CUTE LAB NAME

USC Viterbi School of Engineering

The Inevitable Problem of Rare Phenomena Learning in Machine Translation

Dissertation Defense

by

Thamme Gowda

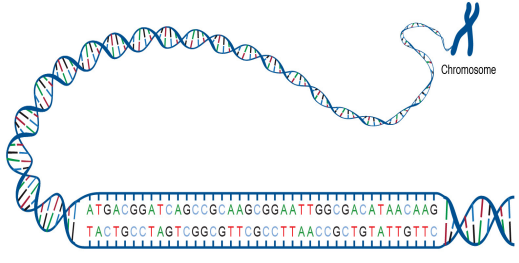
March 28, 2022

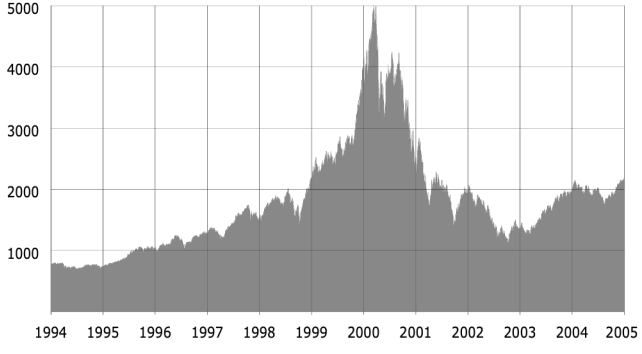
Committee

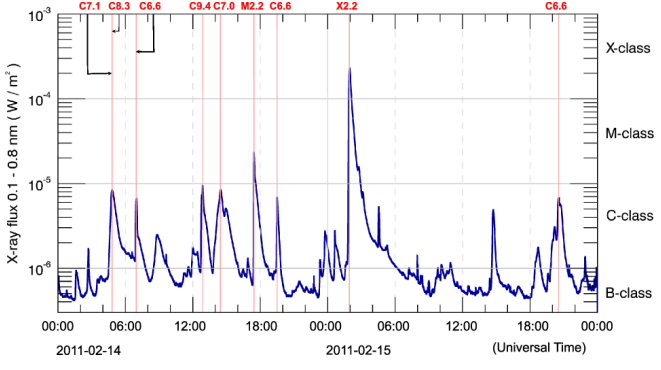
Jonathan May (advisor)
Chris Mattmann
Xuezhe Ma
Aiichiro Nakano
Shri Narayanan
Xiang Ren

Rare Phenomena Learning Problem

- Naturally occurring categorical observations are often imbalanced, some categories are frequent (majority classes), and others are rare (minority classes) E.g., cancer detection, fraud detection
- ML classification methods perform subpar on rare categories, but improved performance on rare categories is necessary in real world applications
- Rare phenomena in sequential data: genome, financial market events, space weather, sensor readings … and natural language text!

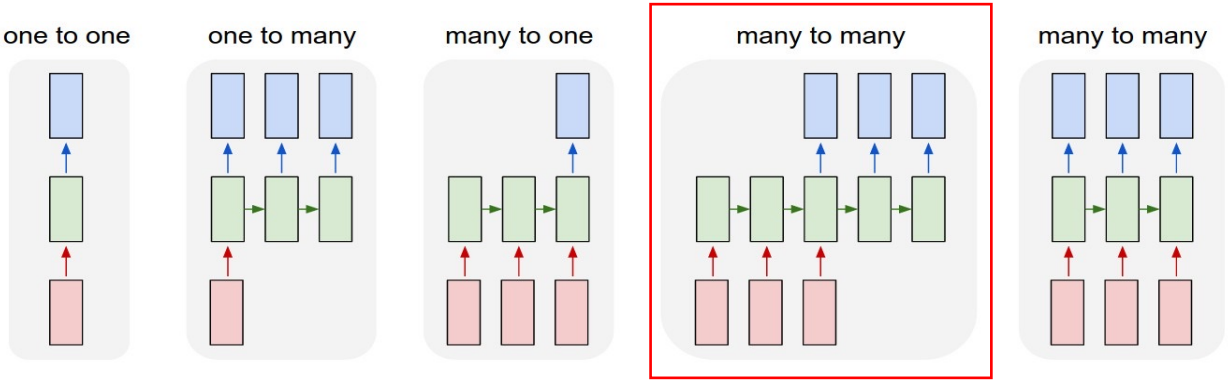






Natural Languages have Imbalanced Types

- A few types are frequent (e.g., stopwords), and many others are rare
- Rare words carry more information content, so they can not be ignored
- **Sequence-to-sequence learning** is a general problem for sequences
  - Transformers[1] were originally shown effective on seq-to-seq task (e.g., MT), but now used on all other taskon sequences
- Focus on machine translation (MT) for the rest of the talk



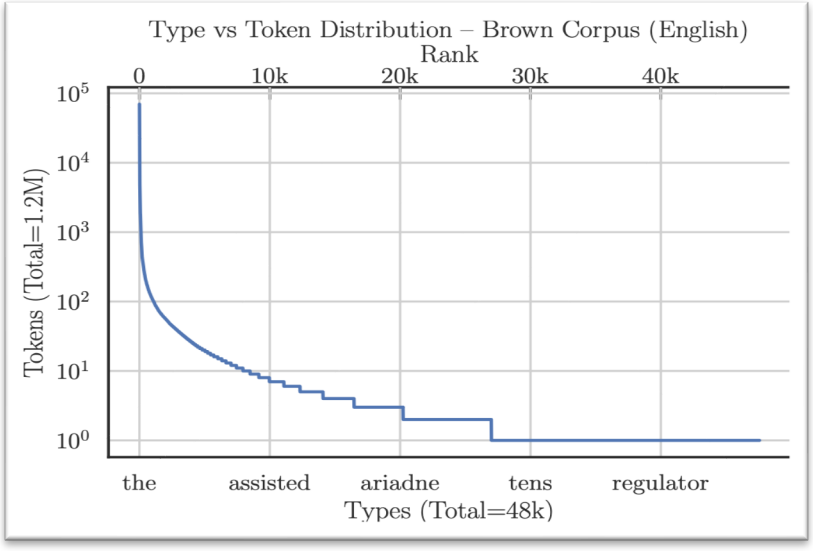
[1] Vaswani et al. 2017, "Attention Is All You Need"

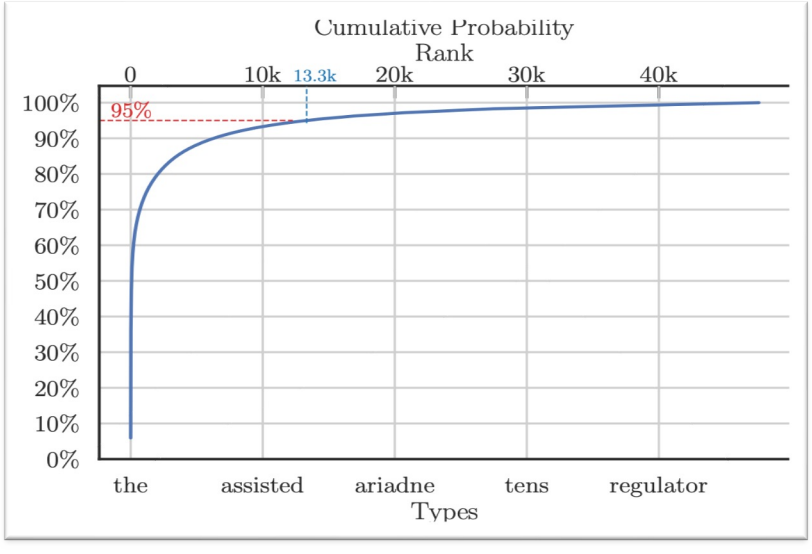
Image Credit: Andrej Karpathy

“ The Long-tail Curse

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] Mark Steedman. 2008. "Last Words: On Becoming a Discipline." CL, https://aclanthology.org/J08-1008/

Thesis Statement

- The need for improved performance on rare categories is ubiquitous across machine learning applications
- In machine translation (MT), this problem is inevitable and manifests in various forms:
    1. Rare words at training
    2. Rare words at evaluation
    3. Rare linguistic styles such as code-switching
    4. Rare languages
- By addressing these areas, we improve our ability to build higher quality, more comprehensive models

Overview

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 – Robustness – code-switching

"Improving robustness in MT via data Augmentation" [Under review]

IV – Rare languages (600+)

"Many-to-English MT tools, data, and pretrained models" [ACL 2021 Demos]

V – Discussion, Future Directions

Thanks

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
- Kenneth Heafield, U Edin

Committee

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

Acknowledgments

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

Computing Resources

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

⇨ I – Rare words in training

"Finding the optimal vocabulary size for NMT" [EMNLP 2020 Findings]

II – Rare words in evaluation

"M

II

"I

IV

"Many-to-English MT tools, data, and pretrained models" [ACL 2021 Demos]

V – Discussion, Future Directions

I – Rare words in training

Finding the Optimal Vocabulary Size for NMT

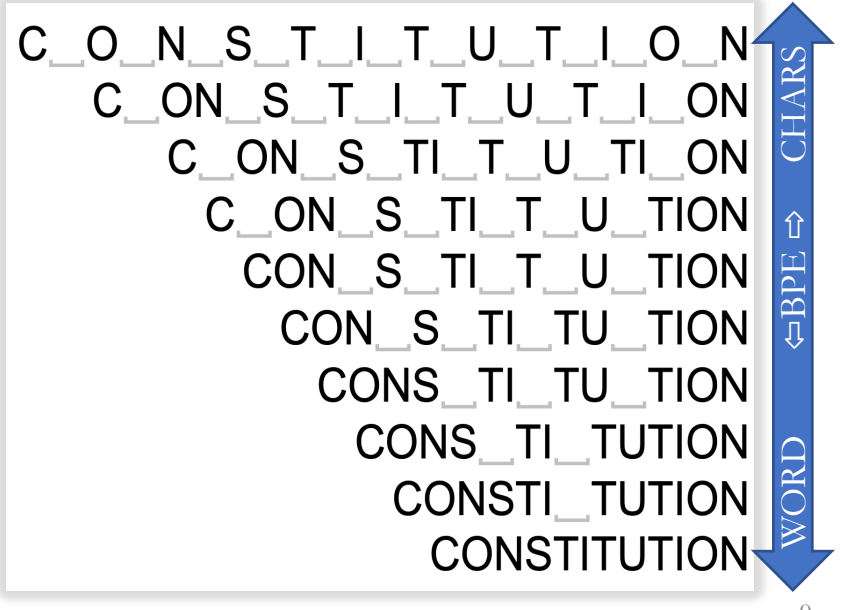
EMNLP Findings 2020

Thamme Gowda and Jonathan May

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

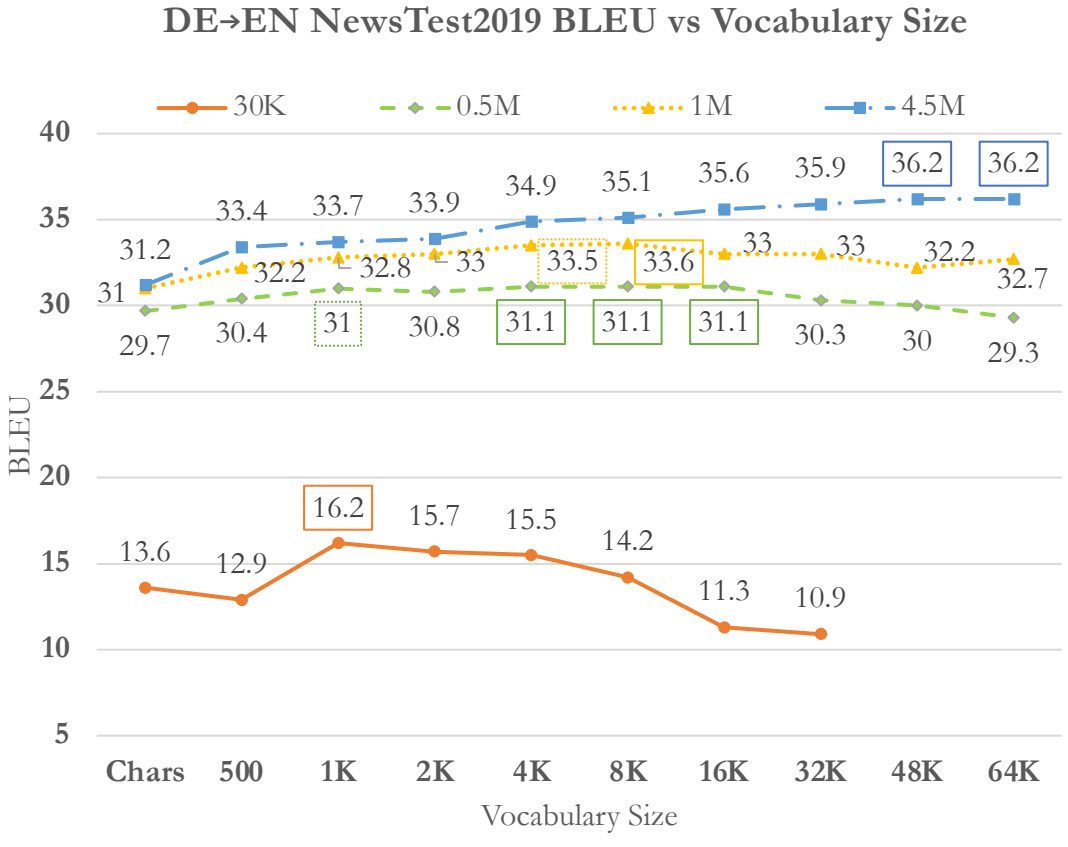
Neural Machine Translation (NMT)

- Sequence transduction, $f: (x_1 x_2 x_3 \dots x_m) \rightarrow (y_1 y_2 y_3 \dots y_n)$
- Maximize $P(y_{1:m} | x_{1:m})$ ⇨ Maximize $\prod_{t=1}^{n} P(y_t \mid y_{<t}, x_{1:m}; \theta)$
- $y_{1:n}$ = Decoder(Encoder($x_{1:m}$))
  - Implementations: RNN (LSTM, GRU), CNN, Transformers
- Q: What is the impact of data imbalance?
- Byte-pair-encoding sub words
  [Sennrich et al 2016[1]]
  - Addresses rare word (OOV) generation problem
- Subwords are obtained by merging most frequent chars and subwords
  - Better performance than chars, words
- Number of merges is a hyper parameter



