPL
- Weiqiu You, U Penn
- Constantine Lignos, Brandeis University

Collaborators/Special Thanks

- Scott Miller, USC ISI
- Shantanu Agarwal, USC ISI
- Joel Barry, USC ISI

Committee Members

- Xuezhe Ma
- Shri Narayanan
- Aiichiro Nakano
- Xiang Ren

Acknowledgments

- USC Employee Benefits (Tuition Assistance)
- DARPA LORELEI
- IARPA MATERIAL
- DARPA LwLL

Computing Resources

- USC Center for Advanced Research Computing (CARC)
- Texas Advanced Computing Center (TACC)

Part-I

Rare Words in Training

Finding the Optimal Vocabulary Size for NMT

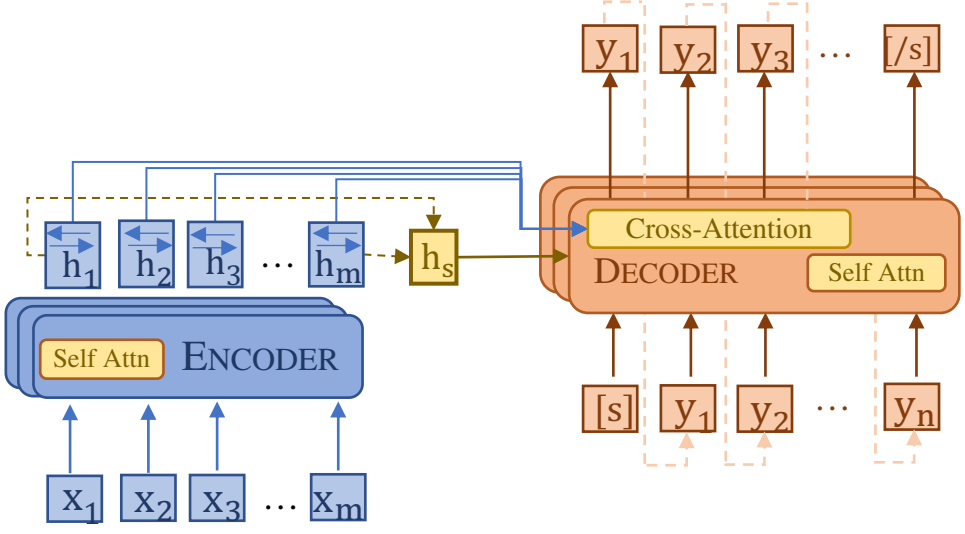
EMNLP Findings 2020

Thamme Gowda and Jonathan May

https://aclanthology.org/2020.findings-emnlp.352/

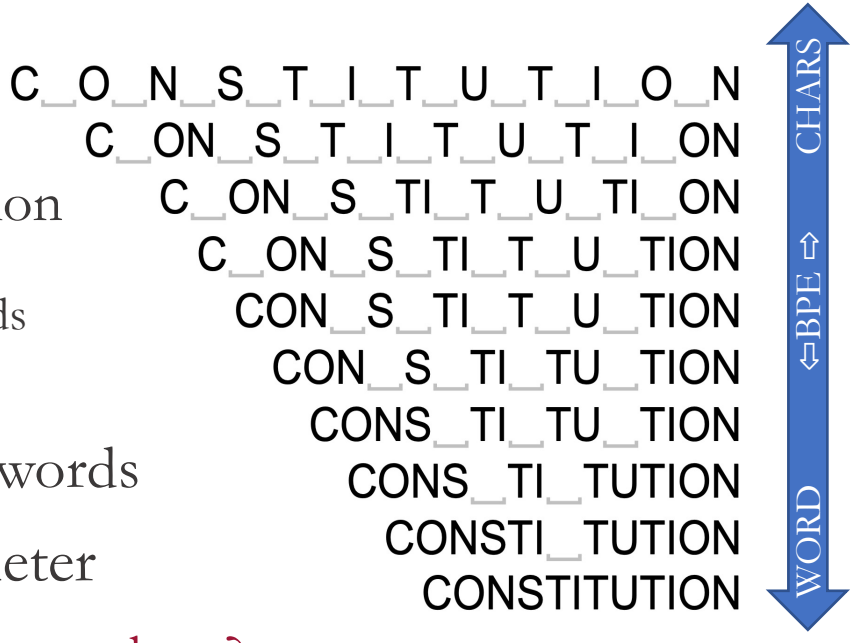
Neural Machine Translation

- NMT, f: $(x_1 x_2 x_3 \dots x_m) \rightarrow (y_1 y_2 y_3 \dots y_n)$
- $y_{1:n} = \text{Decoder}(\text{Encoder}(x_{1:m}))$
- Maximize $P(y_{1:m} | x_{1:m})$ ⇨ Maximize $\prod_{t=1}^{n} P(y_t \mid y_{<t}, x_{1:m}; \theta)$



BPE Subwords

- Out-of-vocabulary/UNK words used to be a hard problem in MT
- Byte-pair-encoding sub words
  ➔ addresses open-vocabulary generation
  [Sennrich et al 2016[1]]
  ➔ gets better performance than chars and words
- Subwords are obtained by merging most frequent chars and subwords
- Number of merges is a hyper parameter
- Why are some vocabulary sizes better than others?

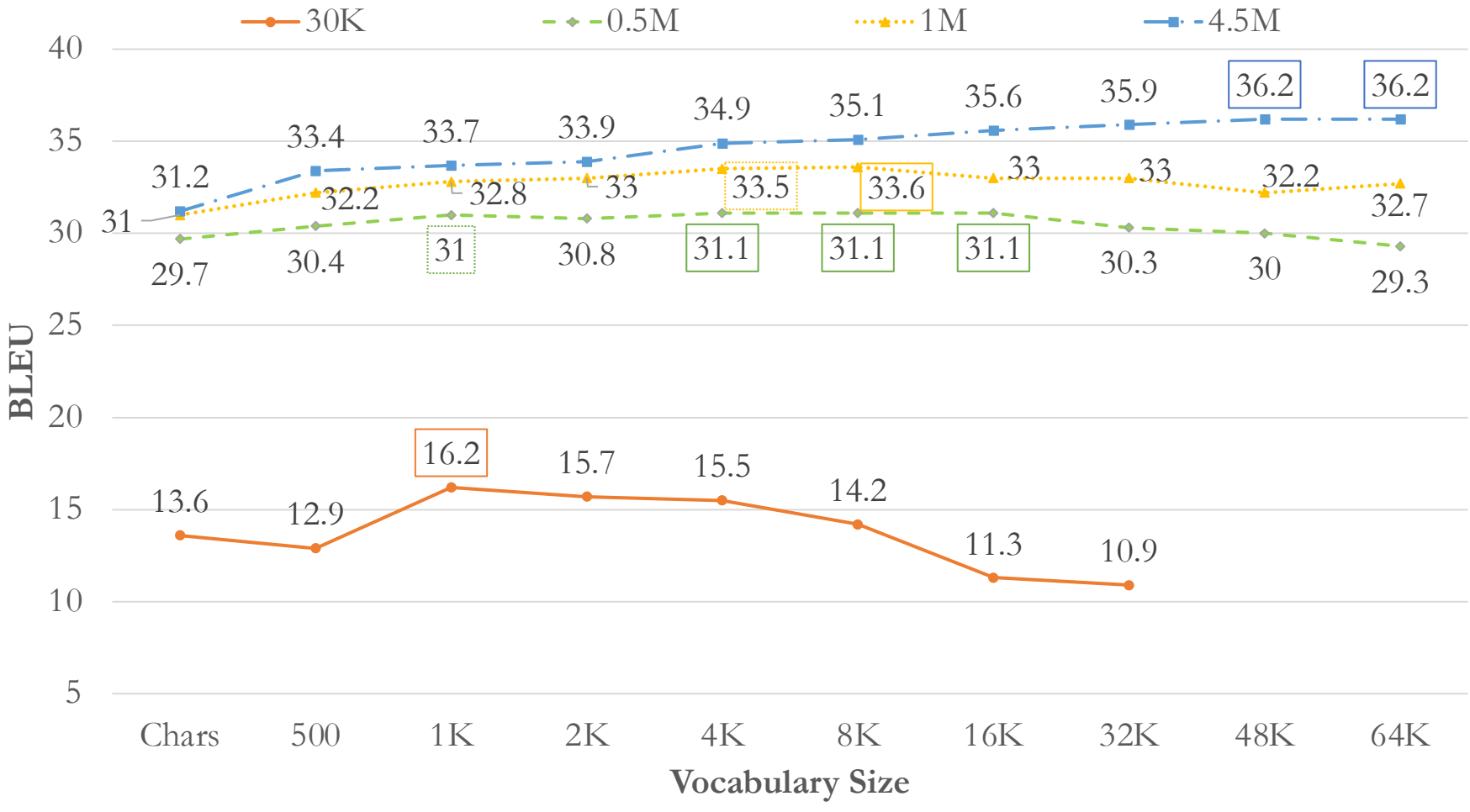


[1] https://aclanthology.org/P16-1162/

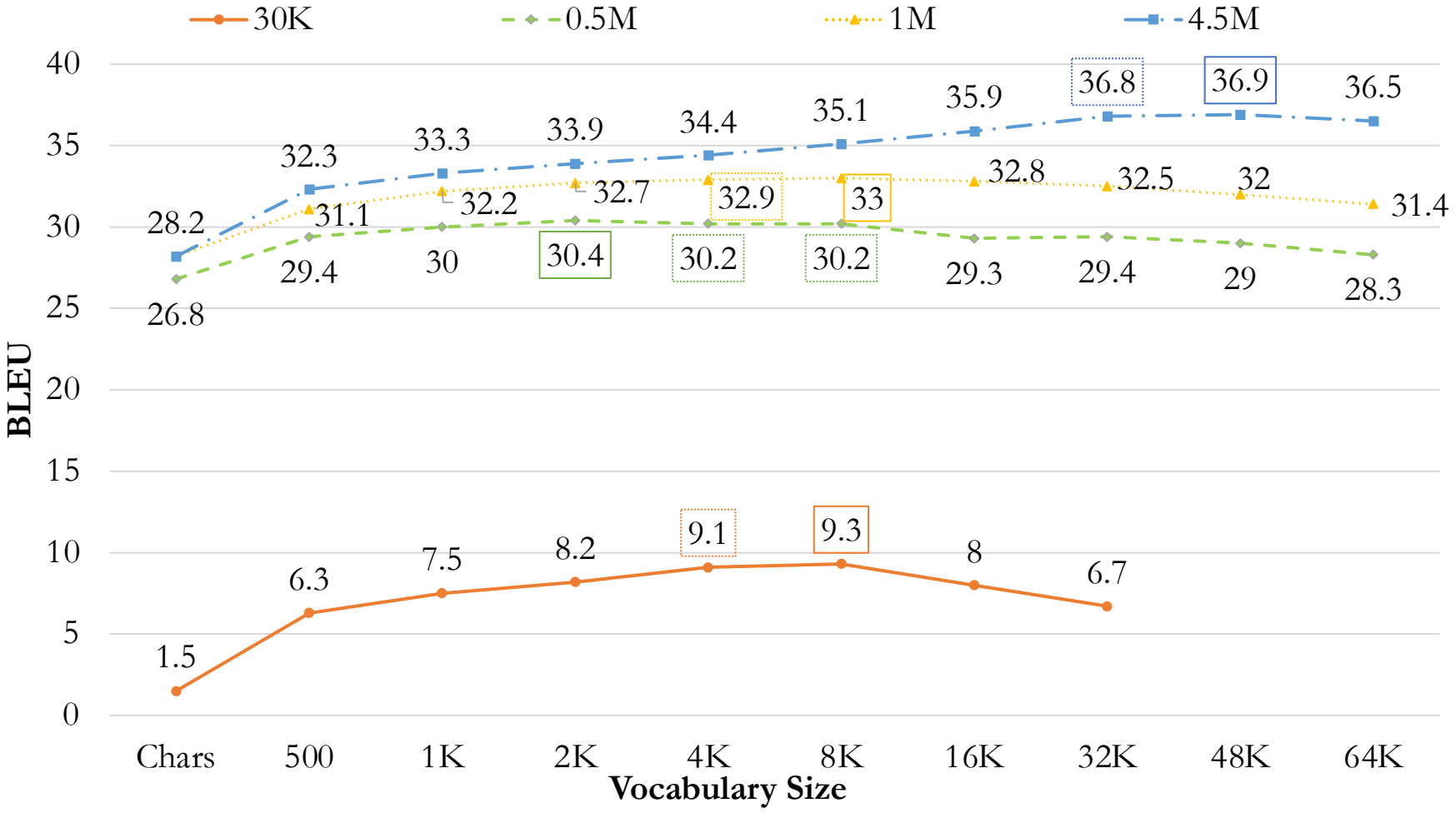
Experimental Setup

- Four target languages: DE→EN, EN→DE, EN→HI, and EN→LI
  - Total of 11 dataset sizes: between 30K and 4.5M sentences
  - x 10 vocabulary sizes: Chars, 500, 1K, 2K, 4K, 8K, 16K, 32K, 48K, and 64K
- **Transformer** with 6 layers, 512 dims, 8 attn heads, 0.1 dropout, …
  - Adam optimizer; 16,000 warmup steps followed by inverted sqrt decay; training stops when the validation loss start to climb up
  - Beam decoder, length penalty, checkpoint averaging …
  - Separate vocabs for source and target
    Decoder's input and output embeddings are tied