Experiments

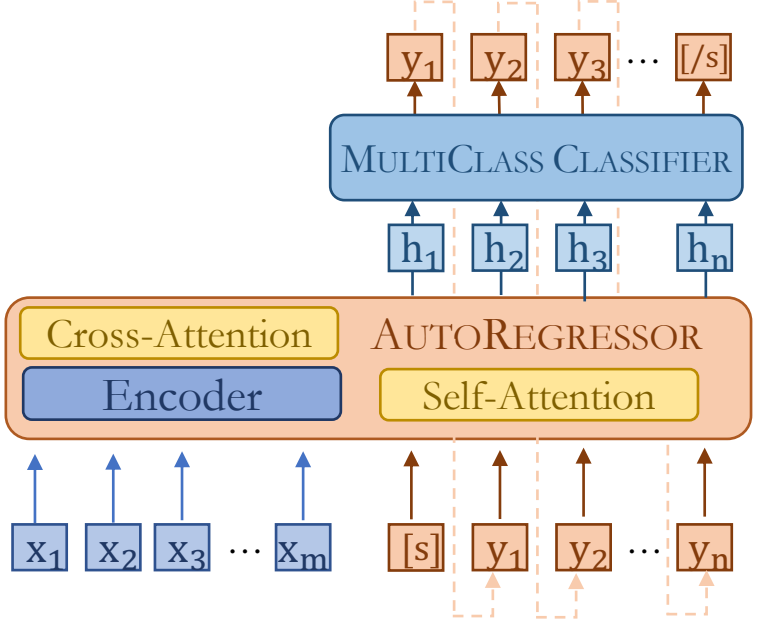
- Four target languages: DE→EN, EN→DE, EN→HI, and EN→LI
- Total of 11 dataset sizes: between 30K and 4.5M sentences
- × 10 vocabulary sizes: chars, 500, 1K, 2K, 4K, 8K, 16K, 32K, 48K, and 64K
- **Transformer** with 6 layers, 512 dims, 8 attn heads, 0.1 dropout, … Trained as per the best practices for training transformers
- DE→EN shown; trend is similar on other language pairs



All BLEU lines are (sort of) concave down on vocabulary size; why?

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



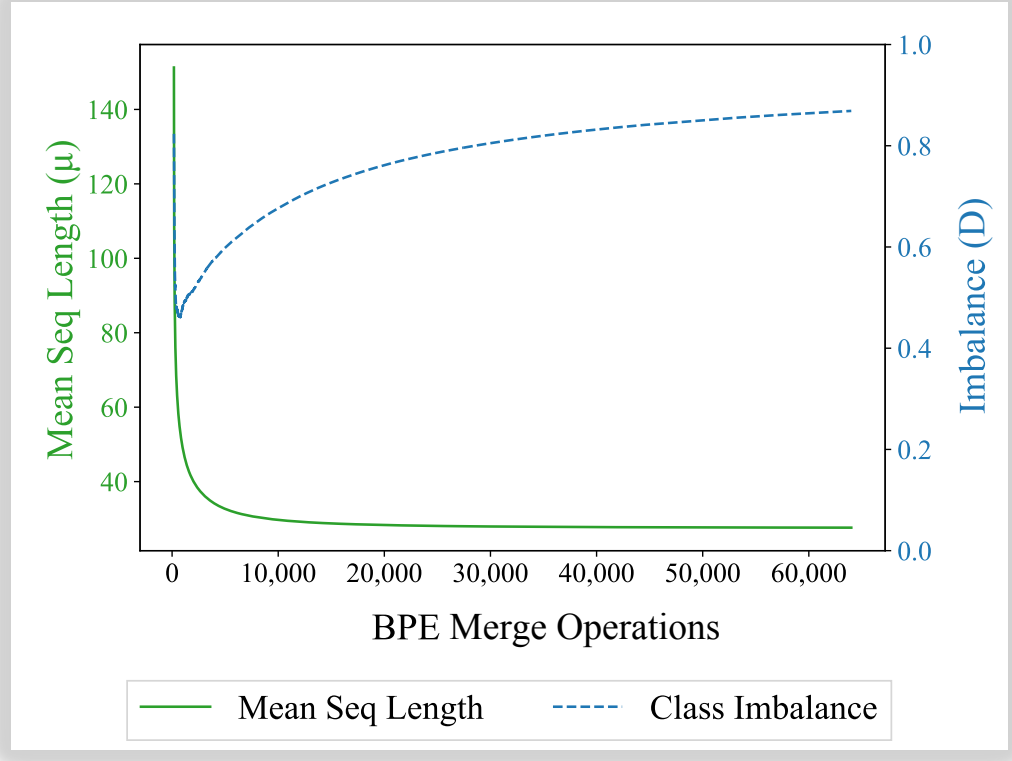
High-level abstraction of Transformer NMT

Imbalanced Classification

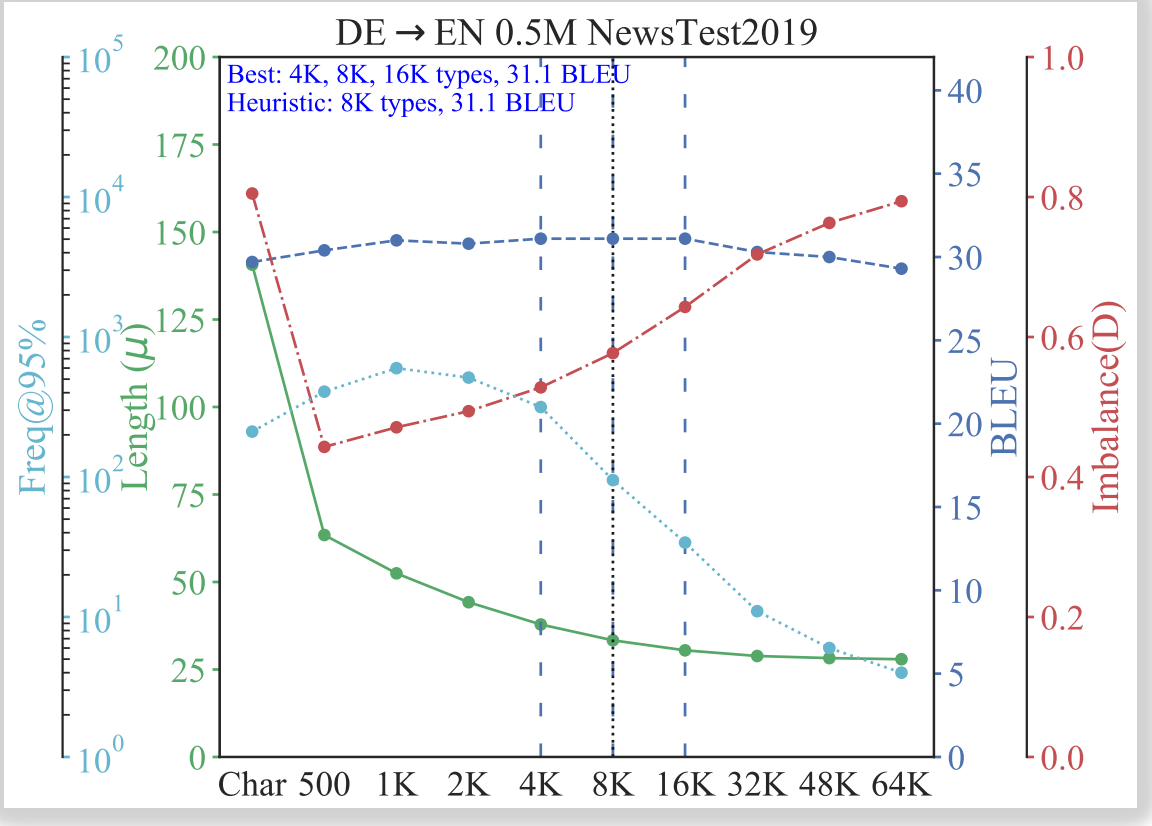
- Type to token distribution is skewed
- Classifiers are known to possess frequency-based biases. Minority classes are often ignored, i.e., poor recall
- Imbalanced classification learning:
  - Sampling methods: not feasible
  - Weighted cross entropy, focal loss, etc., did not improve performance
- Byte pair encoding (BPE): balances classes via splitting and merging

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 vocabulary size is the one that achieves a good trade-off



Heuristic



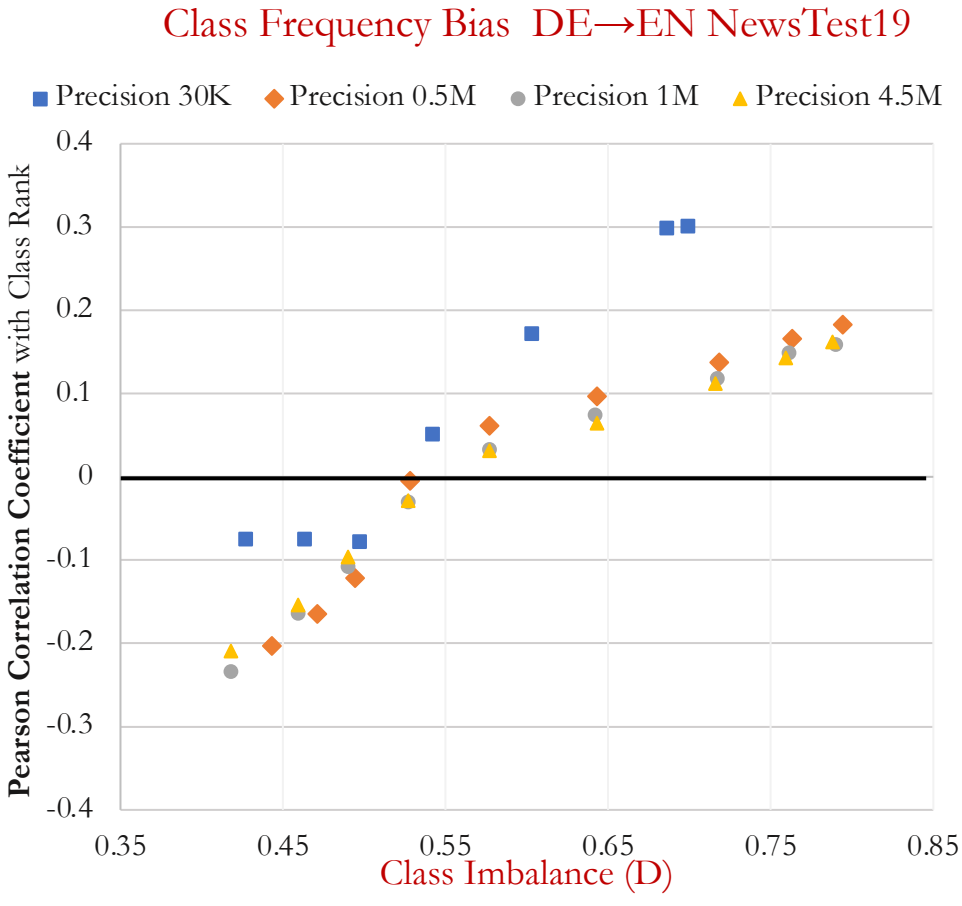
"Use the largest possible BPE vocabulary such that at least 95% of classes have about 100 or more training examples"

Frequency-based Bias on Classes

- Pearson correlation coefficient
- Rank test set classes based on training frequency
- Correlation between Rank and Precision

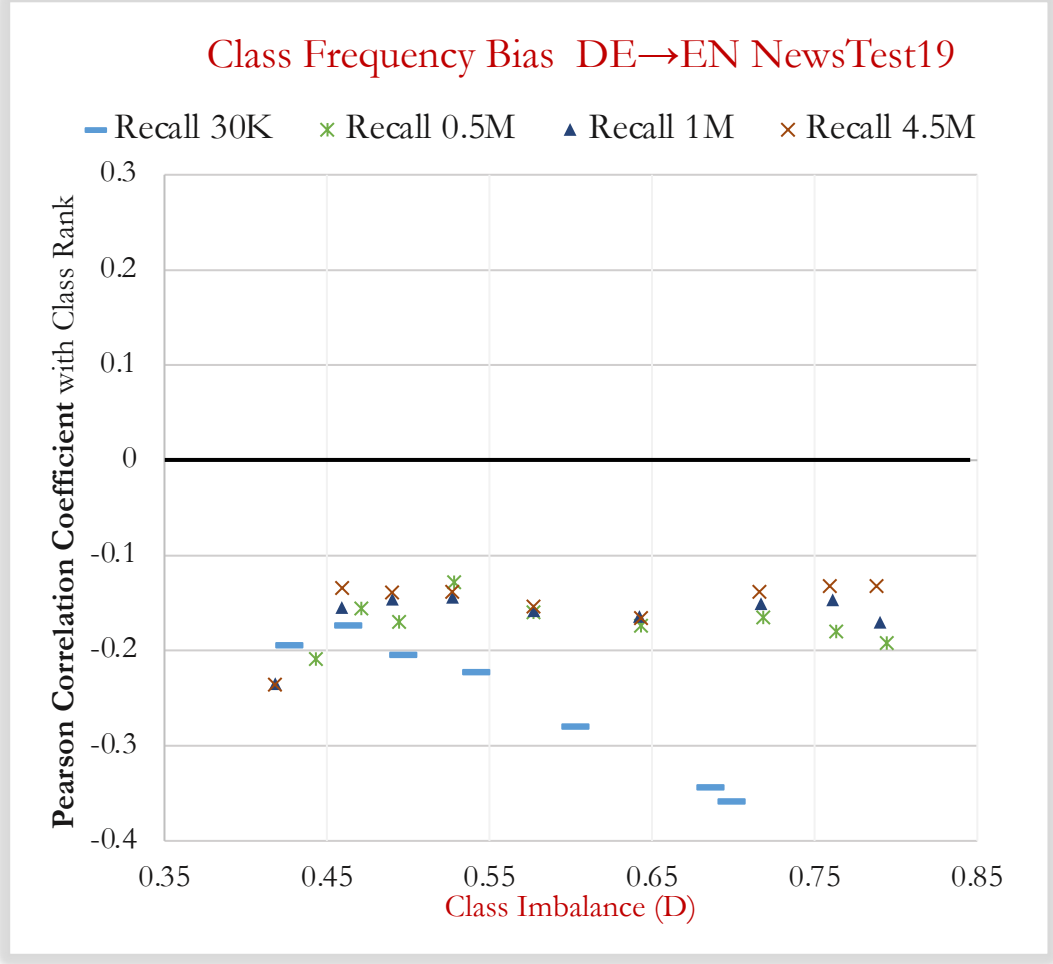
Related: 'the the the the…' problem

- By tuning the vocabulary size, and there by tuning class balance, precision can be made uncorrelated with rank