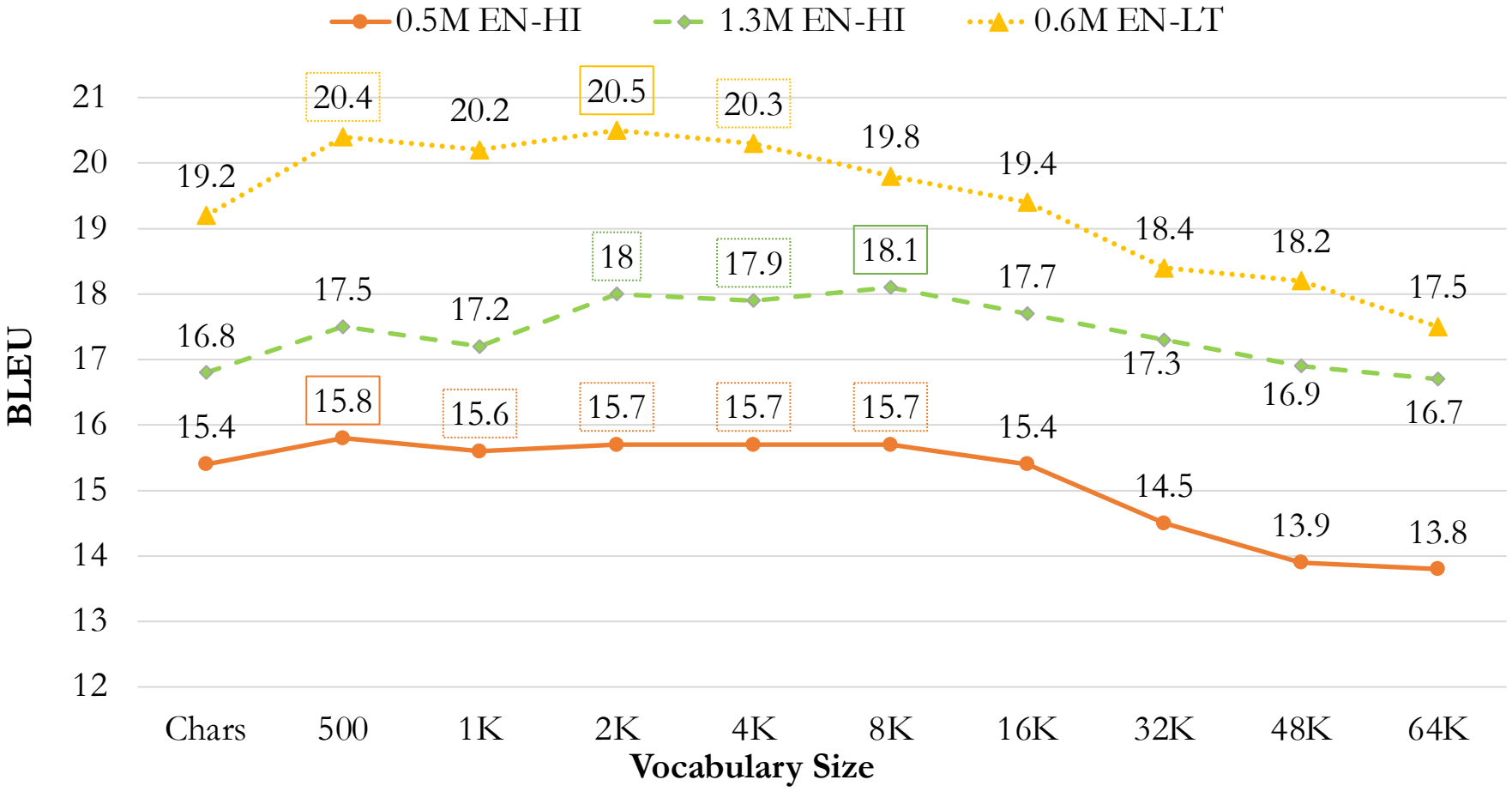
DE→EN NewsTest2019 BLEU vs Vocabulary Size

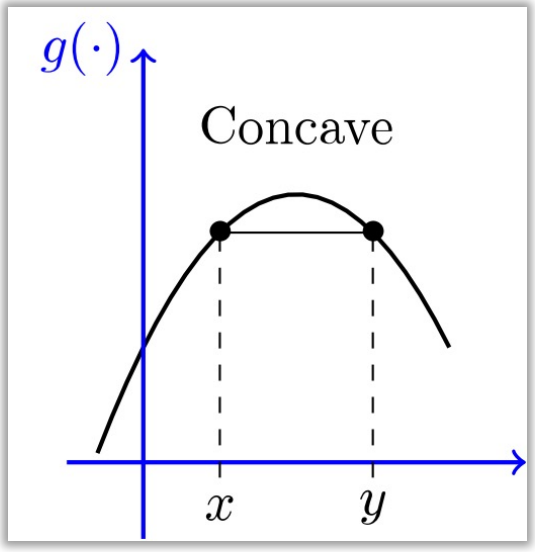


EN→DE NewsTest2019 BLEU vs Vocabulary Size

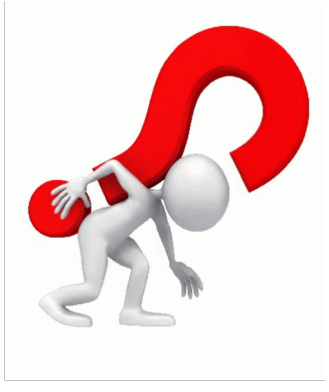


BLEU vs Vocabulary Size
EN→HI IITB Test and EN→LT NewsTest2019





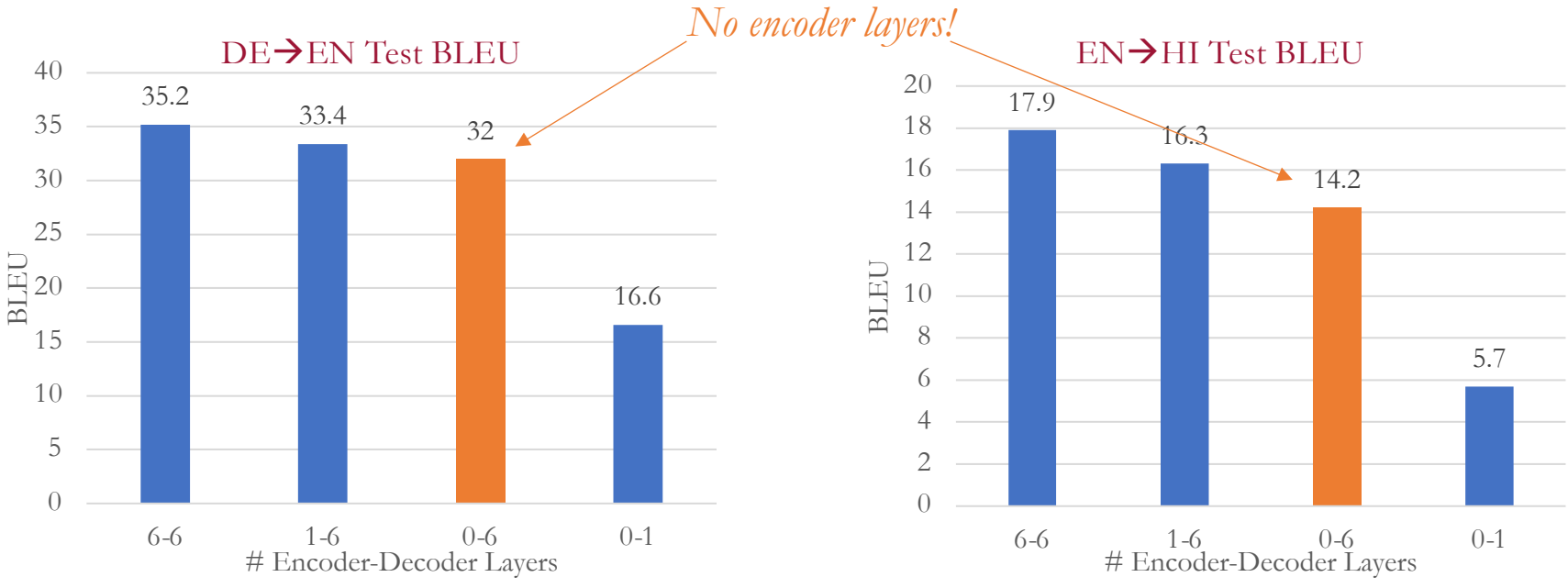
Why are all BLEU lines
(sort of) concave down on vocabulary size?



Transformer Ablation

Hypothesis: Encoder is *not* a must have, but rather a good to have component

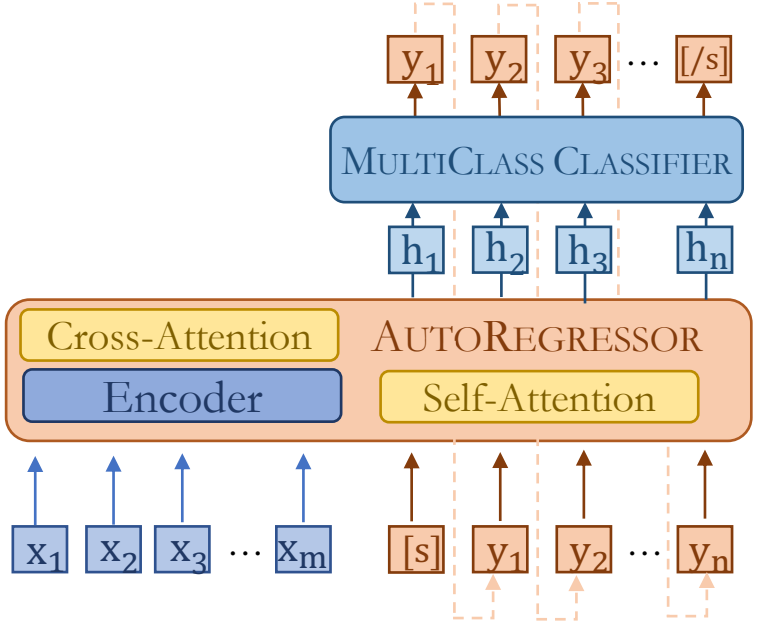
Varying number of layers; DE→EN (shared vocab) and EN→HI (separate vocab)



Most people do not listen with the intent to understand; they listen with the intent to reply. - Stephen Covey, 2004

NMT Abstraction

- $y_{1:n} = \text{Decoder}(\text{Encoder}(x_{1:m}))$
- Maximize $\prod_{t=1}^{n} P(y_t \mid y_{<t}, x_{1:m}; \theta)$
  - ⇨ Maximize $\prod_{t=1}^{n} P(y_t \mid h_t; \theta)$ where $h_t = f(y_{<t}, x_{1:m}; \psi)$
- NMT = MulticlassClassifier + AutoRegressor



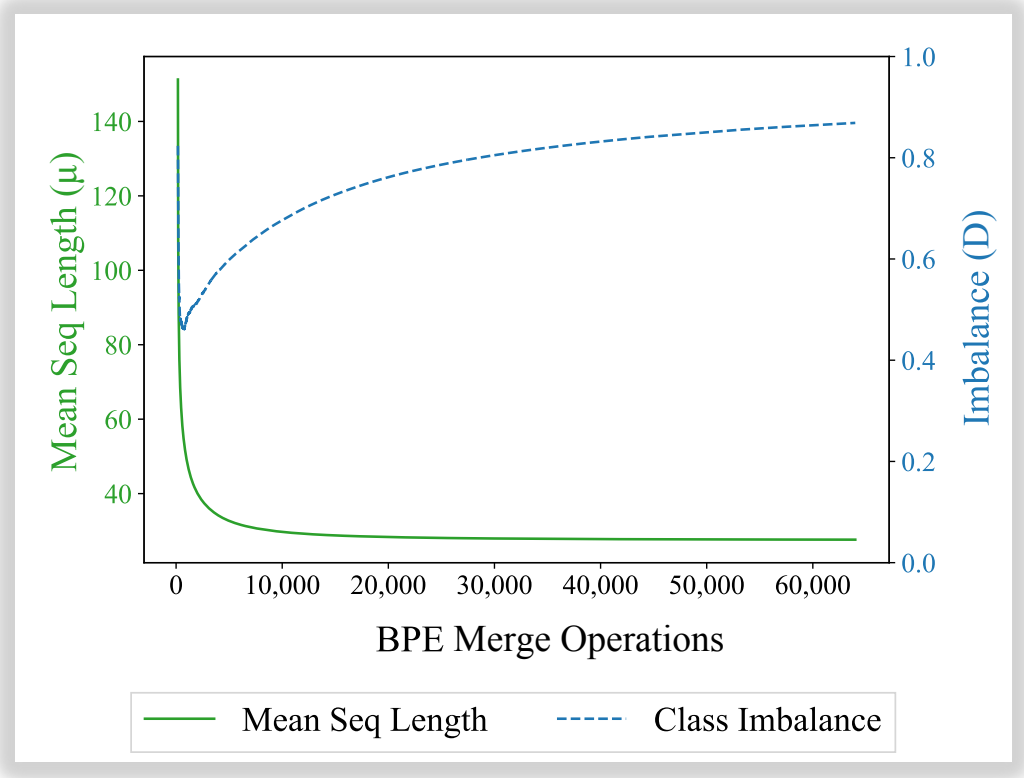
High-level abstraction of Transformer NMT

Imbalanced Classification Problem

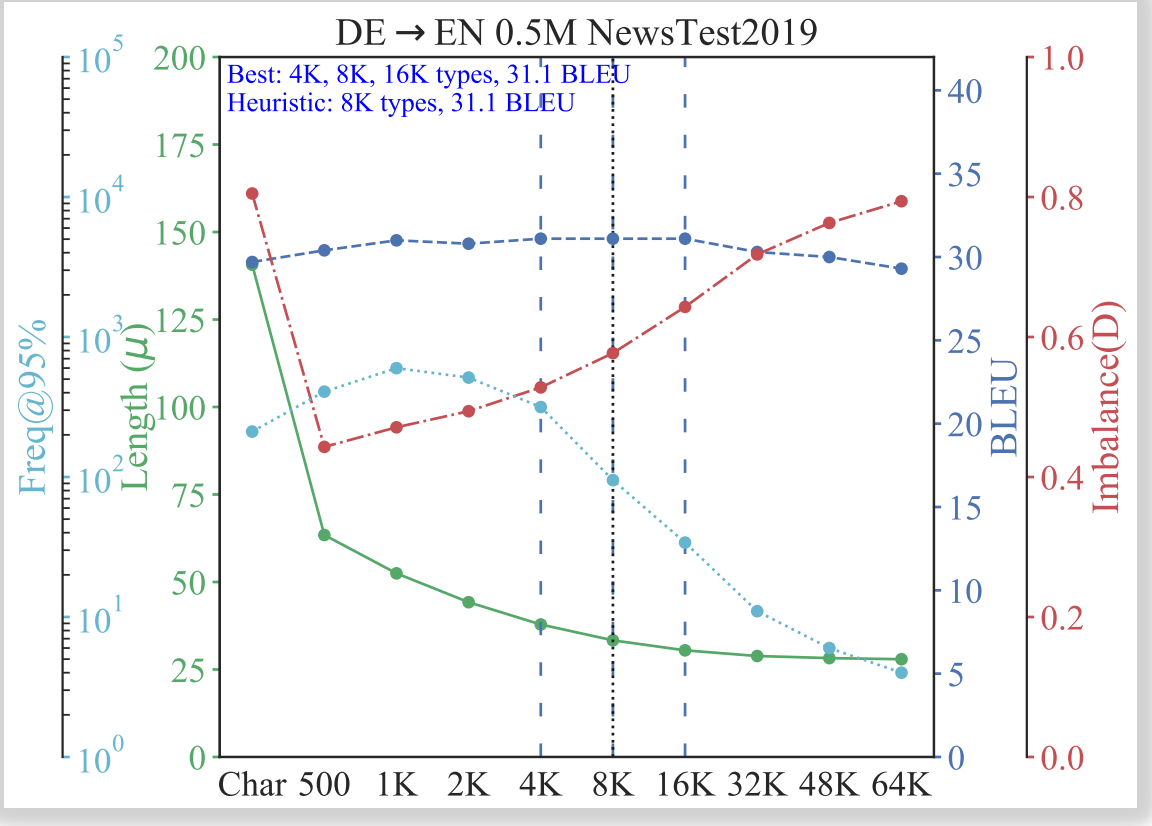
- Classification problem: output variable is discrete e.g., classes, word types
- Imbalanced : unequal frequency distribution
  i.e., some classes are frequent → majority class(es)
  and others are rare → minority class(es)
- Examples
  1. Cancer detection: fewer positive labels than negative
  2. Image segmentation: "background pixel" is more prevalent than foregrounds
  3. NLP: stopword types have more tokens than content word types
- <u>Problem:</u> Frequency based biases degrades overall performance
  - Minority classes are ignored → poor recall
- Special care is required during evaluation → *the topic for Part-II*
- Imbalanced Learning:
  Byte pair encoding (BPE): Balances classes via splitting and merging
  → *Tuning the vocabulary size improves class balance* ✅

Effect of BPE

- As BPE merge operations increase
  - Sentence length decreases
  - Class imbalance increases*
- We need both shorter sequences and smaller imbalance values ⇨
  - Left: balanced but long
  - Right: short but imbalanced
- Best vocab size is the one that reaches a good trade-off



Heuristic



"Use the largest possible BPE vocabulary such that at least 95% of classes have about 100 or more training examples"

Frequency-based Bias on Class Performance

- Pearson Correlation Coefficient
- Rank test set classes based on *training set frequency*

1. Rank vs Precision ($\rho_{\mathcal{R},P}$)

   $\rho_{\mathcal{R},P}$ is positive at high D

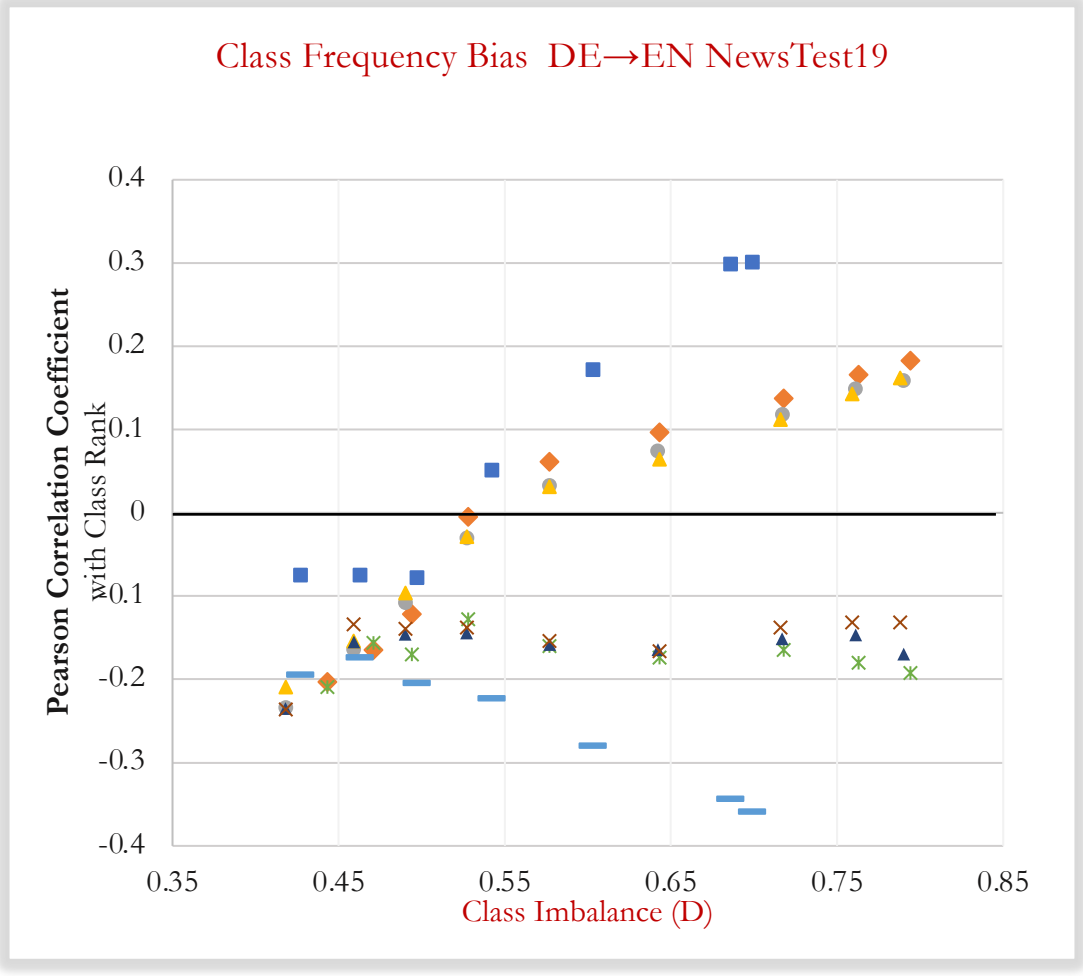
   ⇨ Frequent classes have relatively poor precision

2. Rank vs Recall ($\rho_{\mathcal{R},R}$)

   $\rho_{\mathcal{R},R}$ is negative

   ⇨ Rare classes have poor recall

- Takeaway: Recall of rare classes is still problematic



Part-I Conclusion

Related Work

- Others have focused on ways to search vocabulary size
  - we have given explanation for *why* some sizes are better than others, in addition to a heuristic
- No other work showing frequency-based biases in NMT

Summary

- Imbalance is unavoidable in natural language generation datasets
- We can split [or merge sub-]words, which is effective to handle imbalance
  - One of the reasons why byte-pair-encoding/sub-words is very effective in NMT
- NMT models have frequency-based biases
  - Rare types have lower recall than frequent types

Part-II

Rare Words at Evaluation

All words are important, but some words are more important than others.

Macro-Average: Rare Types are Important Too

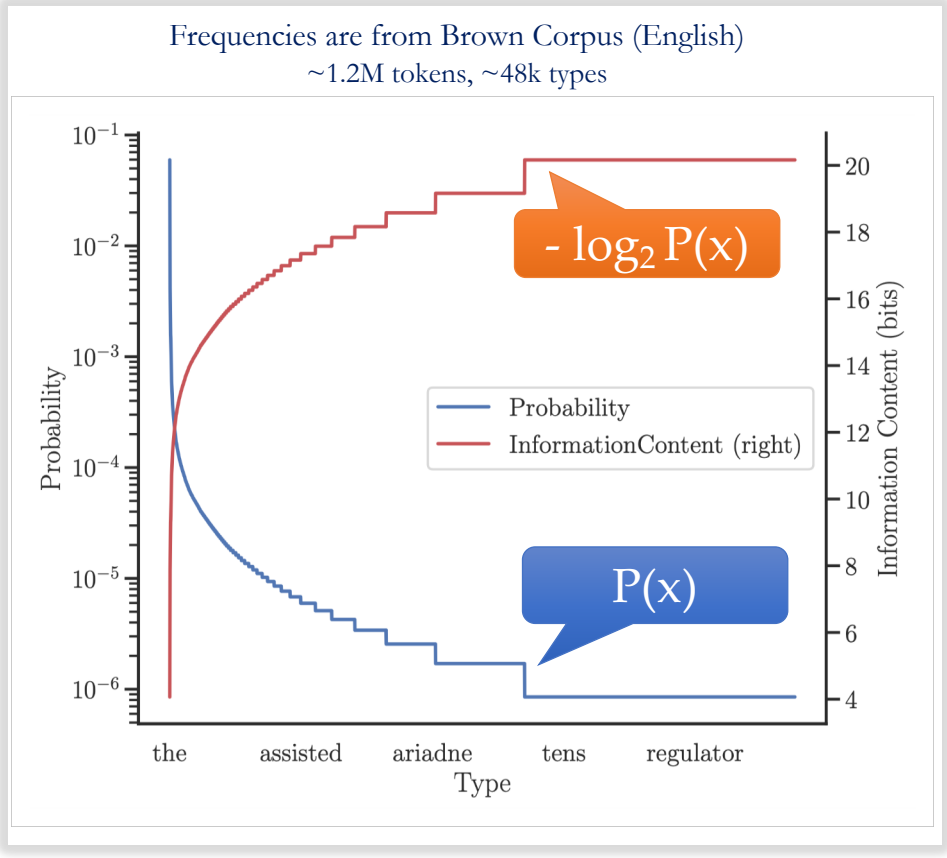
NAACL 2021

Thamme Gowda, Weiqiu You, Constantine Lignos, and Jonathan May

https://aclanthology.org/2021.naacl-main.90/

Motivation / Problem

- Natural language datasets have imbalanced word types
- ML techniques, including NMT, suffer from class imbalance
  - Poor recall for rare types
- Rare types have more information content
- Extra care is required, otherwise, metrics give a false sense of confidence on model performance
  - E.g., cancer detection with 1:99 imbalance; majority label classifier gets 99% overall accuracy, but that's not a useful metric
  - Same is true for NLP



Review of MT Metrics

- [Papineni et al. 2002[1]] $\mathrm{BLEU} = \mathrm{BP} \times \prod_{n=1}^{4} P_n$
  P_n is n-gram precision of tokens. BP is brevity penalty
- [Popović, 2015[2]] $\mathrm{ChrF}_\beta = (1+\beta^2)\frac{\mathrm{ChrP} \times \mathrm{ChrR}}{\beta^2 \times \mathrm{ChrP} + \mathrm{ChrR}}$
  Character n-grams for up to 6-grams
- [Sellam et al. 2020[3]] BLEURT
  BERT, a transformer language model, finetuned to predict human judgement scores on WMT

[1] https://aclanthology.org/P02-1040/ [2] https://aclanthology.org/W15-3049/ [3] https://aclanthology.org/2020.acl-main.704/

Multiclass Classifier Evaluation

Performance of a Class: $F_{\beta;c} = (1+\beta)^2 \frac{P_c \times R_c}{\beta^2 \times P_c + R_c}$