Frequency-based Bias on Classes

- Pearson correlation coefficient
- Rank test set classes based on training data frequency
- Correlation between Rank and Recall
- Rare classes have poor recall



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
  - Useful 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
- Rare types have lower recall than frequent types

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

✅ I – Rare words in training

"Finding the optimal vocabulary size for NMT" [EMNLP 2020

⇨II – Rare words in evaluation

"Mac

III –

"Imp w]

IV – Rare languages (000's)

"Many-to-English MT tools, data, and pretrained models" [ACL 2021 Demos]

V – Discussion, Future Directions

II – Rare words in evaluation

Macro-Average: Rare Types are Important Too

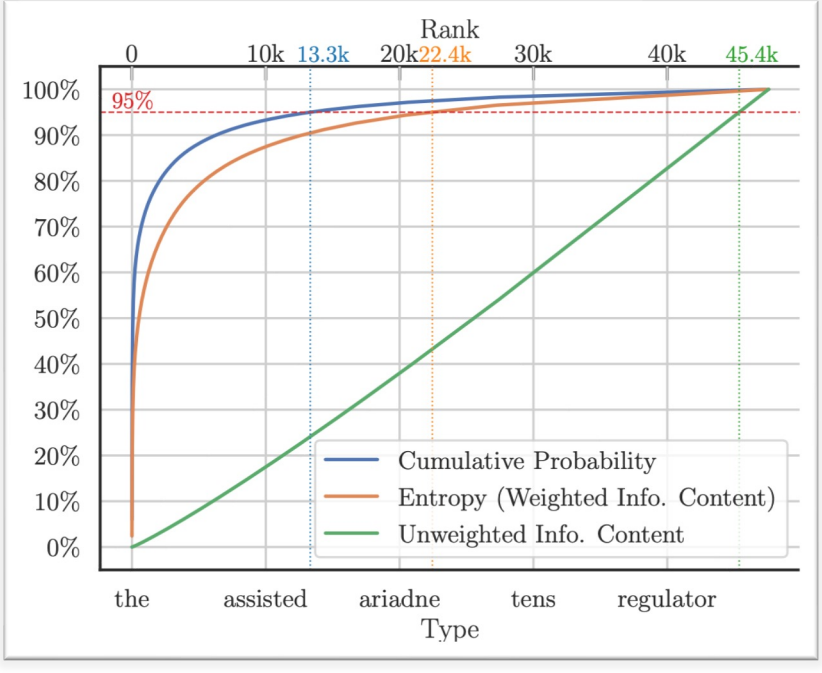
NAACL 2021

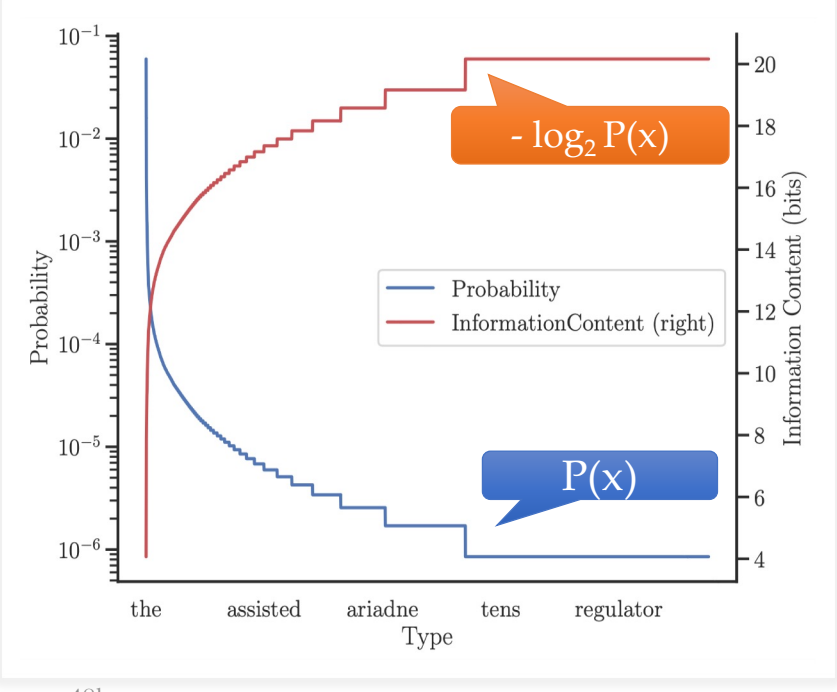
Thamme Gowda, Weiqiu You, Constantine Lignos, and Jonathan May

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

Problem Statement

- Classifier evaluation on imbalanced datasets, two schools of thought:
  (1) Micro: treat each instance equally (2) Macro: Treat each class equally
- Best practice: if classes are imbalanced, and rare classes are important, use Macro
  - In NLP: *type* is *class*, *token* is *instance*; and rare types are important
- Q: What if we apply the best practices of classifier evaluation to MT evaluation?





Brown Corpus (Eng) ~1.2M tokens, ~48k types

Classifier Eval Metrics to MT

- Multi-class performance = average of individual class performances
  - Performance of a class, e.g., F-measure: $F_{\beta;c}$

1. **Macro average:** unweighted

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

i.e., equal importance to each *type*

2. **Micro average**: weighted e.g., frequency

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

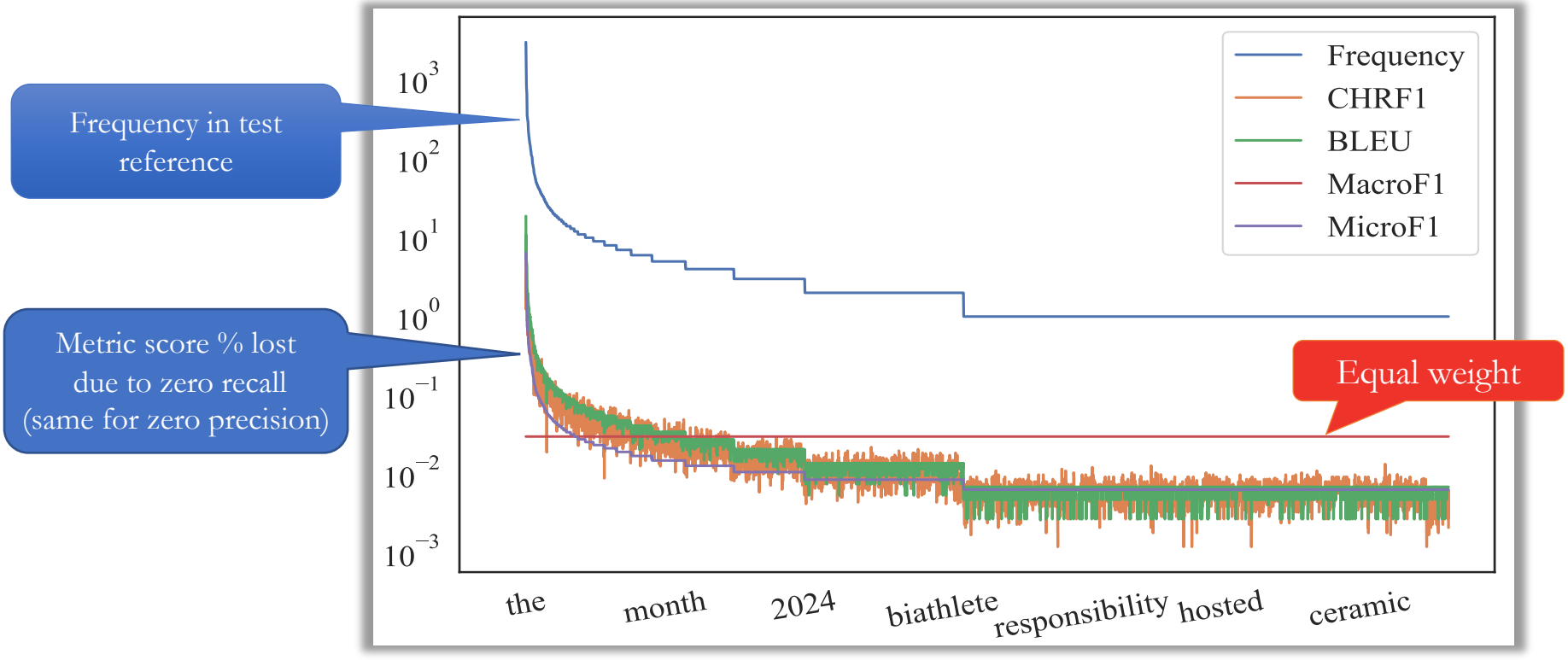
i.e., equal importance to each *token*

NOTE: Micro-F1 ≅ Accuracy

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

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

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

Justification for MacroF1as an MT Eval Metric

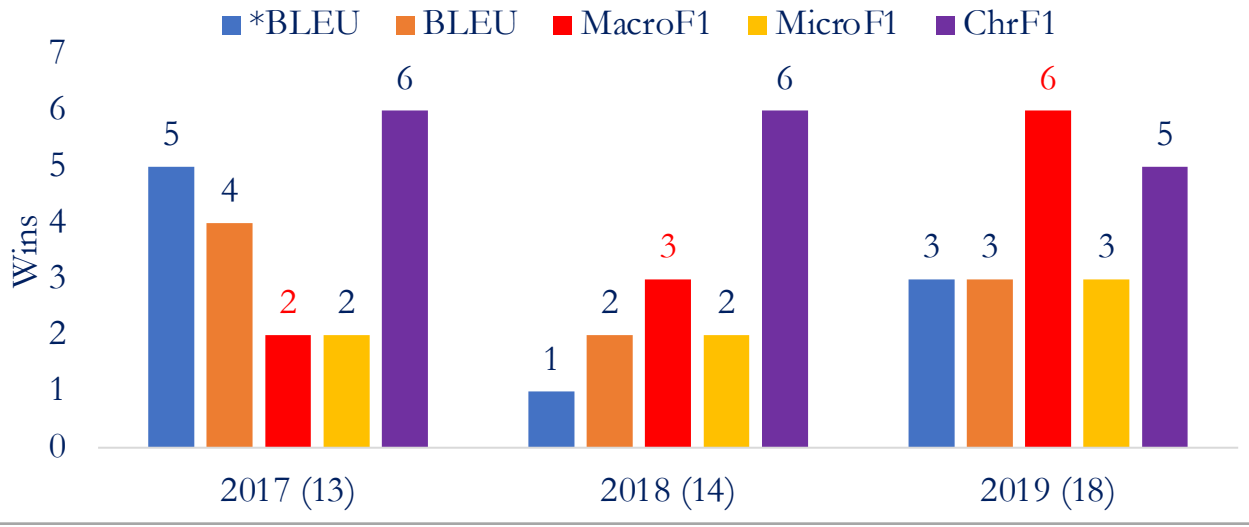
- Compare MacroF1 with
  - BLEU, ChrF1, and MicroF1
  - BLEURT – a model-based metric based on BERT
    - *Model based models have undesirable biases*
- Experiments:
  1. Direct human assessment:
     MT vs Human judgement score correlations on WMT Metrics tasks 2017-19
  2. Downstream CLIR Task metrics:
     IR task with documents and queries in different languages
     - MT vs IR score correlations: CLSSTS 2020: LT-EN, PS-EN, BG-EN
- Findings:
  - Direct evaluation: MacroF1 has upward trend over the years, and wins the highest number of times in 2019
  - Downstream task correlation on CLIR task: MacroF1 is consistently better on all three language pairs we have tested on

Reference:	You must be a doctor.	
Hypothesis:	_____ must be a doctor.	
	He	-0.735
	Joe	-0.975
	Sue	-1.043
	She	-1.100

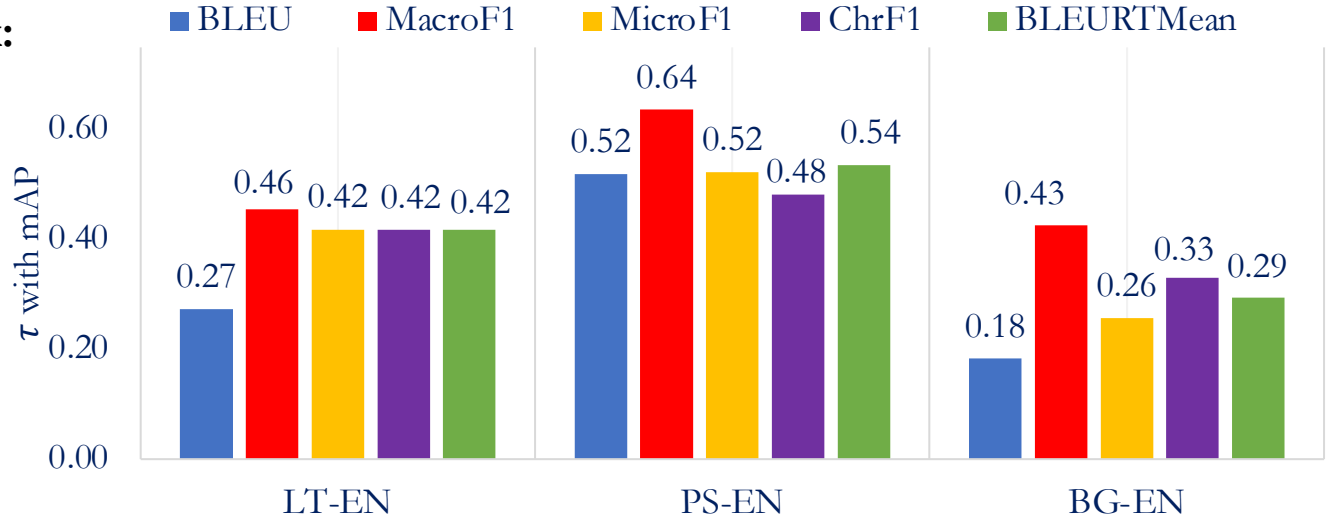
Model based metrics (e.g., BLEURT) have undesirable biases