Where P_c and R_c are Precision and Recall

Multi-class performance = an average of individual class performances

1. **Macro-average:** unweighted i.e., equal importance to each *type*

   $\text{MacroF}_\beta = \frac{\sum_{c \in V} F_{\beta;c}}{|V|}$

2. **Micro-average**: weighted e.g., by frequency: equal importance to each *token*

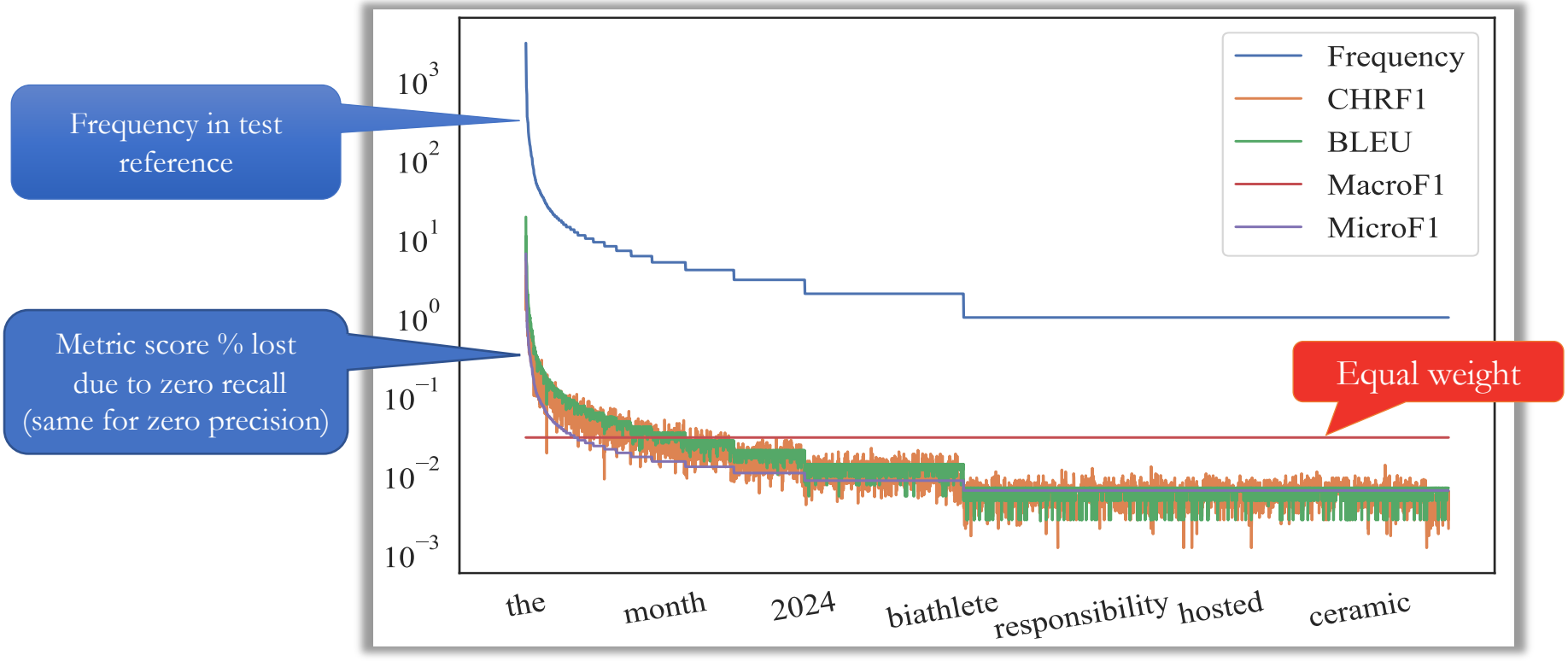
   $\text{MicroF}_\beta = \frac{\sum_{c \in V} w_c \times F_{\beta;c}}{\sum_{c' \in V} w_{c'}}$

   where, weight for class, $w_c = \text{Refs}(c) + k$ for some $k \geq 1$

   - *We use* $k = 1$; Note: if $k \to \infty$, $MicroF_\beta \to \text{Macro}F_\beta$
   - *We use* $\beta = 1$, and scale final scores to [0, 100], just like BLEU

NOTE: Micro-F1 ≅ Accuracy

MacroF1 vs Others



MacroF1 has equal weight for all types (WMT 19 DE-EN NewsTest)

Micro-averaged metrics overlook improvements from rare types, after rounding to one or two decimals

Empirical Justification for MacroF1 as an MT Eval Metric

➔Compare MacroF1 with
- ➢ BLEU and ChrF1
- ➢ MicroF1 – so we can see the difference micro and macro avg makes
- ➢ BLEURT – model-based metric based on BERT
  - *BLEURT has undesirable biases, shown in paper (not talking about here)*

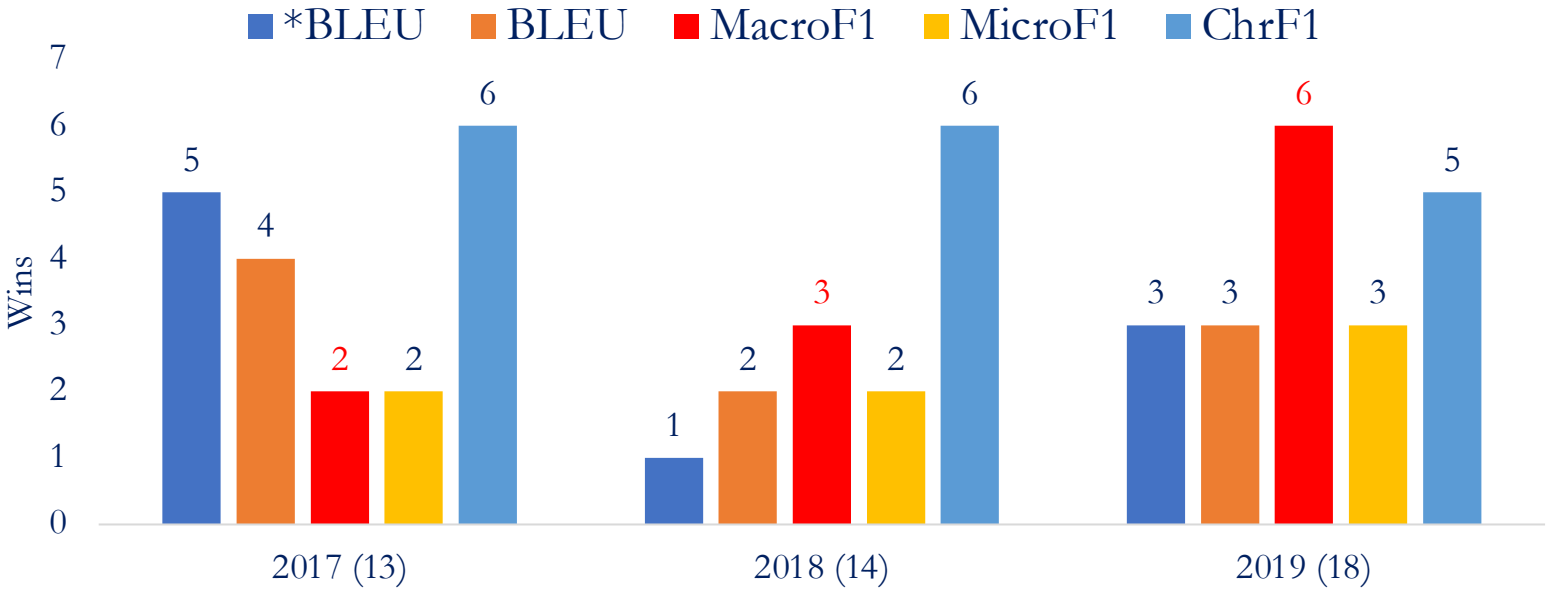
➔ On these tasks:
- **Machine Translation:** direct human assessment: WMT Metrics 2017-19
  MT vs Human judgement score correlations
- **Downstream CLIR Task metrics:** CLSSTS 2020
  MT vs IR score correlations
- **Data-to-text:** Fluency and Semantics: WebNLG

Not talked here

- MacroF1 as ASR metric correlation with IR metrics: (IARPA MATERIAL program)

WMT Metrics: Wins per Metric

- Wins = Number of times a metric scored highest correlation with human judgements
- *BLEU is from the WMT metrics package, precomputed by task organizers
- MacroF1 and MicroF1 use the same tokenizer as BLEU, obtained using SacreBLEU



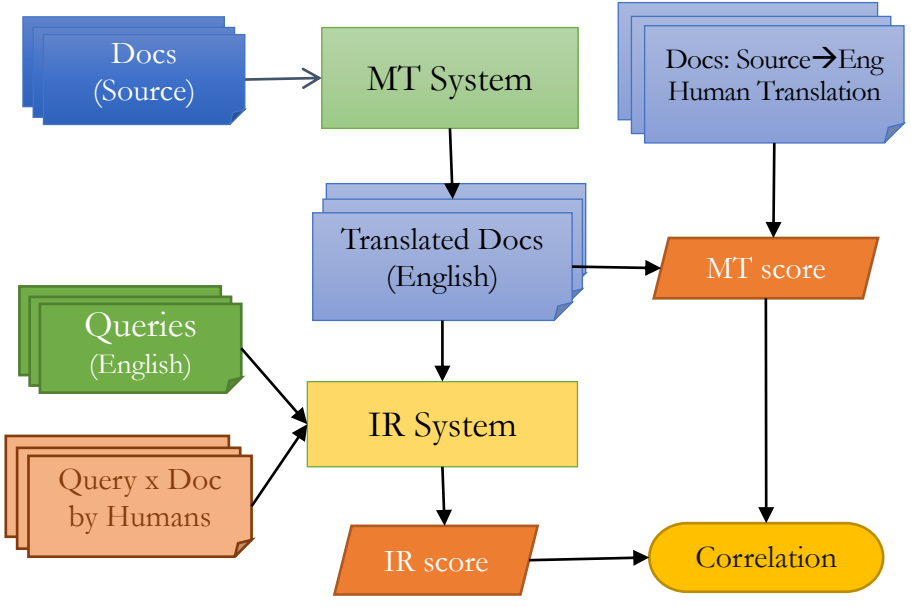
MacroF1 has more wins in the recent year -- when systems are mostly fluent, adequacy is a key discriminator

CLIR Task: Pipeline

CLSSTS 2020 / IARPA MATERIAL

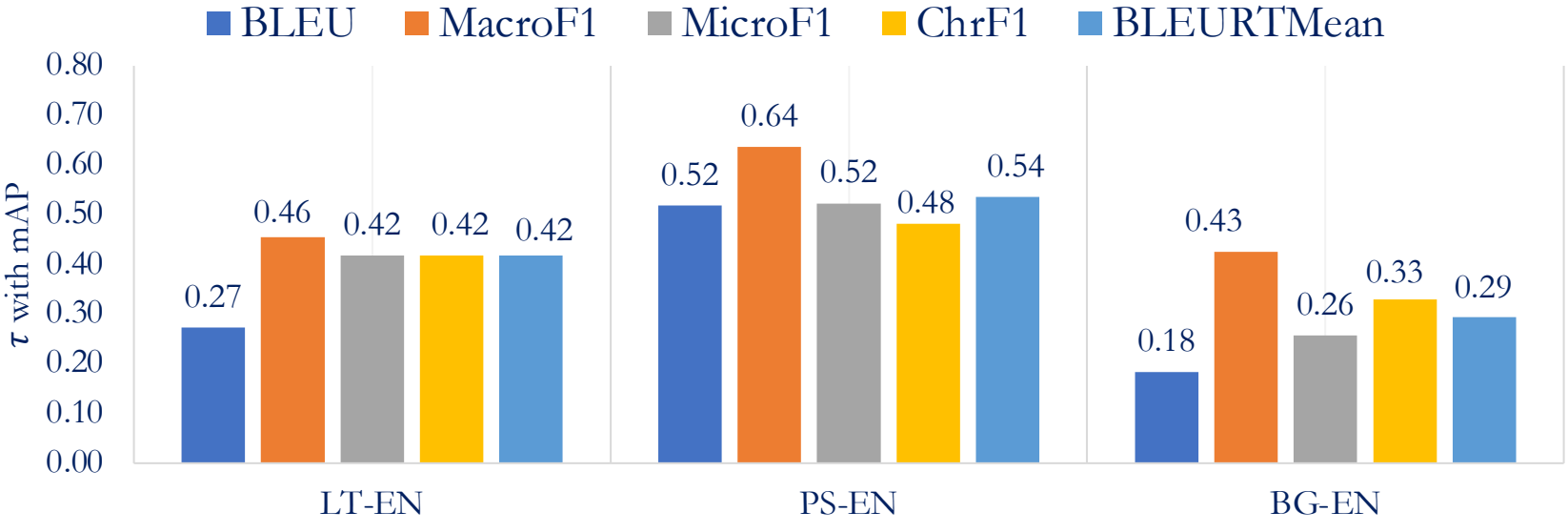
IR task with queries and docs in different languages

1. Build a set of MT models; translate all source documents to the target language, compute MT metric(s)
2. For each MT model's translations, build an IR model, and measure IR metrics
   [Thanks to Joel Barry and Shantanu Agarwal]
3. Find the correlation between the set of MT scores and IR scores. The MT metric having stronger correlation with IR metric(s) is more useful than others.
4. Repeat this on many languages:
   LT-EN, PS-EN, BG-EN



Downstream CLIR Task

- IR task with queries and docs in different languages
- Translate source docs to target language, and match queries with docs
- MT metric having strong correlation with IR metric (e.g., mAP) is more useful

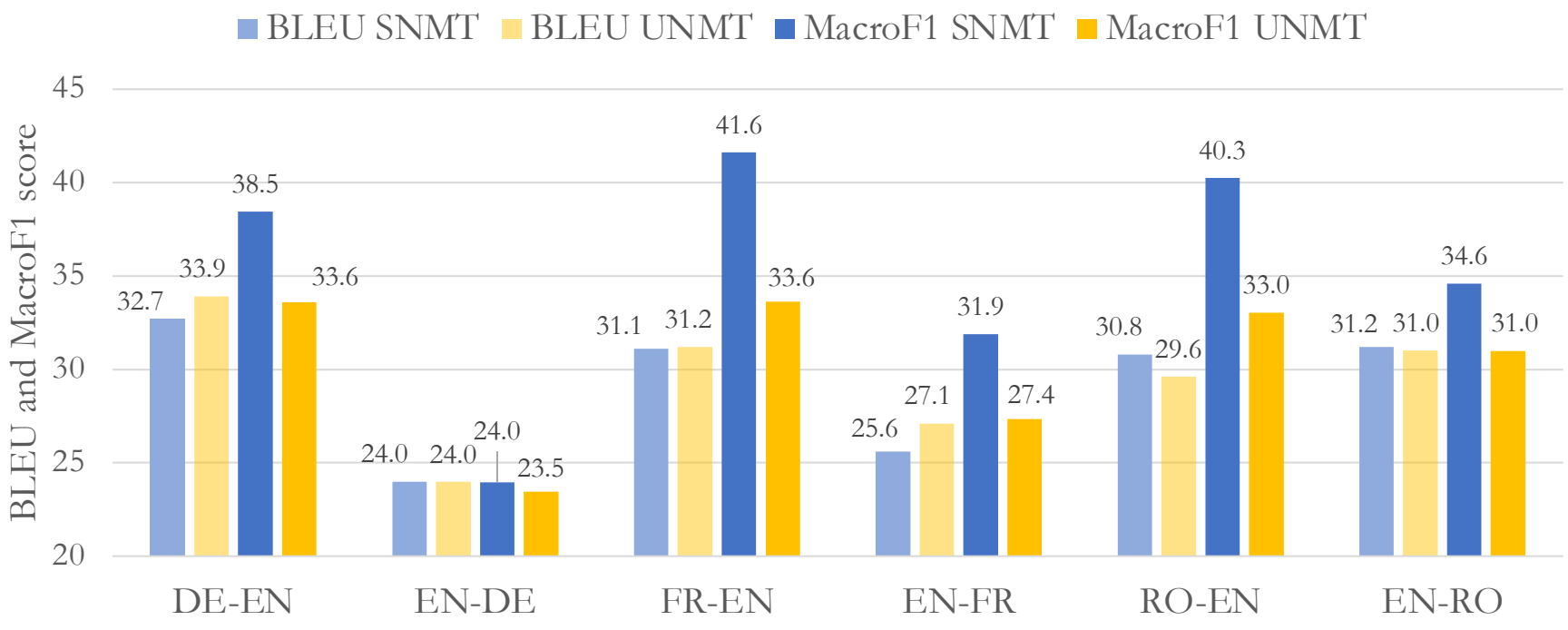


MacroF1 is the strongest indicator of downstream IR task performance

Difference Between Supervised and Unsupervised NMT Performance

(Collaboration with Weiqiu You)

SNMT vs UNMT: BLEU and MacroF1



In terms of BLEU, UNMT and SNMT performance is comparable,*
but MacroF1 shows significant differences between SNMT and UNMT

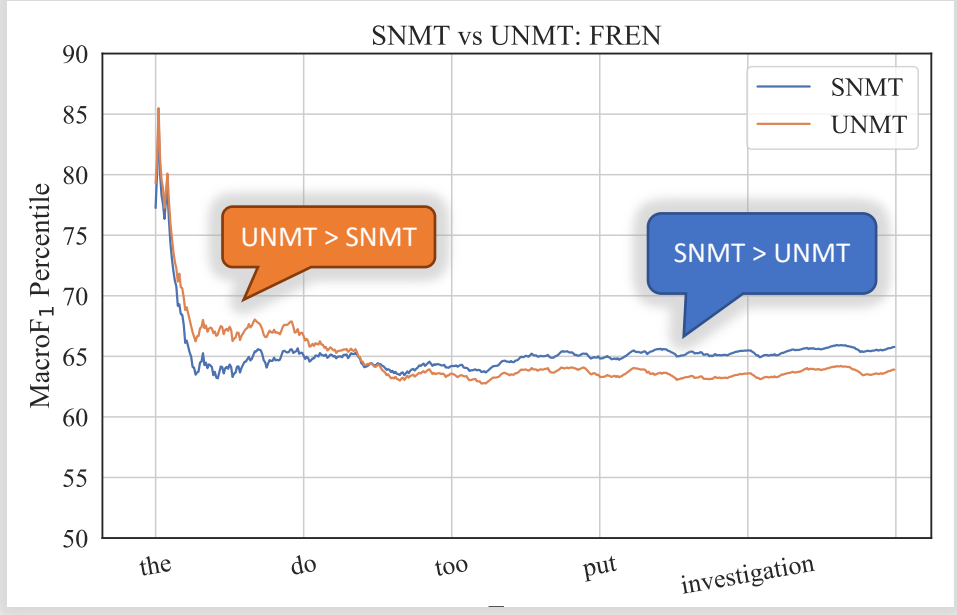
* SNMT systems were chosen to match BLEU scores with UNMT

MacroF1 Difference Between SNMT and UNMT

1. UNMT has better performance on frequent types, but SNMT outperforms on rare types
   ⇨ Approximately same BLEU
   but a huge difference in MacroF1

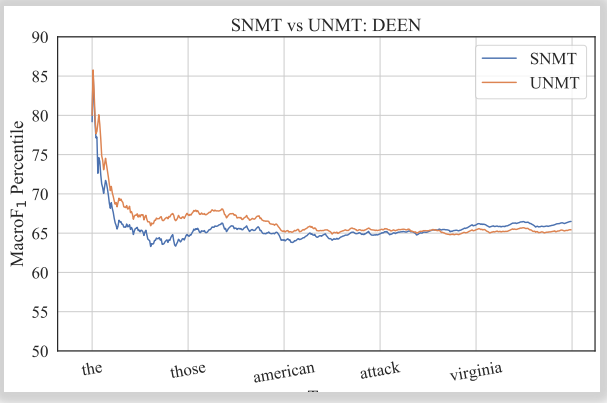
2. Both SNMT and UNMT have lower F1 score for content types than stopwords
   ⇨long-tail curse

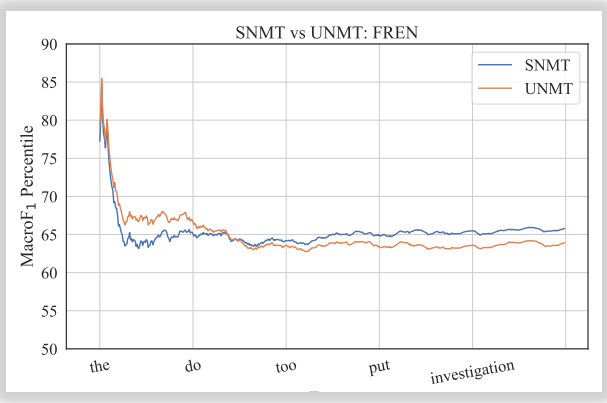
3. Other metrics do-not offer this level of breakdown. Try: BLEU, CHRF, BLEURT

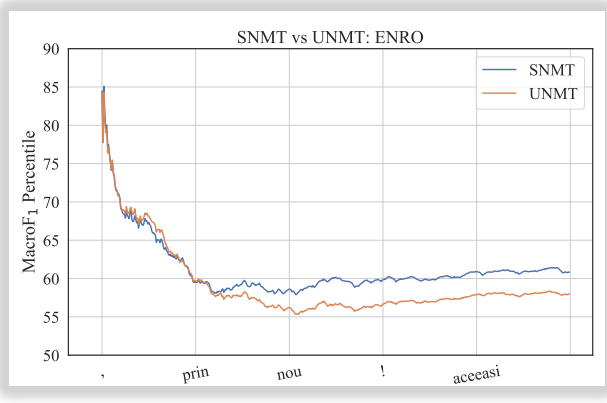


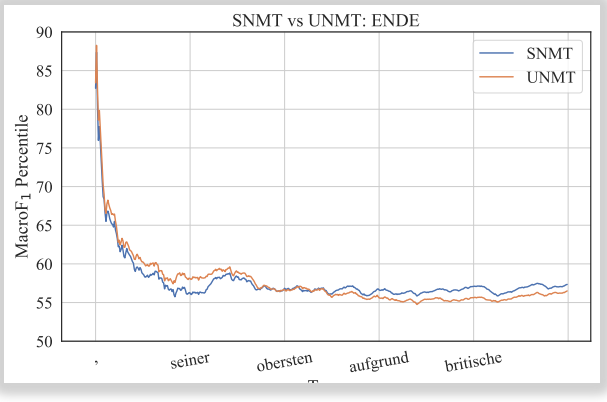
MacroF1 Difference Between SNMT and UNMT

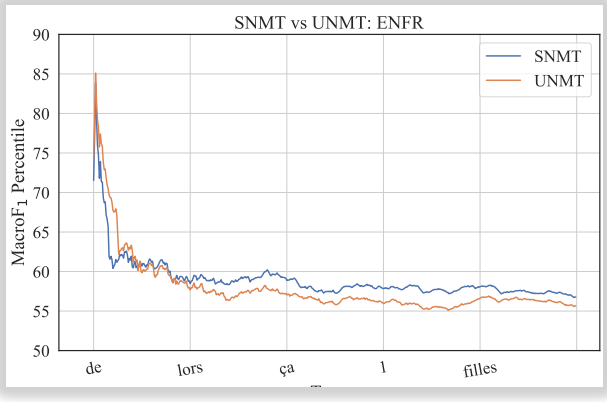
Similar trend across all languages: SNMT is better than UNMT on rare words

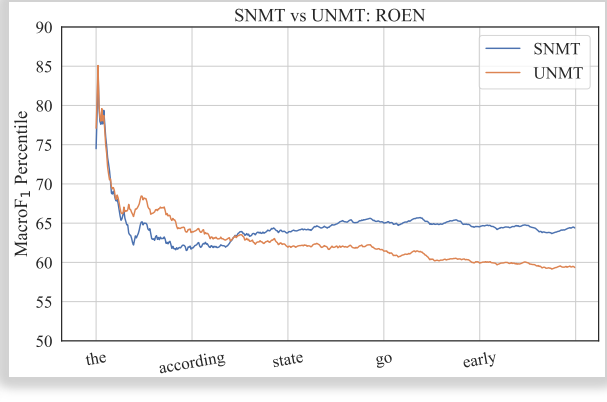












Part-II: Conclusion

Related Work

- A lot of MT evaluation metrics… but missed class/type imbalance
- Recent trend: model-based evaluation metrics, i.e., use one neural model to evaluate another
  - Biases!
  - Uninterpretable scores

Summary

- MacroF1 for MT evaluation
  - Competitive on direct human assessment (when all systems are fluent)
  - Outperforms others on downstream CLIR task
- BLEU and MacroF1 disagreement can be clearly seen on supervised vs unsupervised NMT

Part-III

Rare Languages

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

ACL 2021 Demos

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May

https://aclanthology.org/2021.acl-demo.37/

Language Coverage

- There are 7,000+ known living languages[1]
- Only about 100 languages are supported by popular MT
  - Google[2]: 108
  - Microsoft[3]: 103
- Research MT efforts are also constrained to the top 100 languages
- There are no MT models for thousands of languages

Population Range	Number of Languages			Number of speakers		
	Count	Percent	Cum%	Total	Percent	Cum%
100M - 1B	8	0.1	0.1%	2.8B	40.46	40.46%
10M - 100M	86	1.2	1.3%	2.8B	40.00	80.47%
1M - 10M	313	4.4	5.7%	1B	14.09	94.56%
100k - 1M	977	13.7	19.5%	310M	4.44	99.00%
10k - 100k	1,812	25.5	44.9%	62M	0.89	99.89%
1k - 10k	1,966	27.6	72.6%	7.5M	0.107	99.99%
100 - 1k	1,042	14.7	87.2%	0.5M	0.007	100%
10 - 100	305	4.3	91.5%	12k	0.0002	
1 - 9	114	1.6	93.1%	465	0.00001	
0	314	4.4	97.6%	0	0	
Unknown	174	2.4	100%			
Total	**7,111**			**7B**		

[1] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com . https://web.archive.org/web/20190401105648/https://www.ethnologue.com/statistics/size [2] https://blog.google/products/translate/five-new-languages/
[3] https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting/