Justification

WMT Metrics Task:
System level MT score vs human judgement correlation



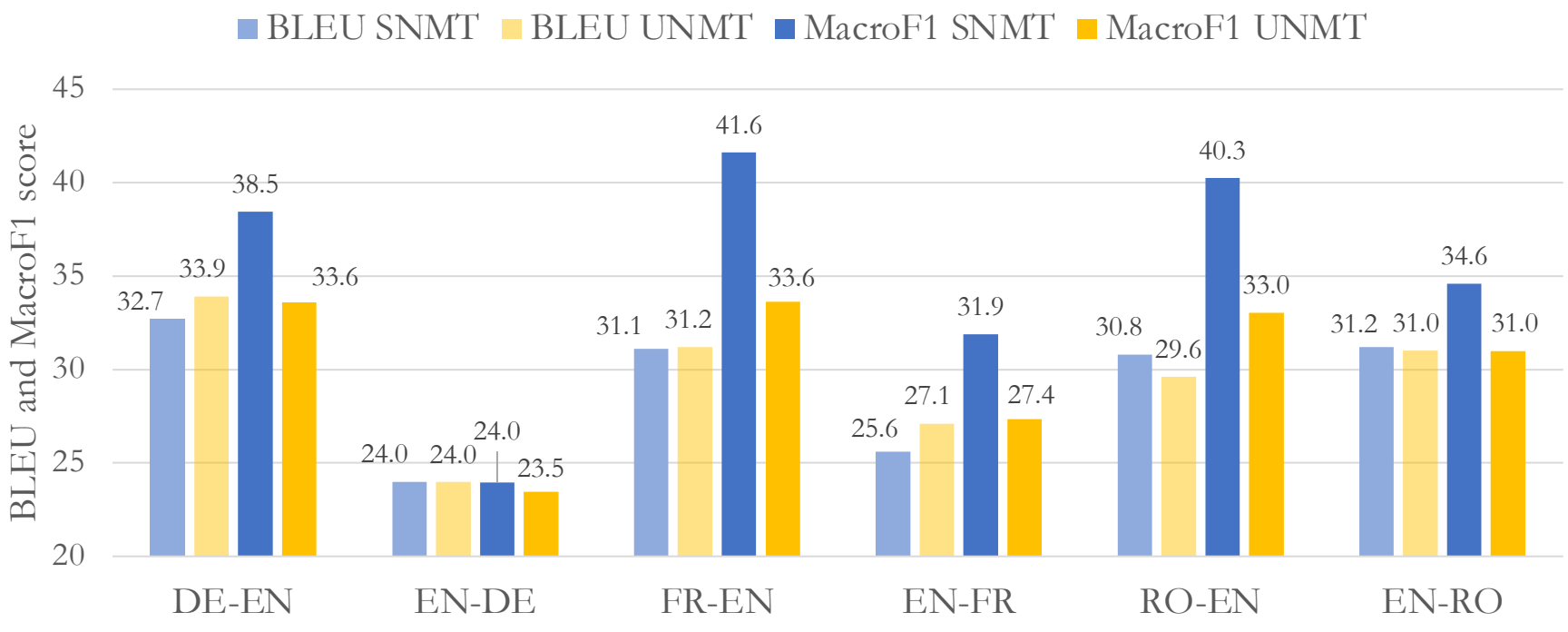
Downstream CLIR Task:
MT vs IR metrics correlation



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

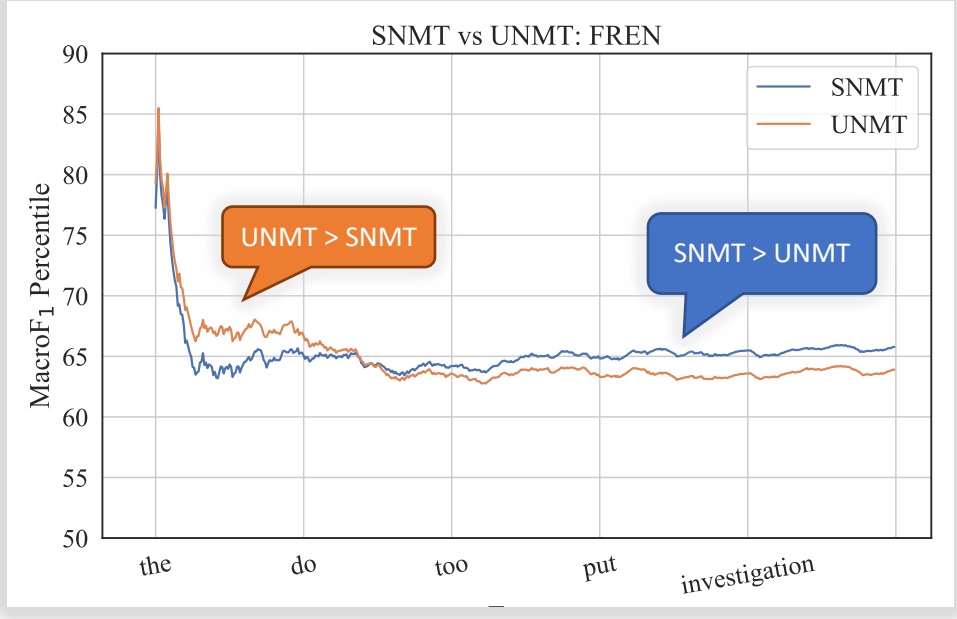
MacroF1 Difference Between SNMT and UNMT

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 MT have lower F1 score for content types than stopwords ⇨ long-tail curse

- Similar trend on EN-FR, EN-DE, DE-EN, EN-RO, RO-EN

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



Only the top 500 types are visualized

Part-II: Conclusion

Related Work

- A lot of MT evaluation metrics… but missed type imbalance
  - Exception: NIST BLEU; but it makes much stronger assumptions about type importance
- Recent trend: model-based evaluation metrics. Undesirable biases and uninterpretable scores

Summary

- Rare types are important too
- Macro F1 for MT evaluation
  - Competitive on direct human assessment (when all systems are fluent)
  - Outperforms others on downstream CLIR task
- Disagreement between BLEU (a micro metric) and MacroF1 can be clearly seen on supervised vs unsupervised NMT

✅ I – Rare words in training

"Finding the optimal vocabulary size for NMT" [EMNLP 202

✅ II – Rare words in evaluation

"Macr

⇨ III –

"Impr view]

IV –

"Many-to-English MT tools, data, and pretrained models" [ACL 2021 Demos]

V – Discussion, Future Directions

III – Rare linguistic styles

Improving Multilingual MT Robustness via Data Augmentation

(Under Review)

Thamme Gowda, Mozhdeh Gheini, and Jonathan May

Problem Statement

- Sometimes,
  - Multilingual speakers switch between languages
  - Part of text is already in target language
- Code switching or language alternation
- Are the current multilingual NMT models robust? No
- Q: How to check robustness? And how to 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.

Code-switching example: Kannada+Hindi

Robustness Checks

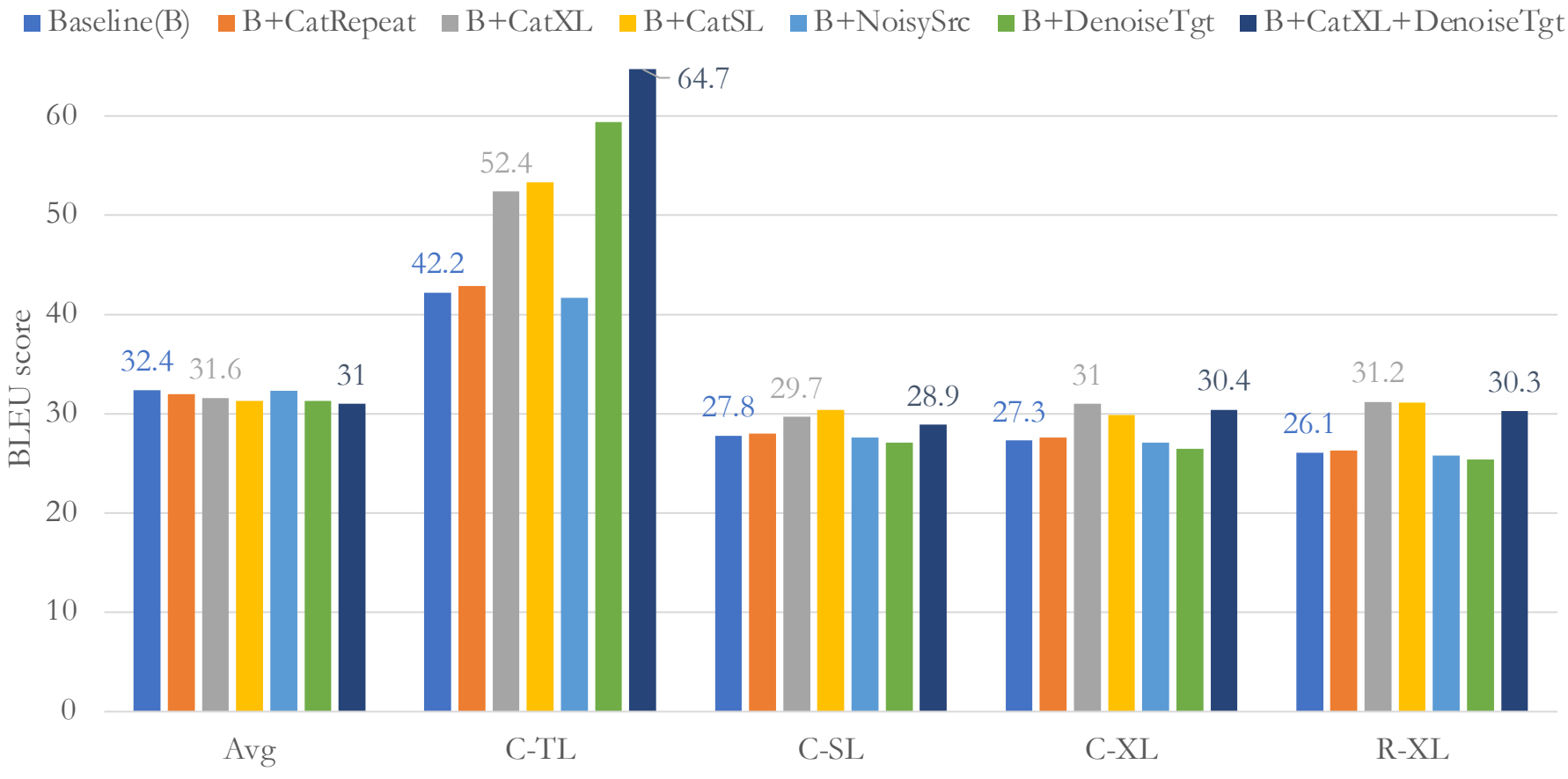
- Behavior testing for NLP in general (Ribeiro et al 2020[1])
  - Simple modifications of test sets: negation, synonym, NER replacement, etc.
  - Tasks: sentiment analysis, duplicate question detection, span detection
  - Idea of behavior testing for MT is interesting; lets apply it in multilingual translation settings
- Create more tests by concatenating test sentences
  - **C-SL**: consecutive **same** language ➔ missed sentence segmentation
  - **C-TL**: consecutive **target** language ➔ partial translation
  - **C-XL**: consecutive **cross**-language ➔ Code switching (inter-sentence)
  - **R-XL**: **random cross**-language ➔ Code switching + random topic switching

[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
- Robustness checking evaluation sets are created as per previous slide
- Training data augmentations:
  1. CatRepeat: concatenate
  2. CatSL: concatenate random sentence in same language
  3. CatXL: concatenate random sentence across languages
  4. NoisySource: noise(source) → target
  5. DenoiseTarget: noise(target) → target
- Noise: 10% of random word drop, random replacements, and random word order shuffle

Results: BLEU



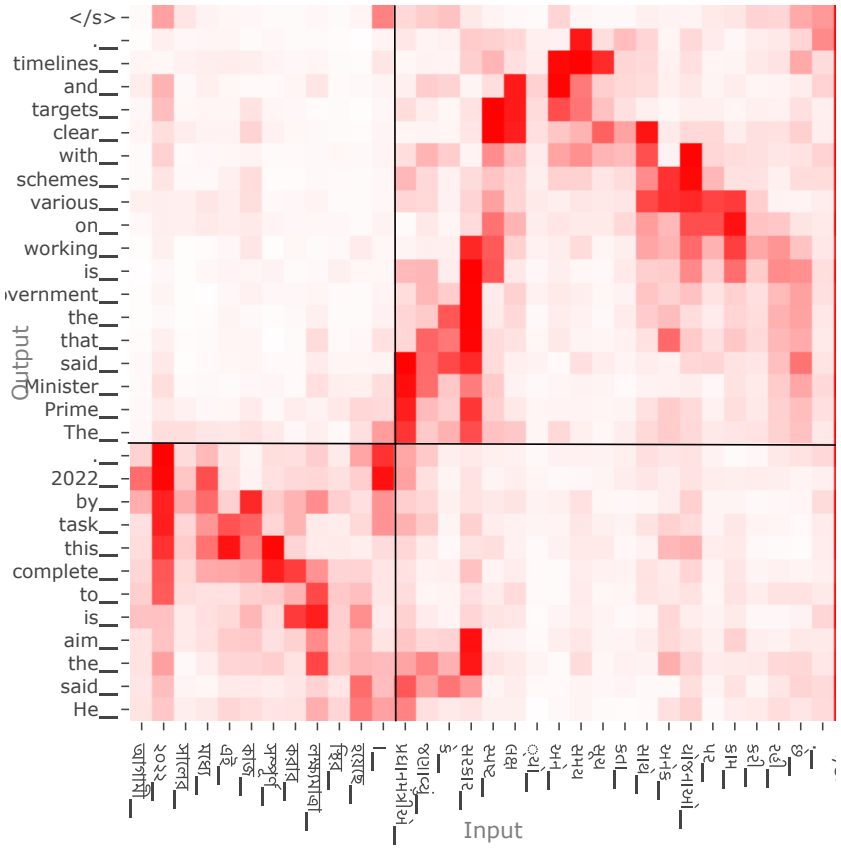
Improvements are not visible on the original (Avg) set, but proposed checklist sets showcase it.