Tools for Scalable NMT

Focus: open-source, reproducibility, and scalability

1. MTData: parallel dataset catalog and downloader
   - 120K+ datasets, hundreds of languages (As of June 2021)
   - 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; one `conf.yml` per experiment
   - All the necessary ingredients for NMT research to production

- `pip install mtdata nlcodec rtg` [4]

[1] https://github.com/thammegowda/mtdata/ [2] https://isi-nlp.github.io/nlcodec/ [3] https://isi-nlp.github.io/rtg/ [4] More NLP tools under PyPI https://pypi.org/user/Thamme.Gowda/

Datasets for 500+ Languages

- MTData created an index of all datasets
- Where are the datasets located? ⇨
- BixTex citation entries are shown whenever available
- Datasets comes in different formats
  (Standardization of language names, IDs etc
  under the hood)

Thanks to Kenneth Heafield (U Edinburgh) for their contributions

Stats as of Dec 2021

Source	Datasets
OPUS	80,830
OPUS_JW300	91,248
Neulab	4,455
Facebook	1,617
ELRC	1,341
EU	1,178
Statmt	699
Tilde	519
LinguaTools	253
Anuvaad	196
AI4Bharath	192
ParaCrawl	126
Lindat	56
UN	30
JoshuaDec	29
Phontron	4
NRC_CA	4
IITB	3
WAT	3
StanfordNLP	3
KECL	1
Total	**182.8K**

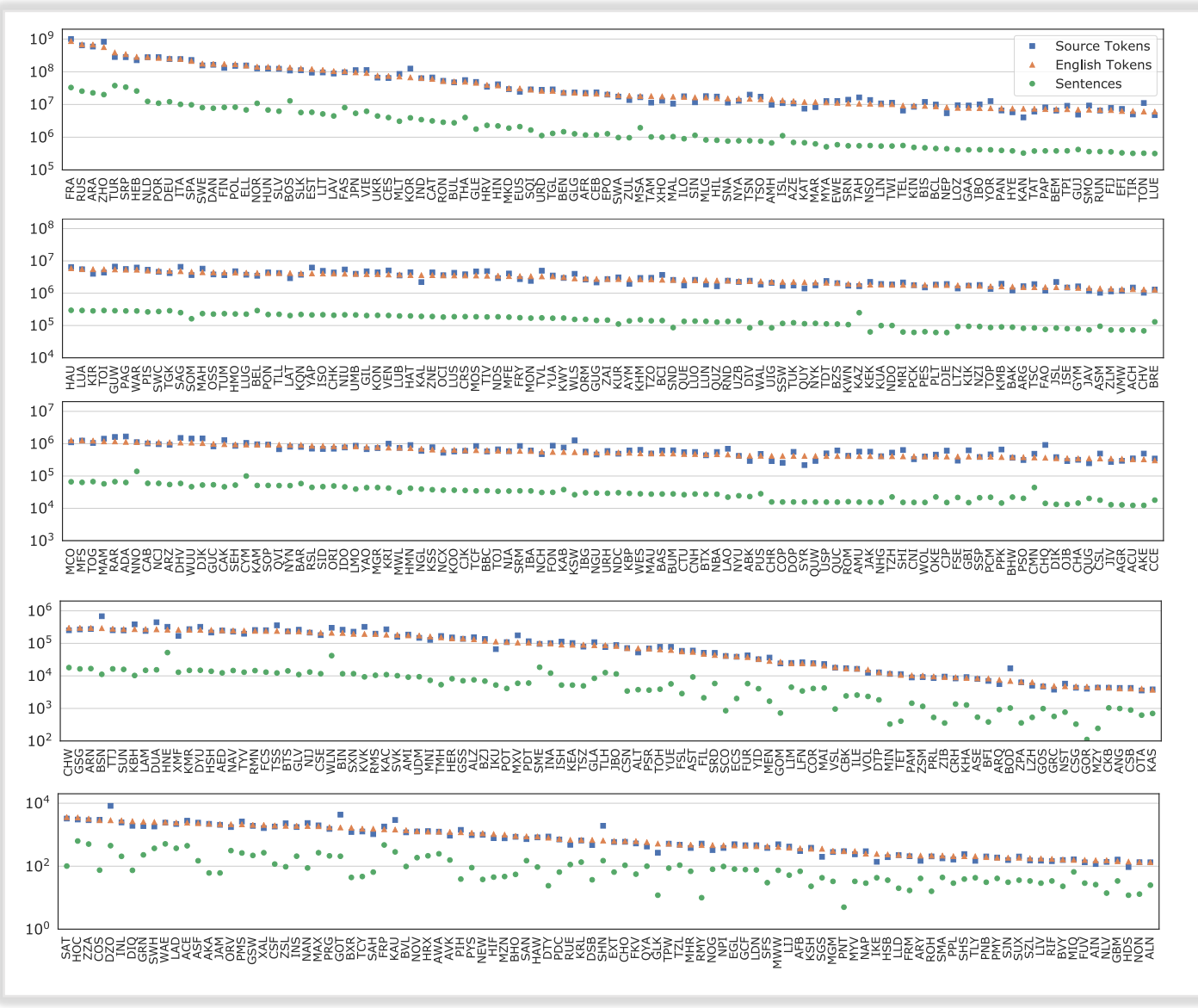
500 → English Translation

- 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
  - Gradient accumulation (5x), Float-16 ops, and distributed training
  - 8x A100 GPUs [Thanks to TACC and Zhao Zhang]
  - 5 days 6 hours for 200k steps

Dataset Sizes: 500 Languages

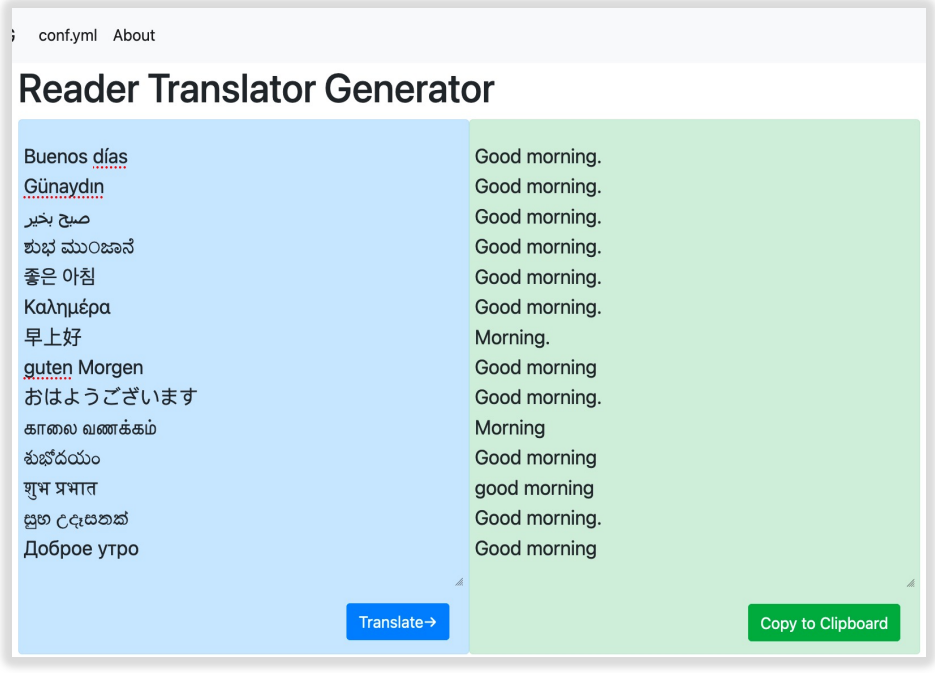
* ISO 639-3 codes



500-Eng : Pretrained Models, Data, Demo

- http://rtg.isi.edu/many-eng/
- Also available for download from docker

```
$ IMG= tgowda/rtg-model:500toEng-v1
$ docker run -p 6060:6060 $IMG
# For GPU support, add: --gpus '"device=0"'
```



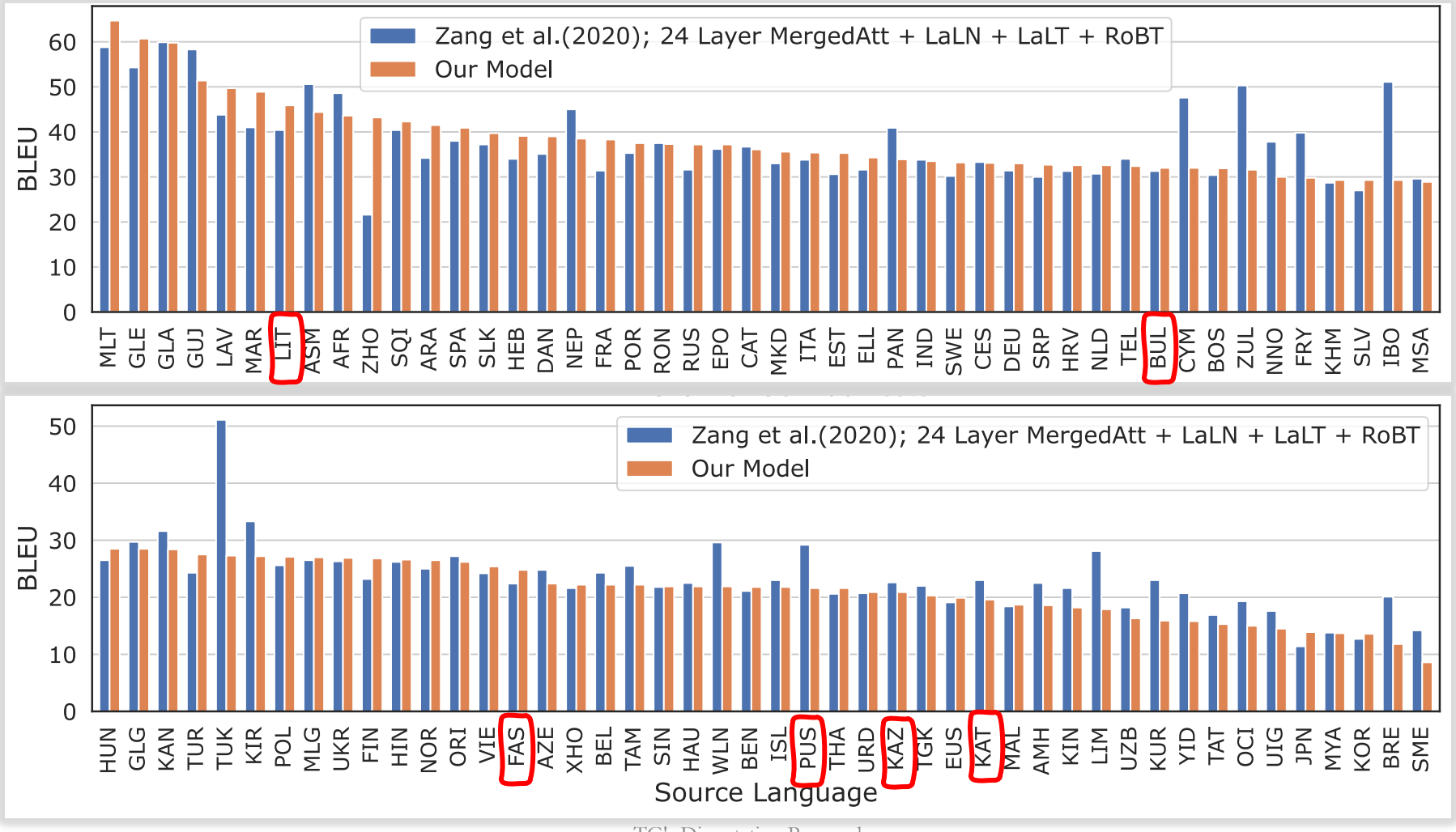
RTG Web Interface: REST API behind the scenes (via AJAX)

See also,

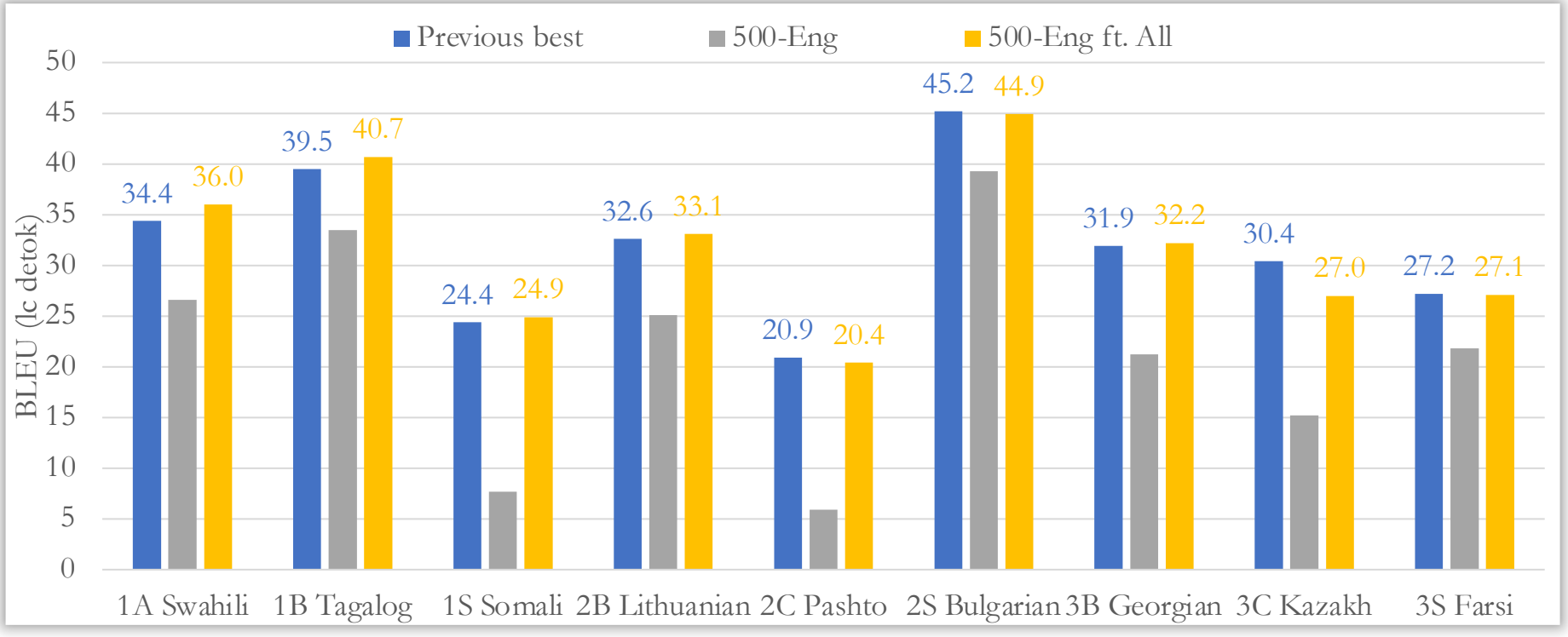
Apache Tika integration: https://cwiki.apache.org/confluence/display/TIKA/NMT-RTG

Spanish translation: https://www.ibidemgroup.com/edu/traduccion-machine-translation-datos-modelos/

Translation Service: BLEU on OPUS-100 Test Set

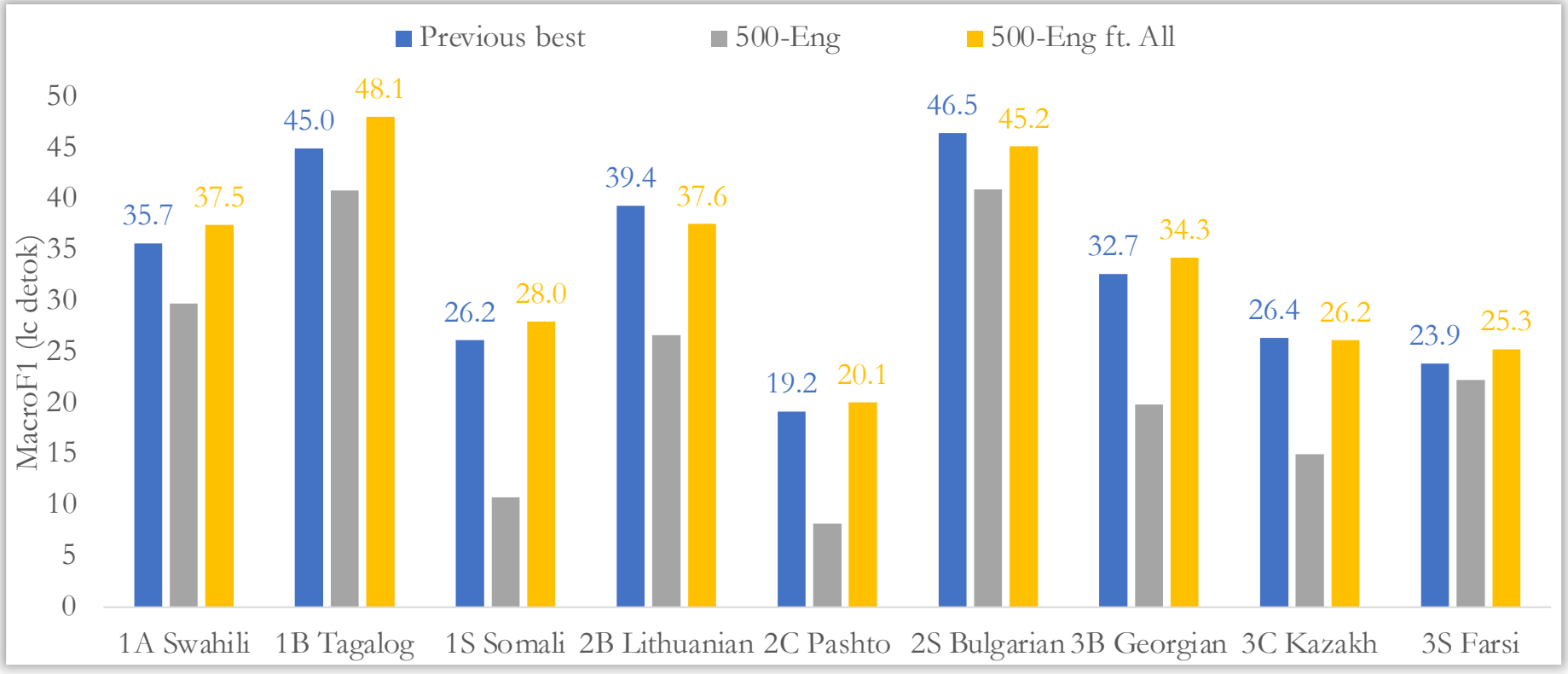


Finetuning: BLEU
on IARPA MATERIAL Datasets (Analysis)



"Previous best" : the best bilingual models used in evals; separate for each language

Finetuning: MacroF1
on IARPA MATERIAL Evaluations (Analysis)



"Previous best" : the best bilingual models used in evals; separate for each language

Part-III Conclusion

Related Work

- Google: 108, Microsoft: 103
  They support many-to-many
- OPUS 100, Facebook AI Research : 100 langs
  Many-to-Eng and Eng-to-Many
- Tatoeba Challenge
  ~500 languages dataset from OPUS;
  Mostly bilingual models

Summary

- Standardization of datasets
- Tools: RTG, NLCodec, MTData
- First ever 500-English multilingual model
- Translation service: available for free via docker
- State-of-the-art on low resource languages, via finetuning on a limited quantity of data

Part-IV

Robustness to Language Switching

Improving Multilingual Machine Translation Robustness via Data Augmentation

(Under review for NAACL 2022)

Thamme Gowda, Mozhdeh Gheini, and Jonathan May.

Problem

- Sometimes,
  - Multilingual speakers switch between languages
  - Part of text is already in target language
  - Sentences are not correctly segmented
- Are the current multilingual NMT models robust? No
- How to
  - Check robustness?
  - Improve robustness?

Image Credit: Amazon US/Funny Quotes Mugs

Original (Kan+Hin)	bandaaginda bari bageeche ke bahar-e iddivi. kahaani ke andhar bandu bidona. kaam bolo, saab.
English Translation	From the time I've reached, we've stayed outside the topic. Let's get into the story. Tell me the work, sir.

Example of language switching Kannada+Hindi

Robustness Checks

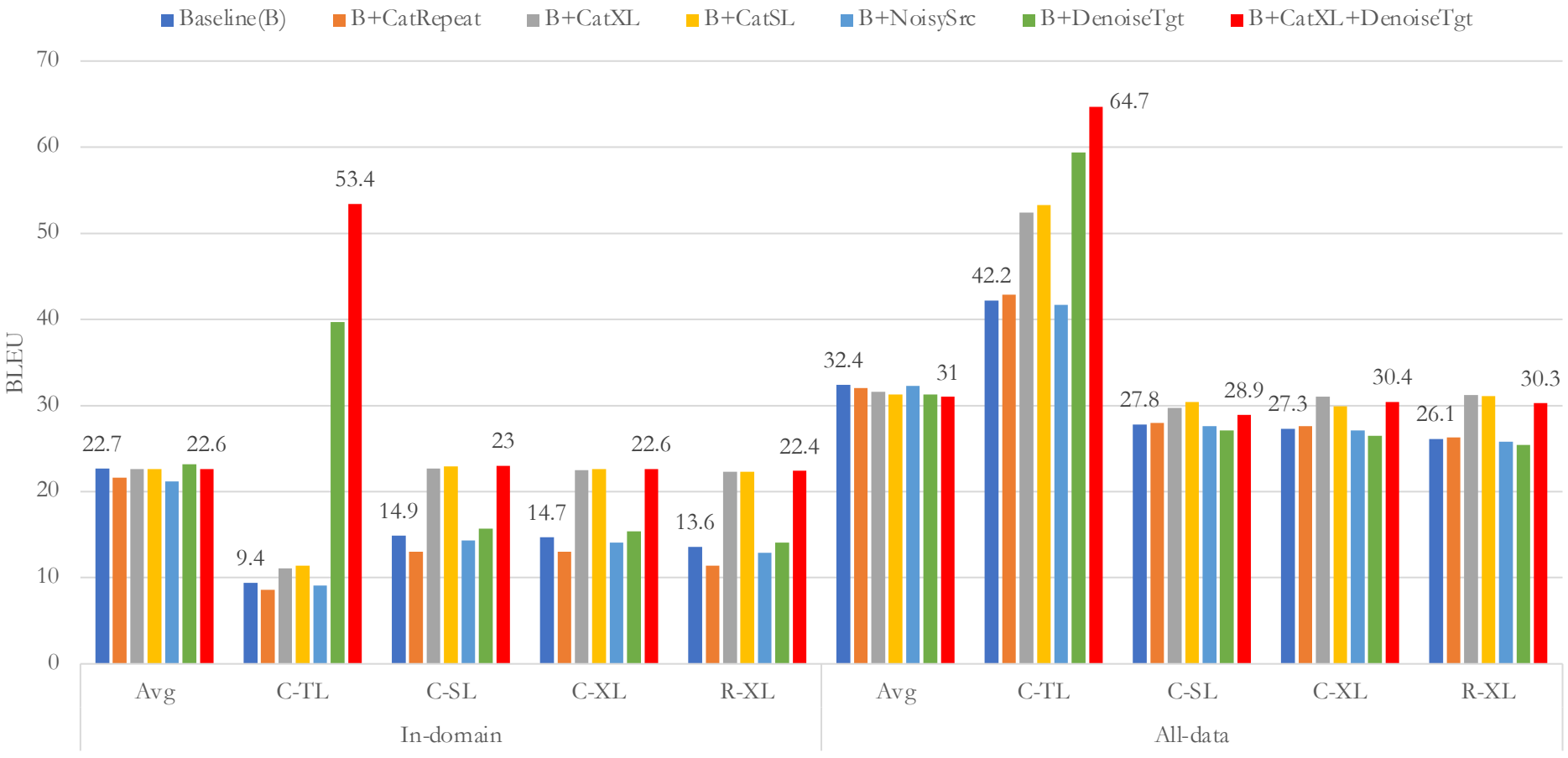
- Ribeiro et al 2020[1] proposed behavior testing for NLP in general
  - Simple modifications of test sets; negation, synonym, NER replacement, etc,
  - Tasks: sentiment analysis, duplicate question detection, span detection
  - *They have not addressed MT task; their modifications are not trivial in MT task*
- But the idea of behavior testing for MT is interesting
- Proposal: Create more Tests by concatenating test sentences
  - **C-SL**: consecutive **same** language
  - **C-TL**: consecutive **target** language
  - **C-XL**: consecutive **cross**-language
  - **R-XL**: **random cross**-language

[1] https://aclanthology.org/2020.acl-main.442/

Experiment Setup

- Workshop on Asian Translation 2021 MultiIndic Task
  - 10 Indian languages → English
  - Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu
- Training data:
  - Limited quantity in-domain data
  - All-data having out-of-domain data (larger)
  - Dev and test sets are from WAT21
- More held-out sets are created as per previous slide (robustness checks)
- Training data augmentations:
  - Concatenation: CatRepeat, CatSL*, CatXL* NOTE: space char between joins
    *: random sentences, SL: same language, XL: cross language
  - Noise: DenoiseTarget, NoisySource
    What noise? 10% of random word drop, replacement and word order shuffle