Qualitative Examples

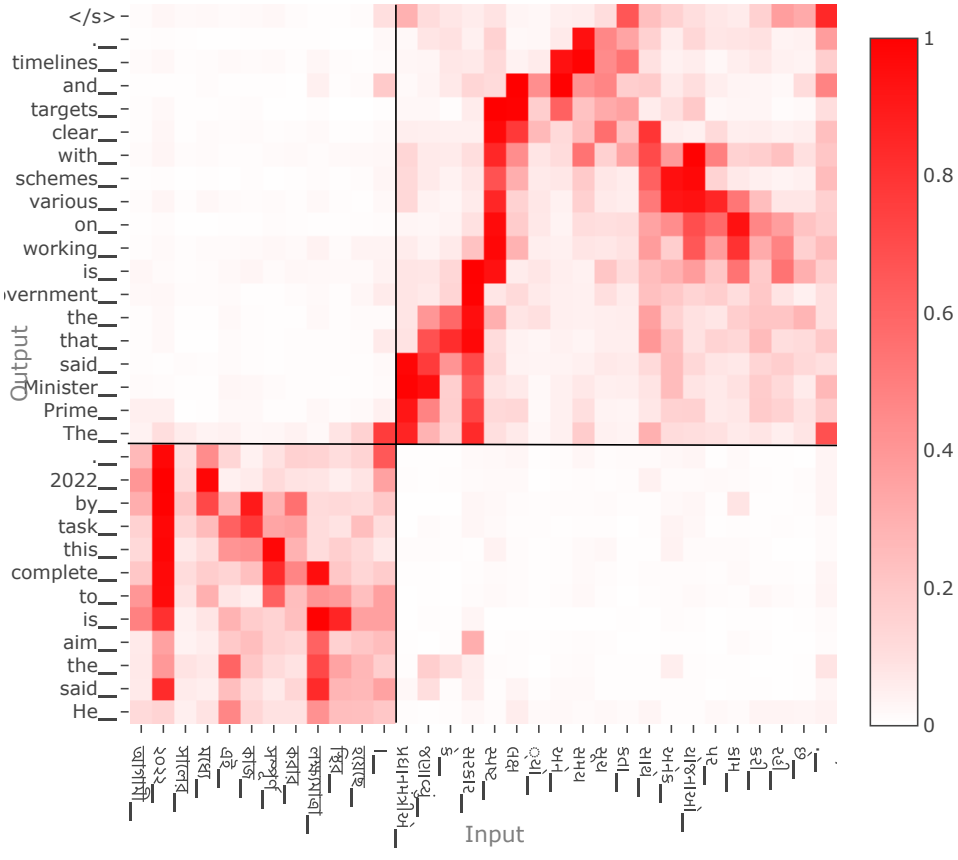
Source	আগামী ২০২২ সালের মধ্যে এই কাজ সম্পূর্ণ করার লক্ষ্যমাত্রা স্থির হয়েছে। પ્રધાનમંત્રીએ જણાવ્યું કે સરકાર સ્પષ્ટ લક્ષ્યો અને સમયસૂચકતા સાથે અનેક યોજનાઓ પર કામ કરી રહી છે.
Reference	He said the aim is to complete this task by 2022. The Prime Minister said that the Government is working on various 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 for a C-XL check (Source: Bengali + Gujarati)

Attention Visualization



Baseline



Model with augmented data (CatXL+Denoise)

Models trained with augmented data learns robust cross-attention mechanism

Part-III Conclusion

Related Works

- Back translation:
  Costly in massively multilingual setup
- Most other robustness works are concerned about bilingual MT, and noisy data, where as our work is about code-switching cases

Summary

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

✅ I – Rare words in training

"Finding the optimal vocabulary size for NMT" [EMNLP 2

✅ II – Rare words in evaluation

"Macro

✅ III – I

"Impro iew]

⇨IV – I

"Many-t emos]

V – Discussion, Future Directions

IV – Rare languages

Many-to-English MT Tools, Data, and Pretrained Models (ACL 2021 Demos)

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

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

Problem Statement

- 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 focused on the top 100 languages
- There are no MT models for thousands of languages
- Q: Can we build MT models for rare languages ?

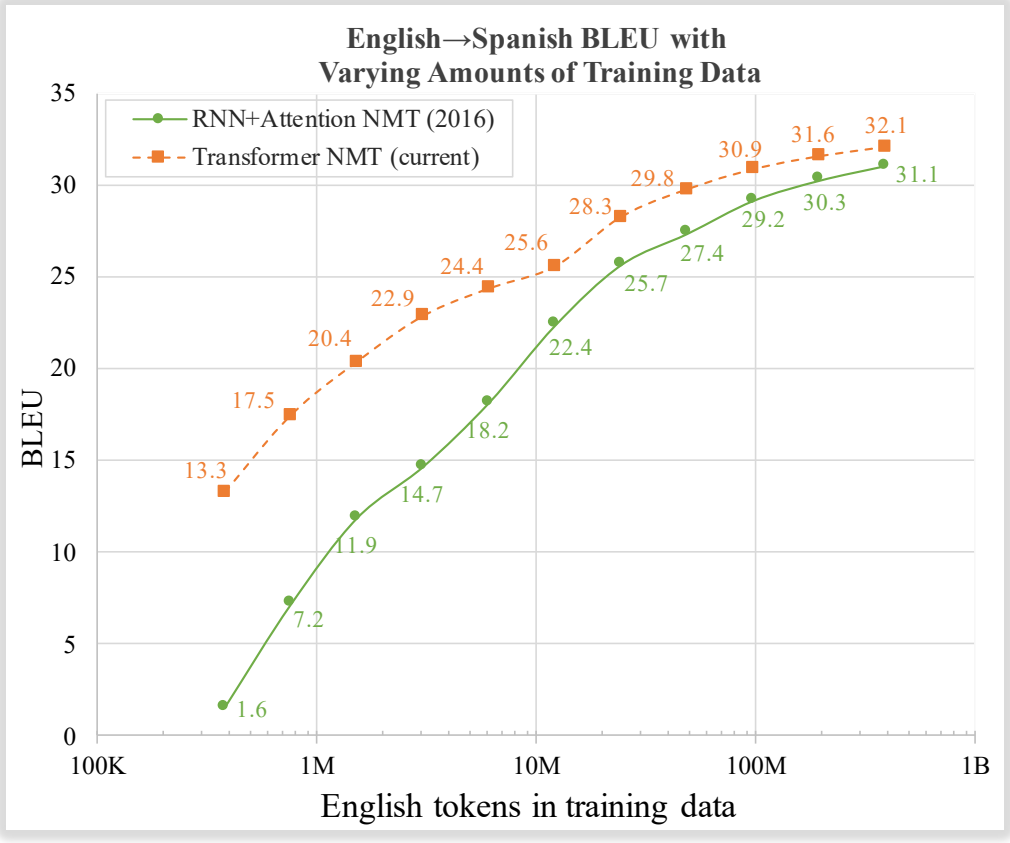
Population Range	Number of Languages			Number of speakers		
	Count	Percent	Cum%	Total	Percent	Cum%
100M - 1B	8	0.1	0.1%	2.8B	40.46	40.46%
10M - 100M	86	1.2	1.3%	2.8B	40.00	80.47%
1M - 10M	313	4.4	5.7%	1B	14.09	94.56%
100k - 1M	977	13.7	19.5%	310M	4.44	99.00%
10k - 100k	1,812	25.5	44.9%	62M	0.89	99.89%
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/

Requirements for Supporting Rare Languages

1. Scalable, label-efficient NMT modeling
   - Thanks, Vaswani et al 2017 !
   - Scaled to 100 of languages
   - Efficient in low resource ⇨
2. Faster hardware
   - Thanks, Nvidia!
3. Datasets
   - *(missing)*

- Some quantity of parallel datasets exist for ~600 languages
- We need tools to put all these together, train models, and make them accessible



In low resource settings, Transformers are more efficient than prior models (simulated on the left side). This opens up

Tools for Scalable NMT

1. MTData: parallel dataset catalog and downloader
   - Publicly listed datasets: OPUS, Statmt.org, Paracrawl, ...
   - 131k datasets and counting, for hundreds of languages [1]
   - https://github.com/thammegowda/mtdata/
2. NLCodec: Vocabulary manager; and database layer
   - PySpark backend for large datasets
   - NLDb: Efficient storage and retrieval layer; parallelizable
   - https://isi-nlp.github.io/nlcodec/
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
   - https://isi-nlp.github.io/rtg/

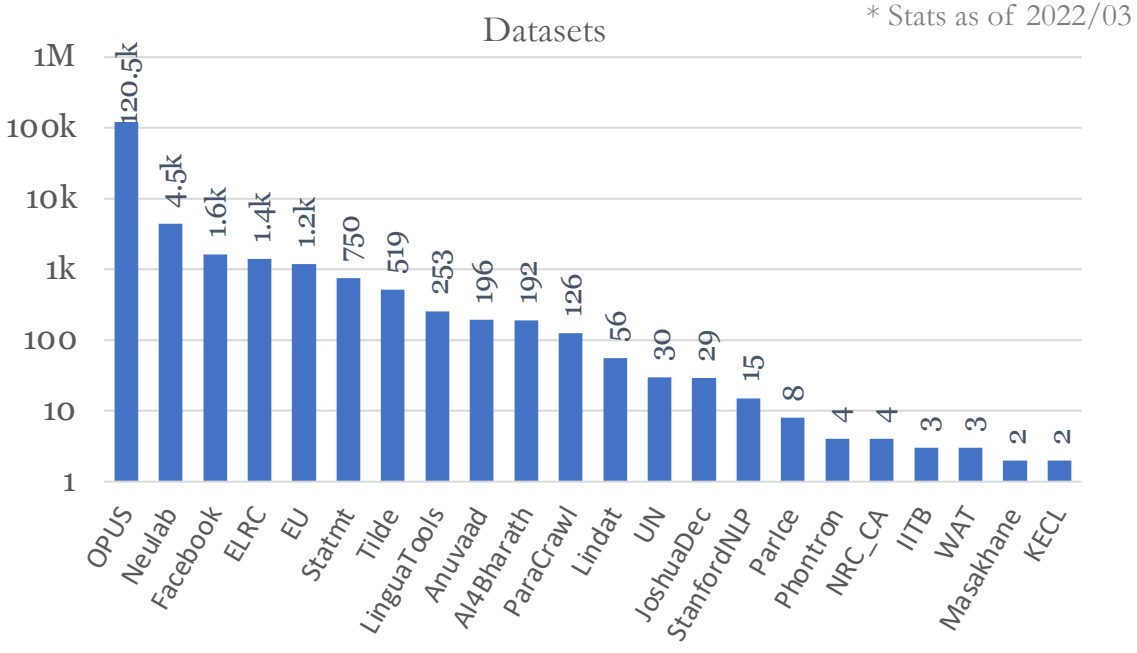
- `pip install mtdata nlcodec rtg`

More NLP tools under my PyPI a/c https://pypi.org/user/Thamme.Gowda/

[1] As of 2022/03; earlier we had 91k more from JW300, but they are taken down due to licensing issues

Datasets for 500+ Languages

- MTData has an index of parallel datasets
- Where are the datasets? ⇨
- Datasets comes in different formats. Standardization of language names, IDs etc
  - ISO 639-3 ~~ISO 639-1~~
  - BCP 47
- BixTex citation entries (whenever available)
- **Recipe**: a set of datasets nominated for train/dev/test
  - Intended to improve reproducibility