Results: BLEU



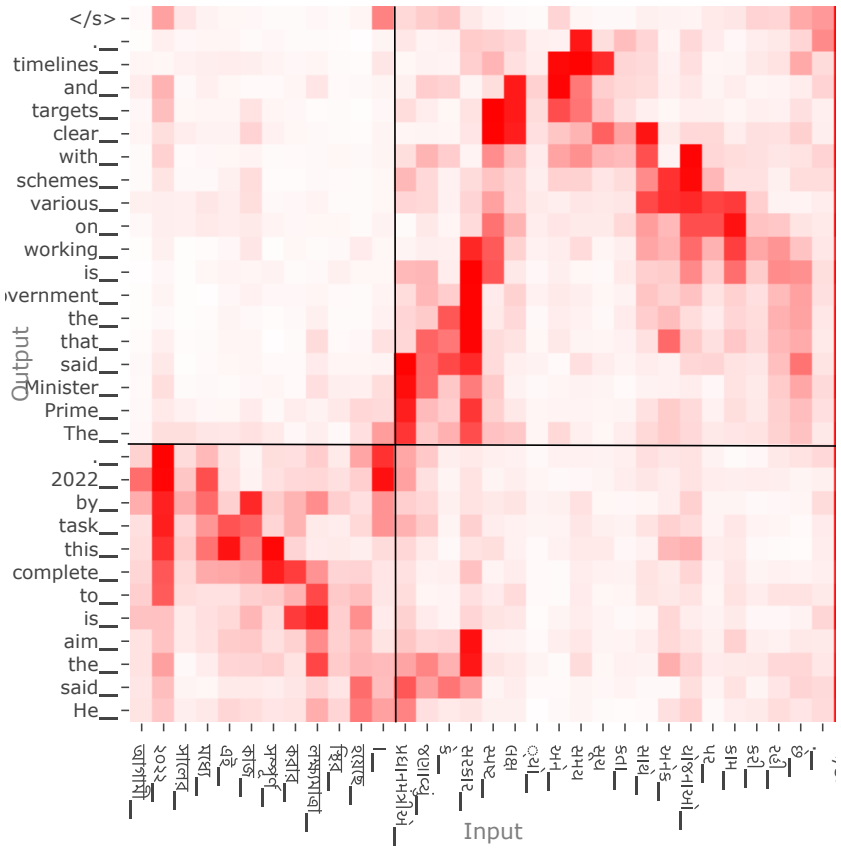
Improvements are not visible on the original (Avg) set, but proposed checklist sets showcase it

Qualitative Examples

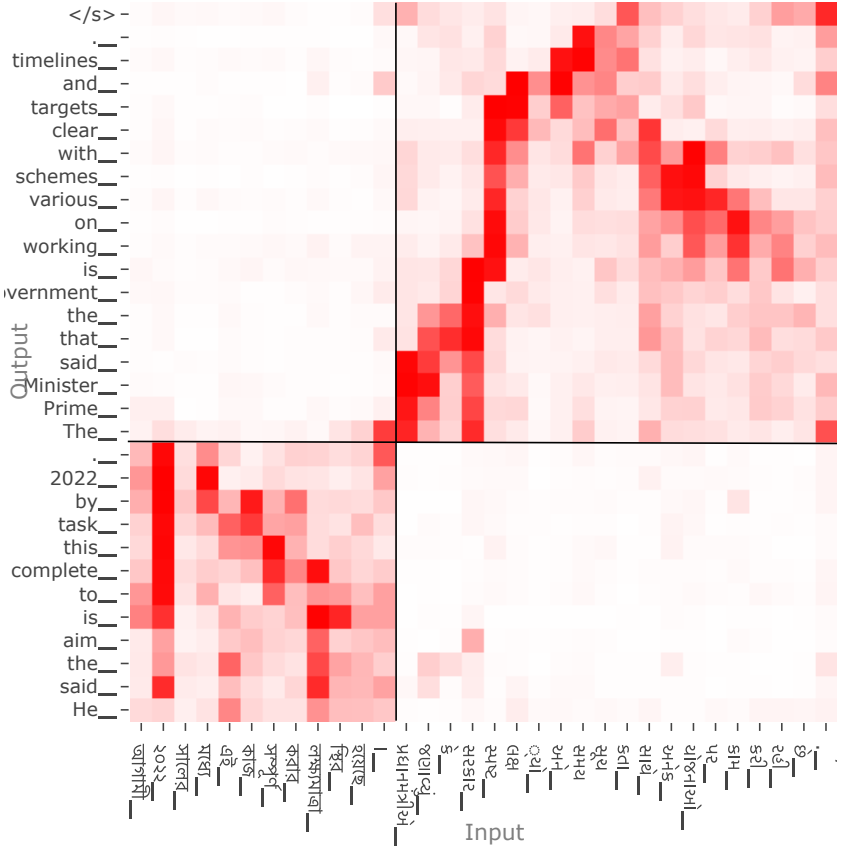
Source	আগামী ২০২২ সালের মধ্যে এই কাজ সম্পূর্ণ করার লক্ষ্যমাত্রা স্থির হয়েছে। પ્રધાનમંત્રીએ જણાવ્યું કે સરકાર સ્પષ્ટ લક્ષ્યો અને સમયસૂચકતા સાથે અનેક યોજનાઓ પર કામ કરી રહી છે.
Reference	He said the aim is to complete this task by 2022. The Prime Minister said that the Government is working on various schemes with clear targets and timelines.
Baseline	He said the Government is working on several schemes with clear objectives and timelines.
B+CatRepeat	The target is to be completed by 2022, the Prime Minister said that the Government is working on several schemes with clear targets and timelines. is the of of of of of of of of of of of of
B+CatXL	The target is to complete it by 2022. The Prime Minister said that the Government is working on a number of schemes with clear targets and timelines.
B+CatSL	We have set a target to complete this task by 2022. The Prime Minister said that the Government is working on a number of schemes with clear objectives and timelines.
B+NoisySrc	The Prime Minister said that the Government is working on several schemes with clear objectives and timelines.
B+DenoiseTgt	He said the Government is working on several schemes with clear objectives and timelines.
B+CatXL +DenoiseTgt	We have set a target of completing it by 2022. The Prime Minister said that the Government is working on a number of schemes with clear targets and timelines.

Example translations from models trained on all-data

Attention Visualization



Baseline

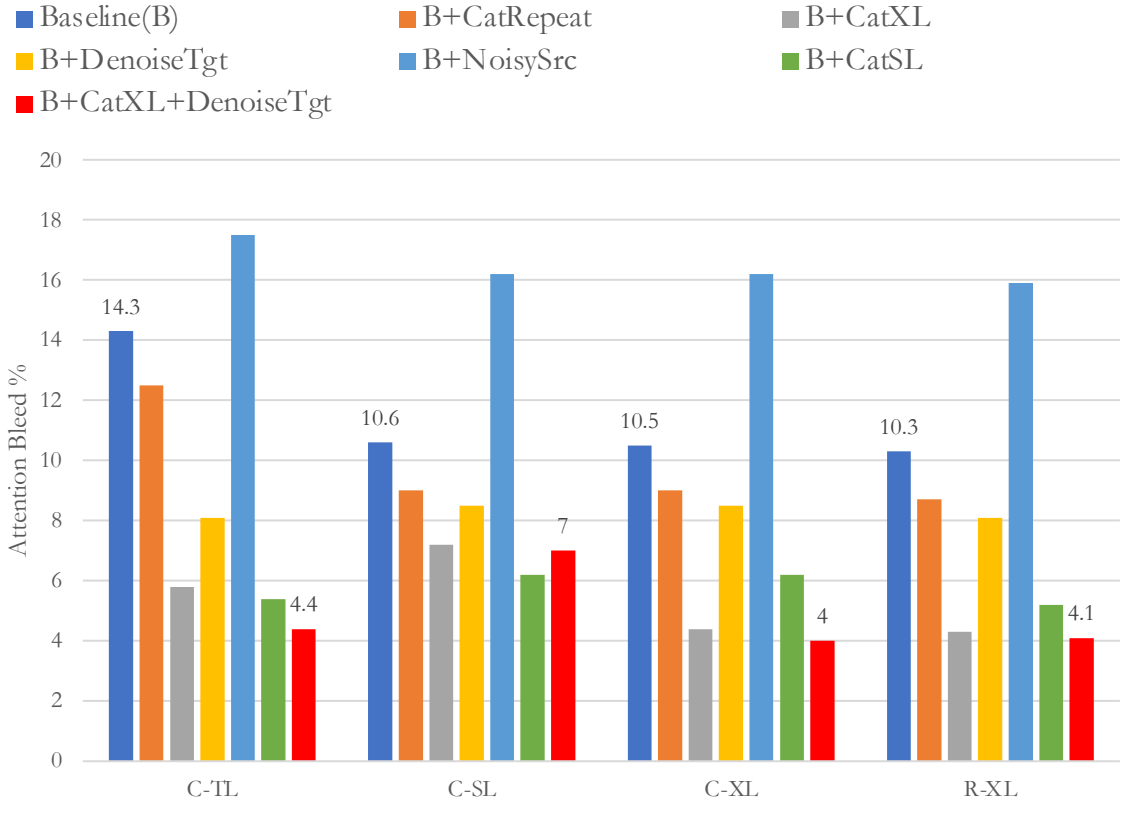


Model with augmented data (CatXL+Denoise)

Models trained with augmented data learns better cross-attentions

Attention Bleed

- Cross-attention mass crossing sentence boundaries in concatenated test sentences
  - Lower is better
- Average *attention bleed* across
  - All sentences
  - Transformer layers
  - Attention heads
- Models trained on augmented sentences achieve lower bleed
  ➔ Learn better attention



Attention bleed in models trained on all-data

Part-IV Conclusion

Related Works

- Back translation:
  Costly in massively multilingual setup
- Most other robustness works are concerned about bilingual MT, and noisy data

Summary

- Current multilingual MT models are not robust to language switching
- Proposed robustness checks
- Investigated augmentation methods to improve robustness
- Models trained with sentence concatenation and denoising achieve
  - Best scores on robustness tests
  - Learn better attention mechanisms

Current/WIP Research

- *[WIP]* Combine Part-III and Part-IV:
  - Robustness across 500+ multilingual NMT
    In part-III, we used dataset from WAT21 shared task
- *[WIP]* Part-III revised:
  - +100 more languages (Up to 600 languages)
  - More datasets have been found on web and included in `mtdata`

Implications

	Before	After
NMT Architecture	Autoencoder: i.e., Encoder-Decoder	Autoregressor + Classifier [1]; more emphasis on the target vocabulary and data imbalance
Vocabulary Size	Did not know why some are best; Arbitrarily chosen or via grid search for each dataset	Heuristic to auto adjust vocab size! Byte pair-encoding (BPE) size is chosen to minimize sequence lengths and improve class balance [1]
Evaluation Metrics	Treat each 'token' equally; Stopwords have more weight	Treat each 'type' equally [2]; All types have equal weight. Address data imbalance at evaluation
Scaling NMT	To ~100 languages	To ~500 languages [3] Bunch of useful tools, datasets; Standardization of dataset IDs
Multi-lingual lang. switching	Not robust	Robust to language switching [4] Can translate text that start in one language and finishes in another. Robustness to partly translated text

[1] **Gowda** and May, *'Finding the optimal vocabulary size for NMT'*, EMNLP 2020 Findings
[2] **Gowda** et al, *'Macro-average: Rare types are important too'*, NAACL 2021
[3] **Gowda** et al, *'Many-to-English tools, data, and pretrained models'*, ACL 2021 Demos
[4] **Gowda** et al, *'Improving multilingual MT robustness via data augmentation,* [under review/NAACL2022]

These projects helped in shaping some of the ideas and skills for PhD work!

Works Outside MT [Since MS @ USC]

IE / NER

Mehrabi, N., **Gowda, T.**, Morstatter, F., Peng, N., & Galstyan, A. (2020, July). 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* (pp. 231-232).

Pan, X., **Gowda, T.**, Ji, H., May, J., & Miller, S. (2019, November). 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)* (pp. 56-66).

Wagstaff, K., Francis, R., **Gowda, T.**, Lu, Y., Riloff, E., Singh, K., & Lanza, N. (2018, April). Mars Target Encyclopedia: Rock and Soil Composition Extracted from the Literature. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Image Classification

Gowda, T., Hundman, K., & Mattmann, C. A. (2017, June). An approach for automatic and large-scale image forensics. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security* (pp. 16-20).

Wagstaff, K. L., Lu, Y., Stanboli, A., Grimes, K., **Gowda, T.**, & Padams, J. (2018, April). Deep Mars: CNN classification of mars imagery for the PDS imaging atlas. In Thirty-Second AAAI Conference on Artificial Intelligence.

Web Data Mining

Hundman, K., **Gowda, T.**, Kejriwal, M., & Boecking, B. (2018, December). Always lurking: understanding and mitigating bias in online human trafficking detection. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 137-143).

Gowda, T., & Mattmann, C. A. (2016, July). Clustering web pages based on structure and style similarity (application paper). In *2016 IEEE 17th International conference on information reuse and integration (IRI)* (pp. 175-180). IEEE.

Mattmann, C. A., Yang, G. H., Manjunatha H., **Gowda, T.**, Zhou, A. J., Luo, J., & McGibbney, L. J. (2016). Multimedia metadata-based forensics in human trafficking web data.

Bias in ML models

Scaling up

Next Steps in Ph.D.

Ph.D. requirements: 60/60 ✅

[02/15] Many-English v2
(Robustness across 500+100 languages)

[02/28] Currently under-review paper publication

[02/28] Finish dissertation draft

[03/15?] Review the draft with committee members and address comments

[March?] Dissertation defense

Submit manuscript
to grad school

USC deadlines:

- Spring 22 : 03/18
- Summer 22 : 07/01

Q & A

USC Viterbi
School of Engineering

The Inevitable Problem of Rare Phenomena Learning in Machine Translation

Dissertation Proposal

by

Thamme Gowda

Committee

Jonathan May (advisor)
Chris Mattmann
Xuezhe Ma
Aiichiro Nakano
Shri Narayanan
Xiang Ren