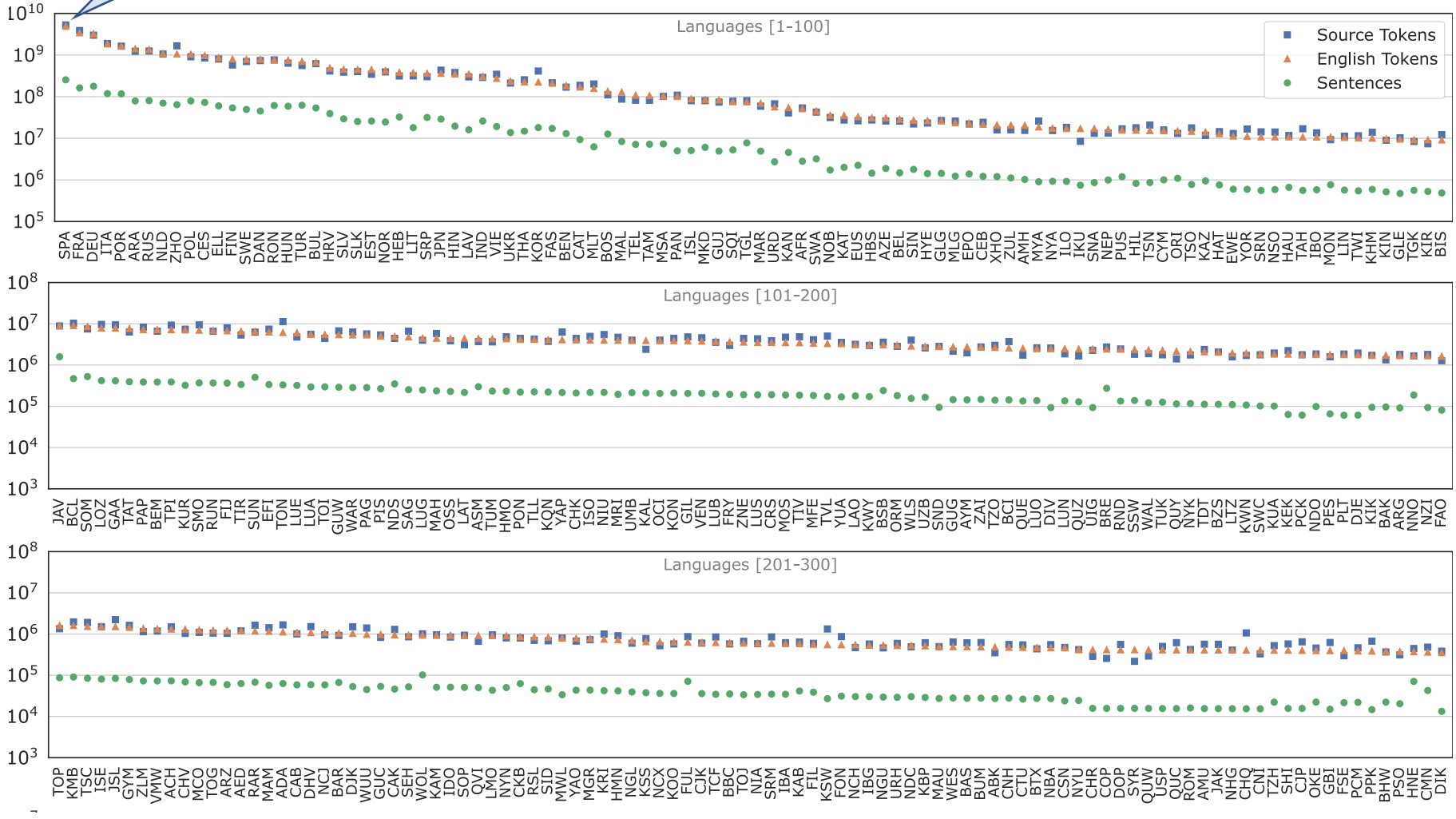
New: WMT 2022 recipes
https://www.statmt.org/wmt22/mtdata/

Thanks to Kenneth Heafield (U Edinburgh) for their contributions

Many → English Translation

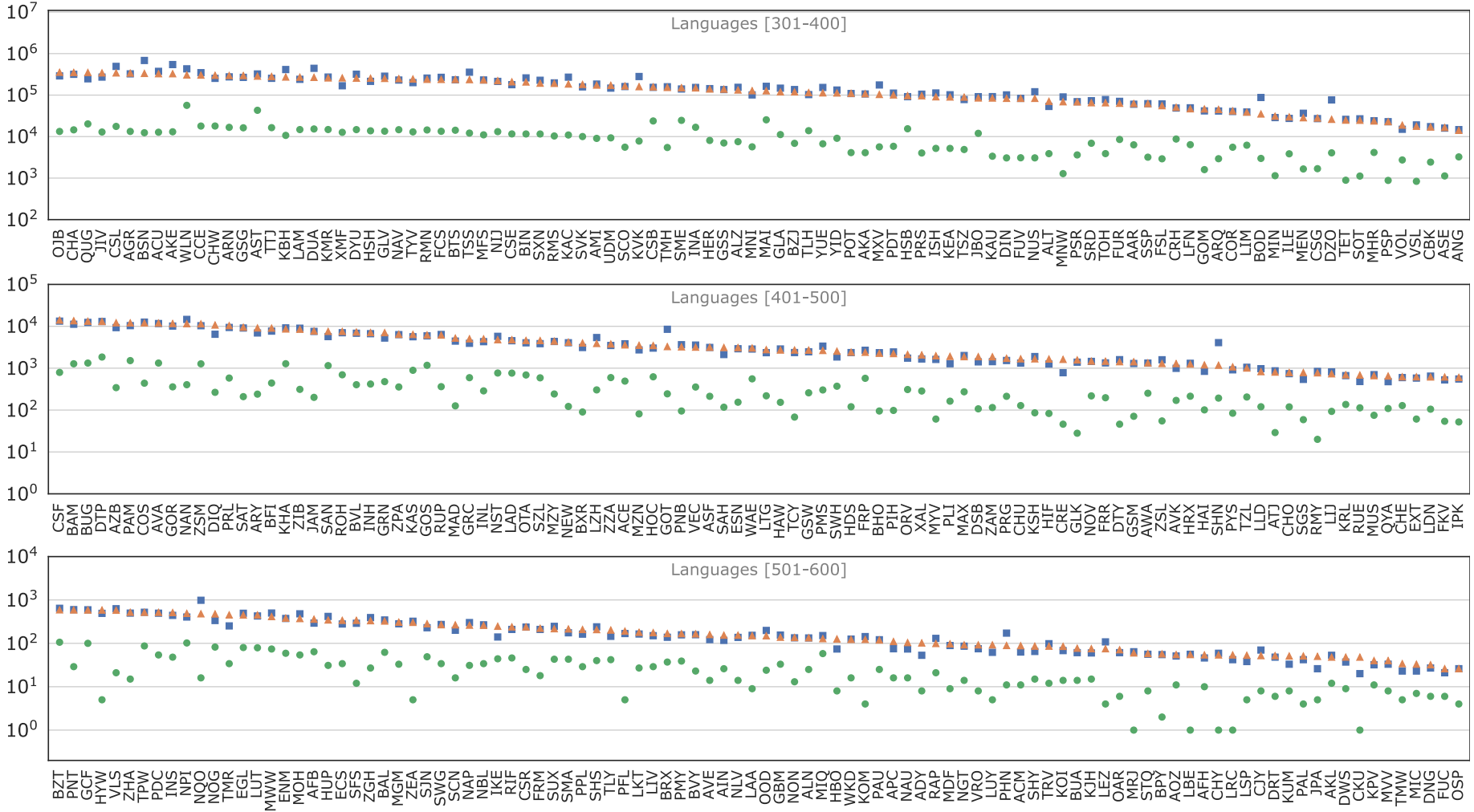
- Using `mtdata` download datasets, followed by deduplication, cleaning, excluding sentences from known test sets, …
- V1: 500-to-English [Fall 2020 – Spring 2021]
  - Dataset: 500+ languages: ~474M sentence pairs; 9B toks on each side
  - Model: Transformer: 768d, 9 encoder and 6 decoder layers
    - 512k source and 64k target embeddings
    - 539M params; ~73% in source embedding, 21% in target embedding
- V2: 600-to-English [Spring 2022]
  - Dataset: 600+ languages: 2.3B sentence pairs; ~37B toks on each side
  - Model: same as V1; except, 1024 dims
- V2.1 : V2 model finetuned on code switching augmentation corpus
  - At most 200k random sentences are selected per language
  - Augmentation methods: (1) Random sent concatenation, (2) Denoise target

V2 Dataset Statistics: [1-300]



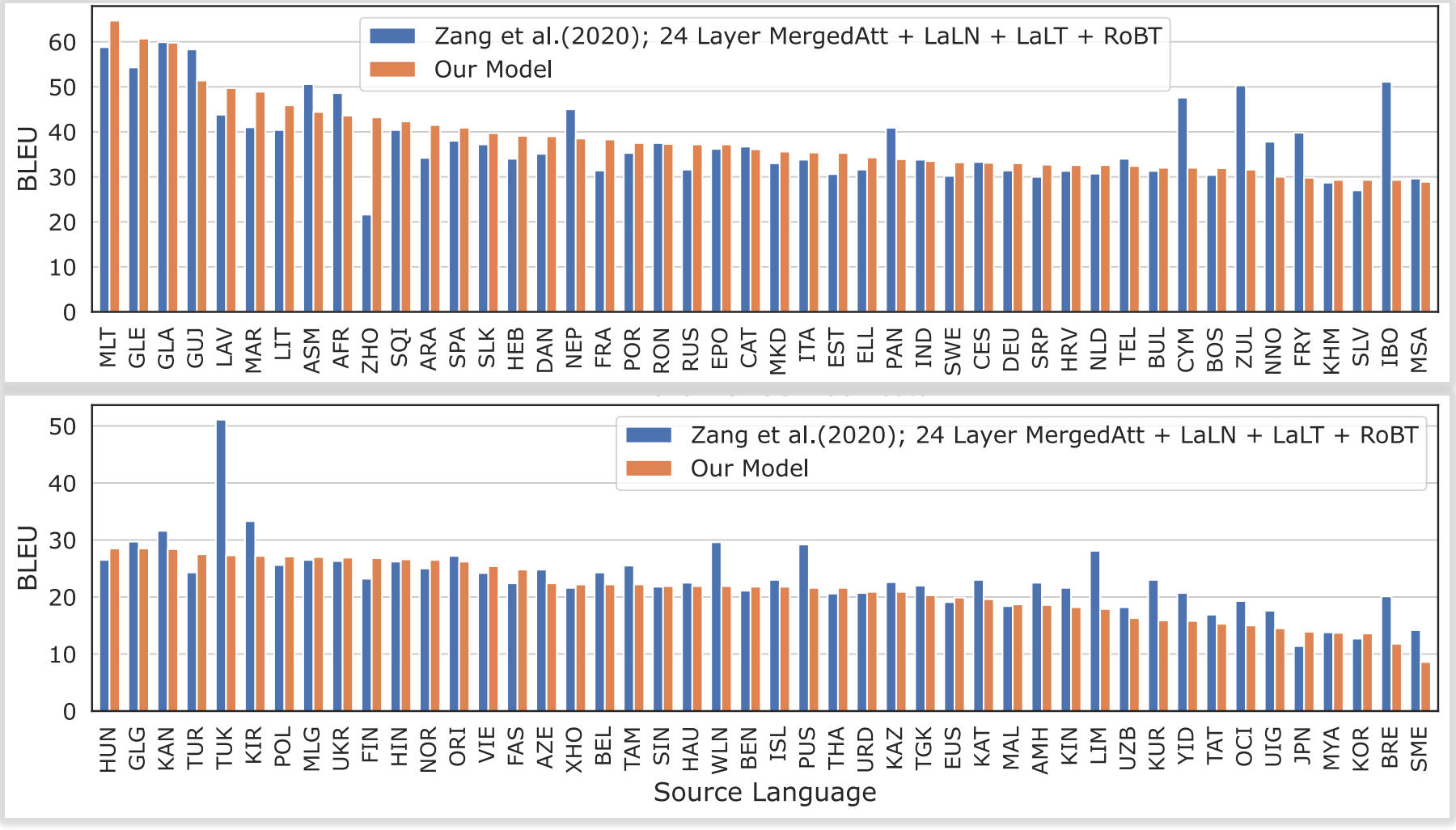
http://rtg.isi.edu/many-eng/data-v2.html

V2 Dataset Statistics: [301-600]

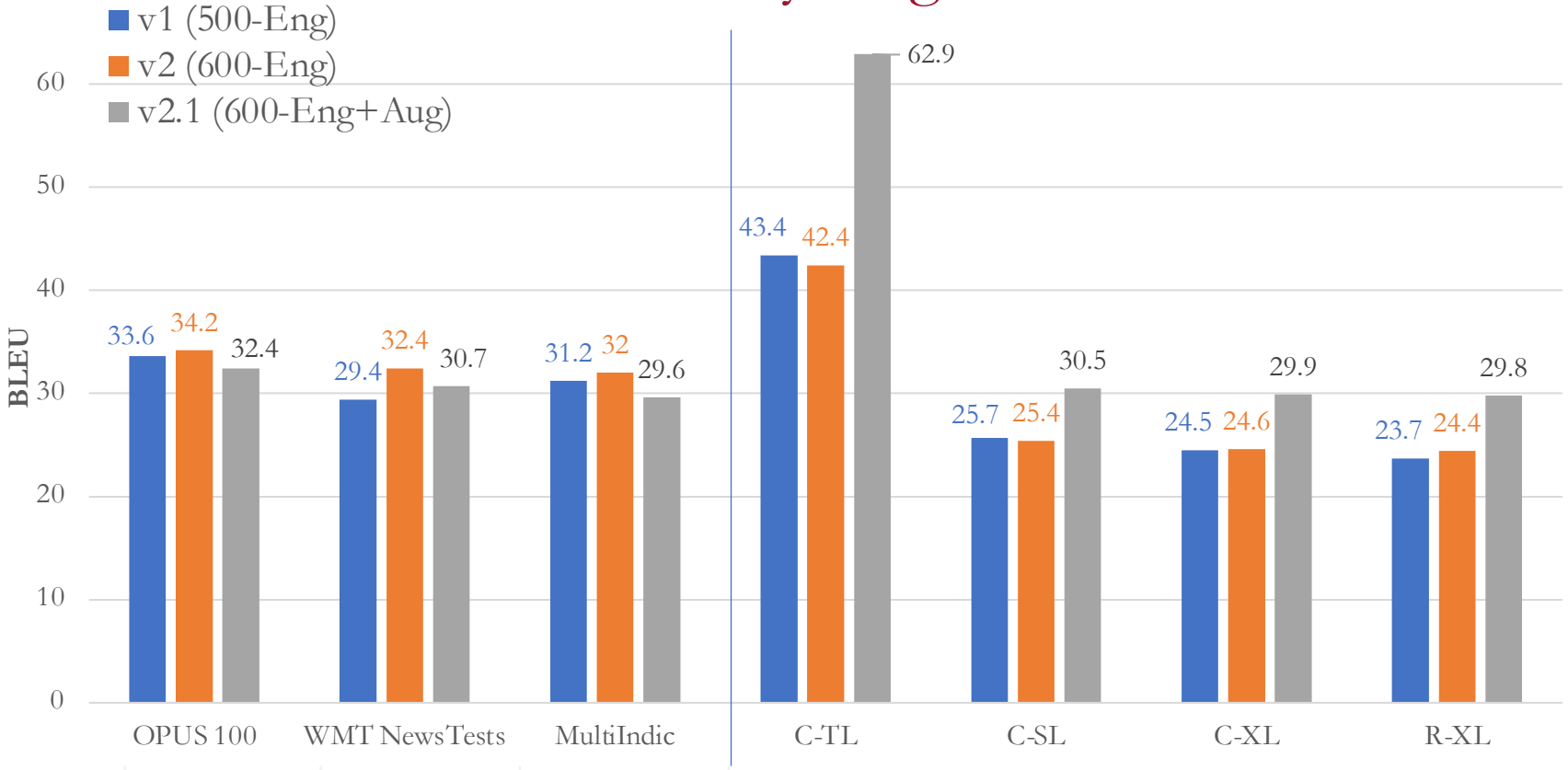


http://rtg.isi.edu/many-eng/data-v2.html

V1 Model's BLEU on OPUS-100 Test Set



Results: All many-English models

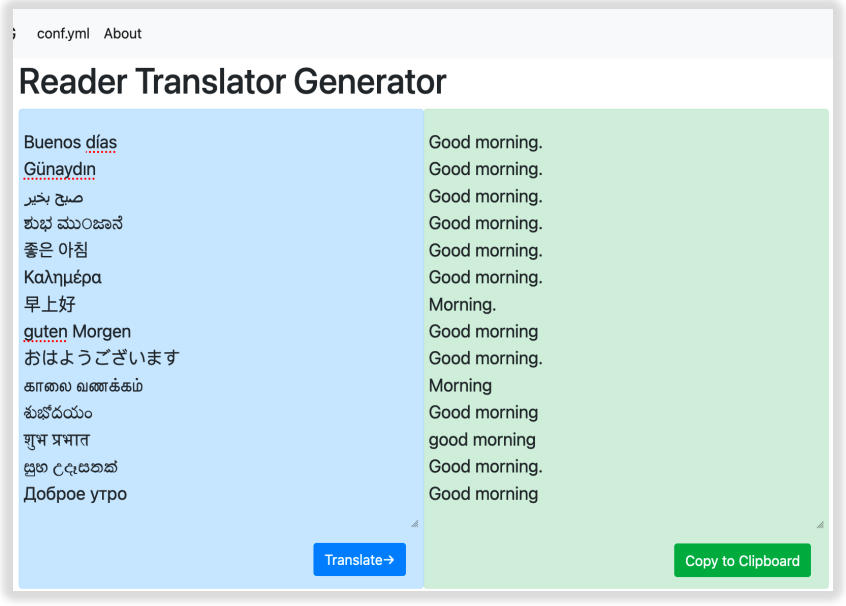


v2 model is better on single sentence translation; v2.1 model is better on robustness checks

An End-to-end MT System

- Demo: http://rtg.isi.edu/many-eng/
  - Models: rtg.isi.edu/many-eng/models/
- Docker images

```
$ IMG=tgowda/rtg-model:500toEng-v1
$ IMG=tgowda/rtg-model:600toEng-v2.0
$ docker run -p 6060:6060 $IMG
# For GPU backend, --gpus '"device=0'"
```



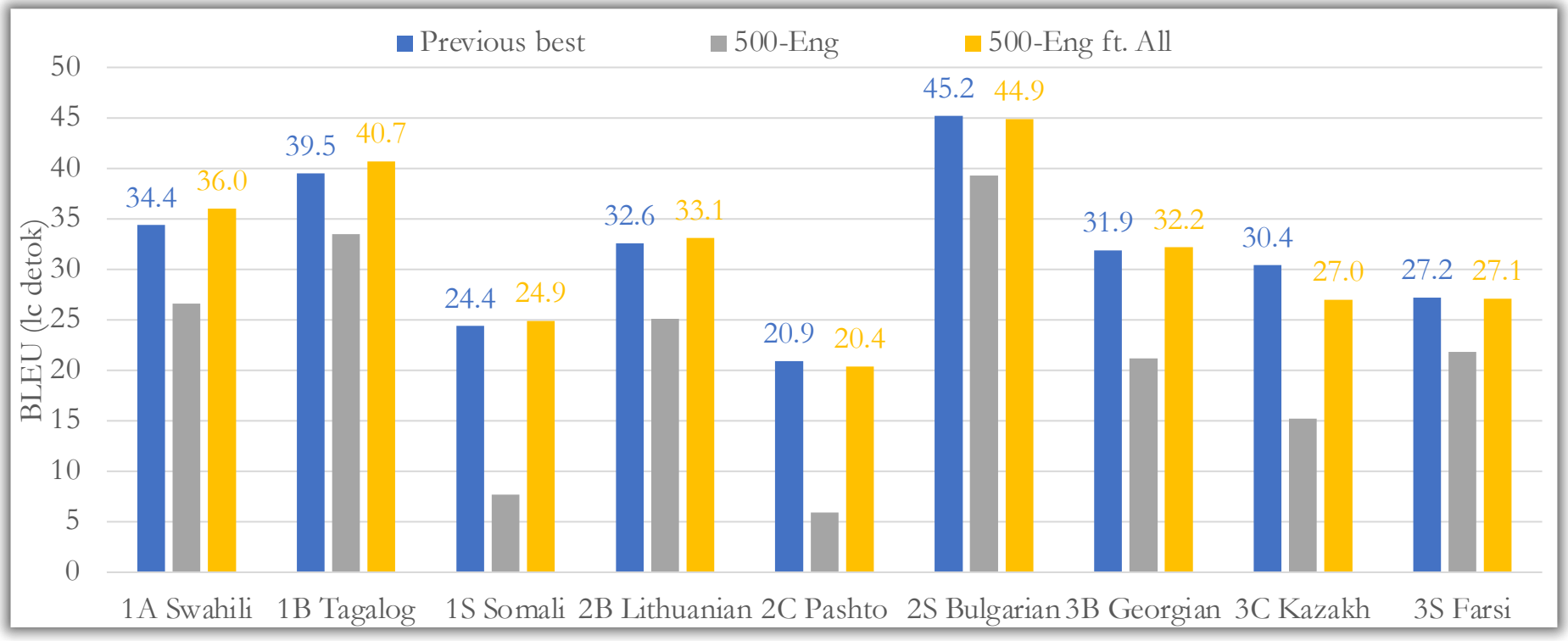
RTG Web Interface: REST API behind the scenes (via AJAX)

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/

Finetuning: BLEU
on IARPA MATERIAL Datasets (Analysis)



"Previous best" : the best bilingual models used in IARPA MATERIAL evaluations; separate model for each language

Part-IV 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

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

✅ 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 –

"Impr view]

✅ IV – Rare languages (600+)

"Many-to-English MT tools, data, and pretrained models" [ACL 2021 Demos]

⇨V – Discussion, Future Directions

Part V
Discussions and Future Directions

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 ~600 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]

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

Works Outside MT [After Joining 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 .

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. The SEXI workshop on the 9th ACM conference on Web Search and Data Mining (WSDM 2016). ACM

Bias in ML models

Rare phenomena learning

Future Directions

- Other seq-to-seq tasks with natural languages
  - Speech recognition, summarization, captioning, dialogue generation… rare words are important too. But these models are evaluated using micro metrics.
  - Dialogue generation may yield more diverse responses if long tail is emphasized
- Other ways to mitigate data imbalance severity
  - Masked language models (e.g., BERT): masking strategies that consider data imbalance into account
  - Label smoothing (LS) is effective in practice. Since LS alters class distribution, it maybe possible to of this method to improve class balance
  - Weighted loss functions: focal loss, dice loss etc., for sequence-to-sequence
  - Maximum Entropy Principle:
    Training with the uniform/balanced distribution: "*Prepare for the worst*",
    Evaluation on Zipfian/skewed dataset: *"Hope for the best"*
- Other sequential data: genome sequences, stock market events, space and earth weather forecasting, rare events in sensor readings, etc.,

CUTE LAB NAME

Q & A

USC Viterbi School of Engineering

The Inevitable Problem of Rare Phenomena Learning in Machine Translation

Dissertation Defense

by

Thamme Gowda

Committee

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

Thanks

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

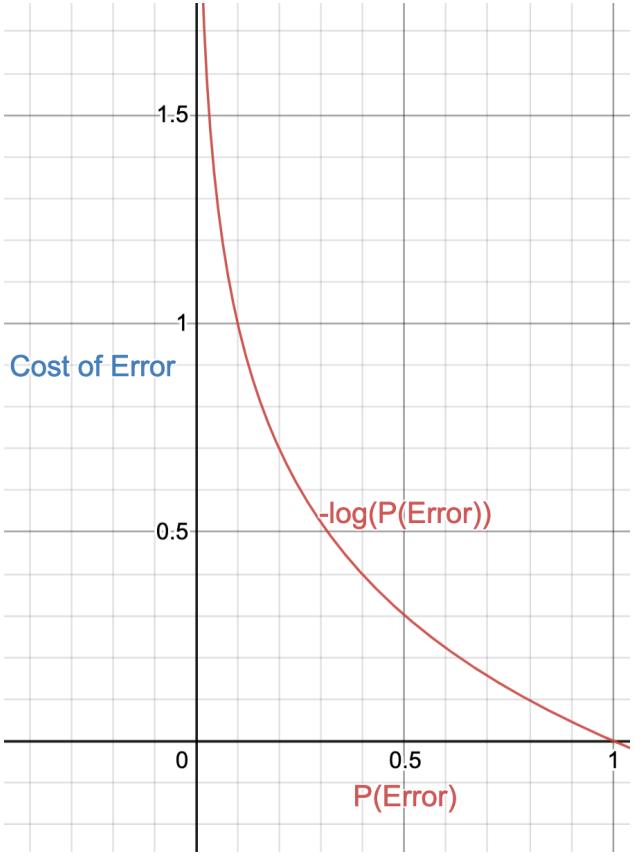
- USC Employee Benefits (Staff Tuition Assistance)
- DARPA LORELEI
- IARPA MATERIAL
- DARPA LwLL

Computing Resources

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

PS. The Other Curse

"The more of the local stuff we get right, the more users will come to trust the software, and hence the more noticeable long-range dependencies will become, and the more upset people will get if they are deceived by a wrong analysis" – Mark Steedman, 2008

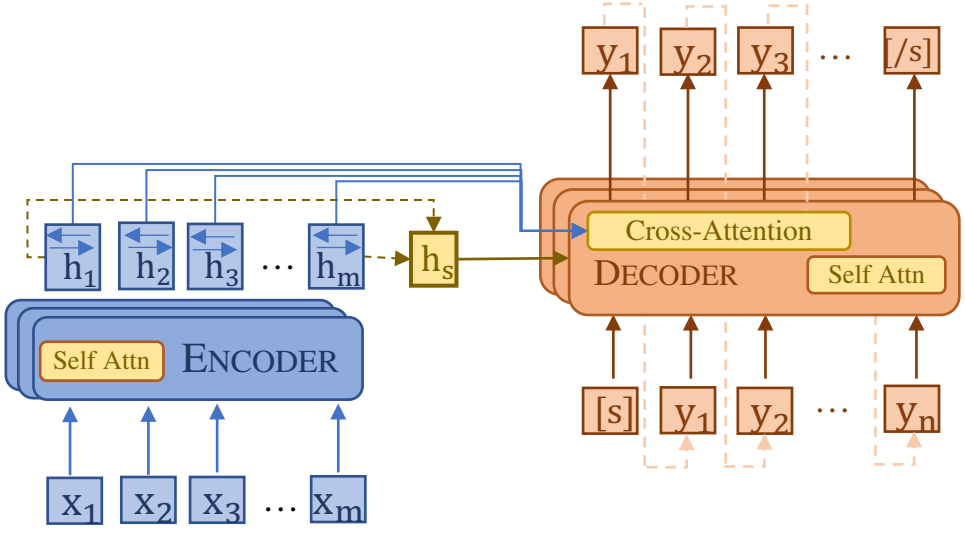


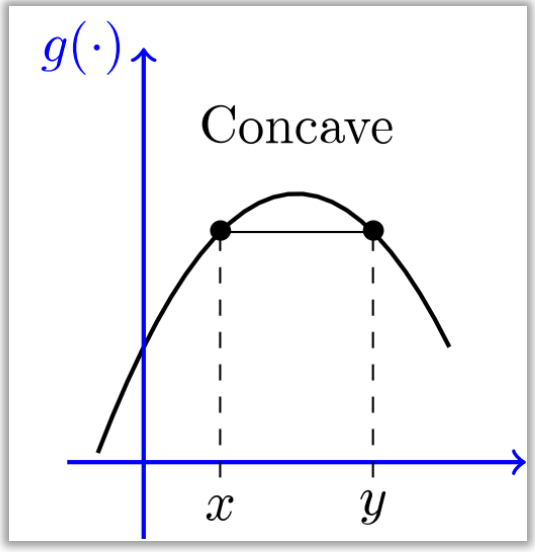
As erroneous predictions become rare, the cost of each error (e.g., how much it will upset users) grows exponentially high

END

Neural Machine Translation

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





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

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/

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/

Part-III

Robustness to Language Switching

Improving Multilingual Machine Translation Robustness to Code-Switching via Data Augmentation

Thamme Gowda, Mozhdeh Gheini, and Jonathan May. (Under review)

Part-IV

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/

Implications: A 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]

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`

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:

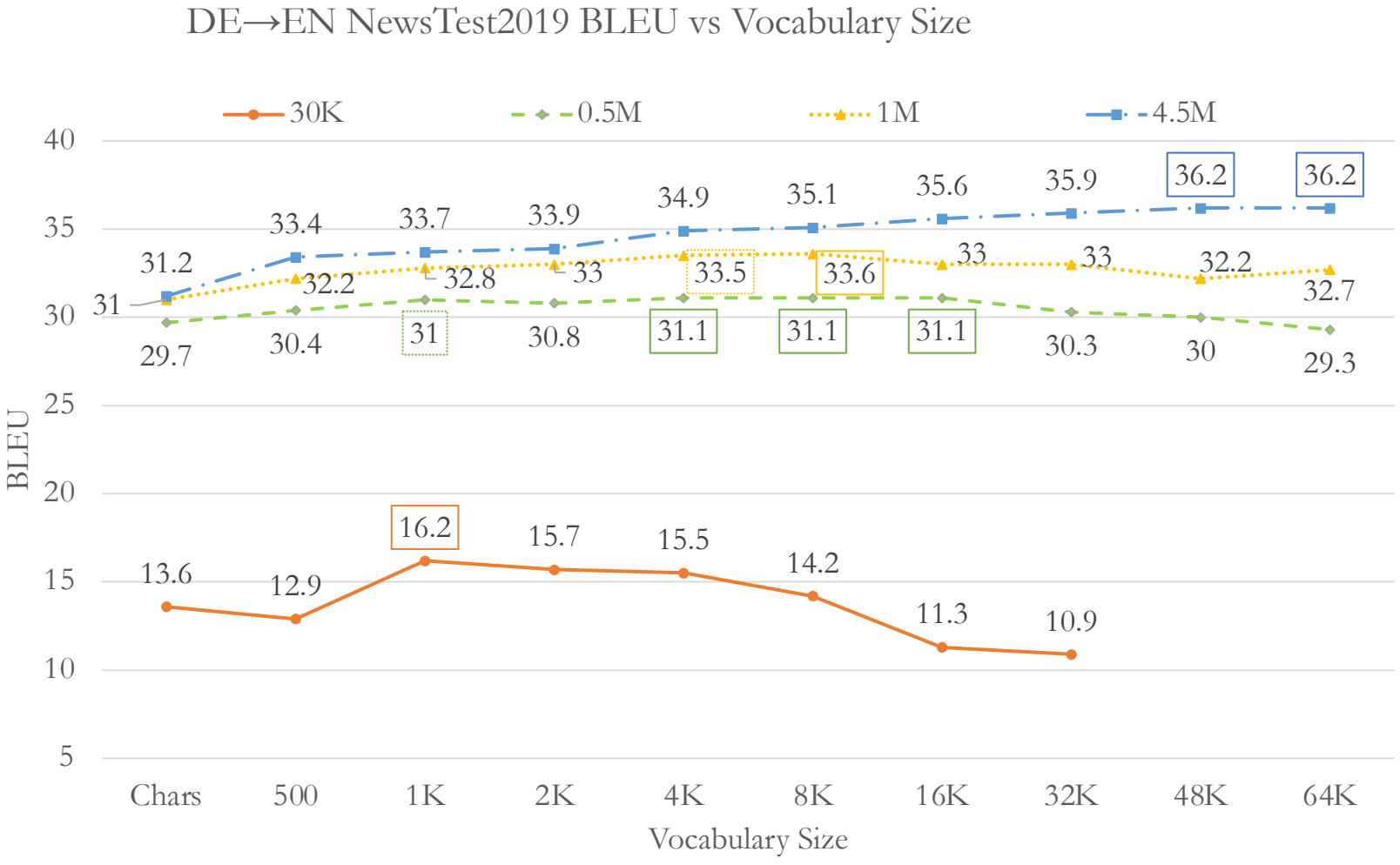
- Spring 22 : 03/18
- Summer 22 : 07/01

Dataset Standardization

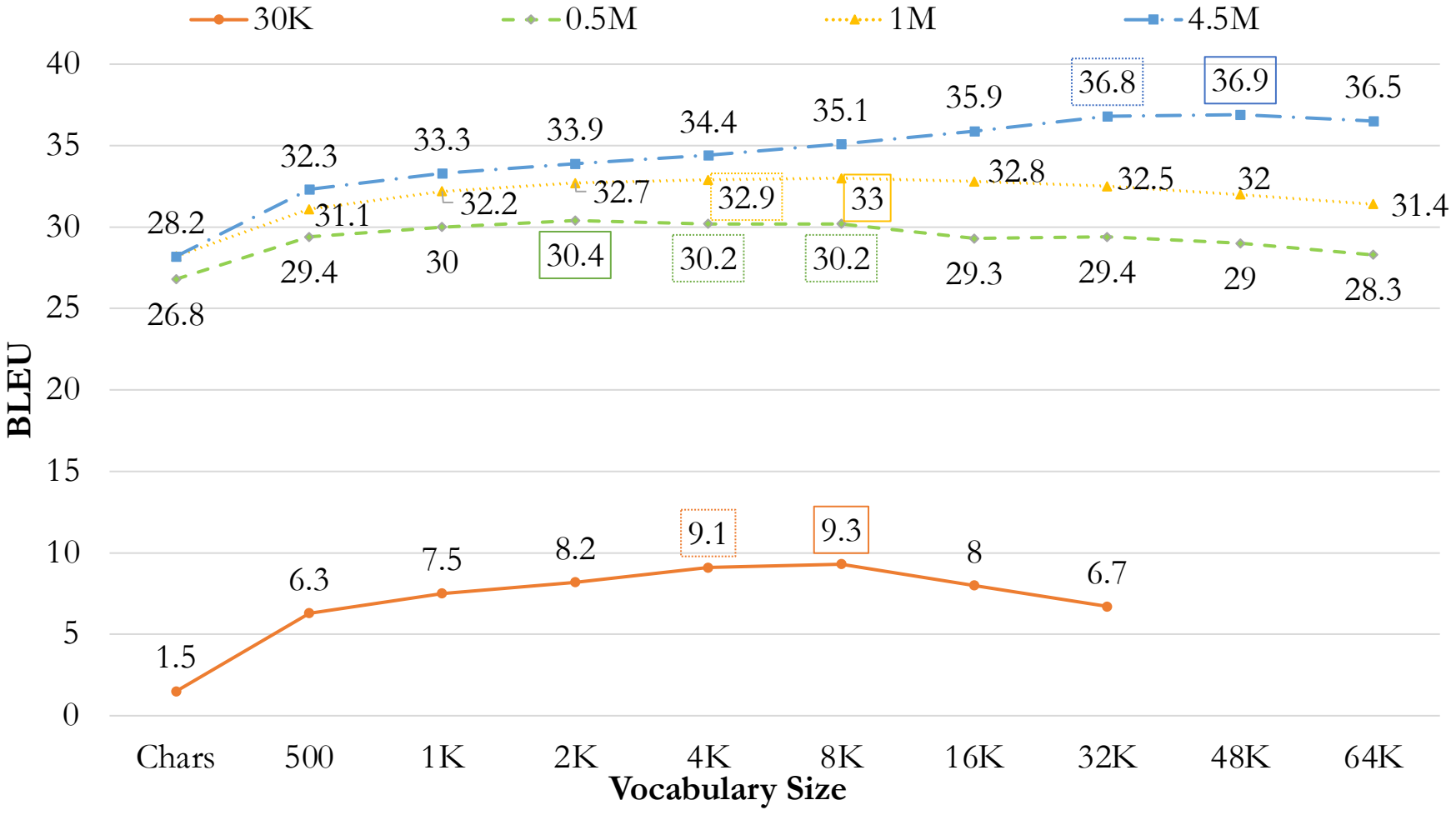
- Standard way to reference/identify languages
  - ISO 639-1: two letter code e.g., en, de, hi, zh, .. 184 codes
  - ISO 639-3: three letter code e.g., eng, deu, hin, zho, … 7,867 codes
- BCP 47: Language [Script] [Region] aka. IETF language tag [1]
  - Language: ISO 639-1 (for backward compatibility) and ISO 639-3 (for low resource)
  - Script: ISO 15924; optional, suppressed if default e.g, Latn, Cyrl …
  - Region: ISO 3166-1; optional, e.g, US, GB, IN, CA, …
- Mtdata uses BCP47 like ID, but ISO 639-3 for all languages
  - `eng_Latn_US` ➔ `eng_US`
  - `mon_Cyrl_MN`
  - `kan_Knda` ➔ `kan` ; `kan_Latn` ➔ `kan_Latn`
- Dataset ID: `<Group>-<name>-<version>-<lang1>-<lang2>`
  - `Paracrawl-paracrawl-8-deu-eng`
  - `OPUS-paracrawl-7.1-deu-eng`

[1] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

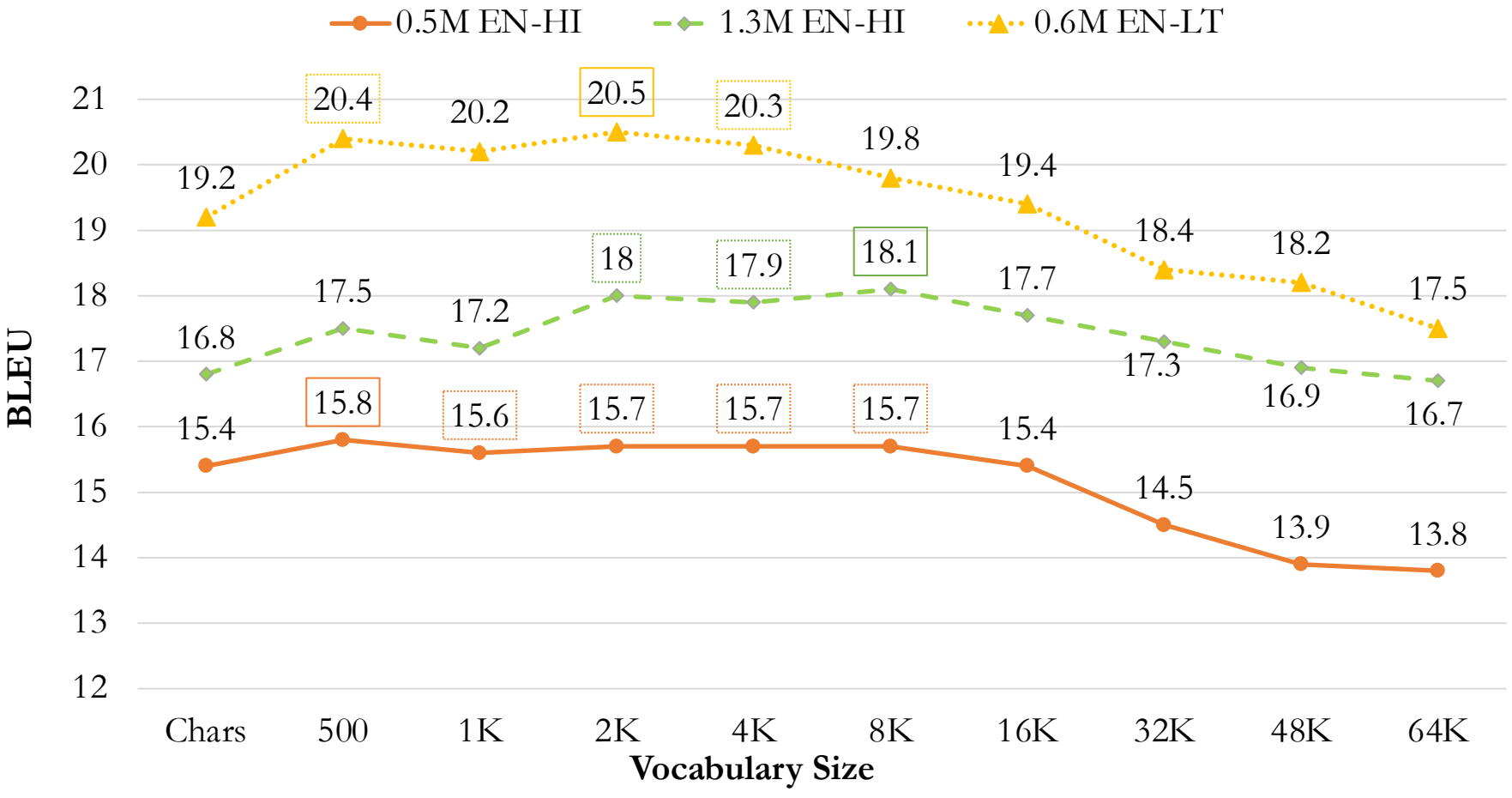
Thanks to Kenneth Heafield for guidance on this topic


DE→EN NewsTest2019 BLEU vs Vocabulary Size
30K
0.5M
1M
4.5M
BLEU
40
35
30
25
20
15
10
5
31.2
33.4
33.7
33.9
34.9
35.1
35.6
35.9
36.2
36.2
31
32.2
32.8
33
33.5
33.6
33
33
32.2
32.7
29.7
30.4
31
30.8
31.1
31.1
31.1
30.3
30
29.3
13.6
12.9
16.2
15.7
15.5
14.2
11.3
10.9
Chars
500
1K
2K
4K
8K
16K
32K
48K
64K
Vocabulary Size

EN→DE NewsTest2019 BLEU vs Vocabulary Size



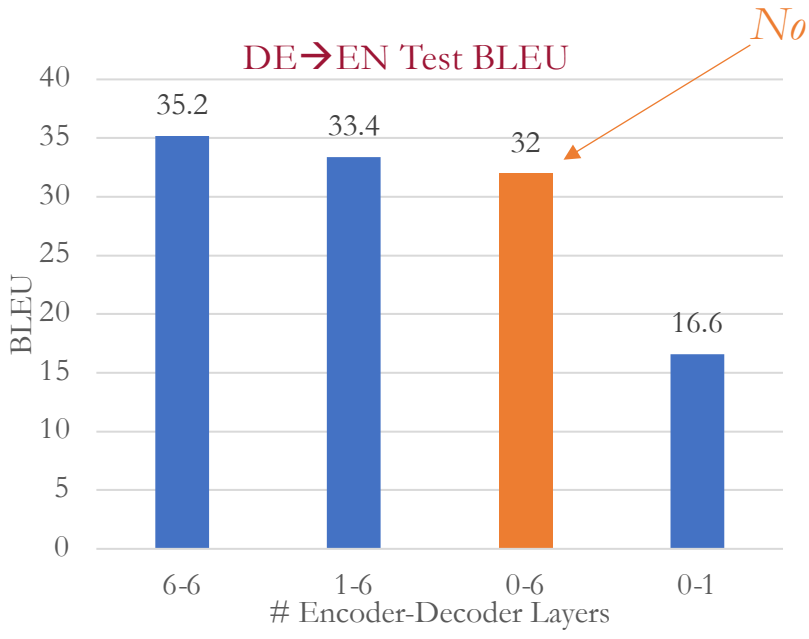
BLEU vs Vocabulary Size
EN→HI IITB Test and EN→LT NewsTest2019



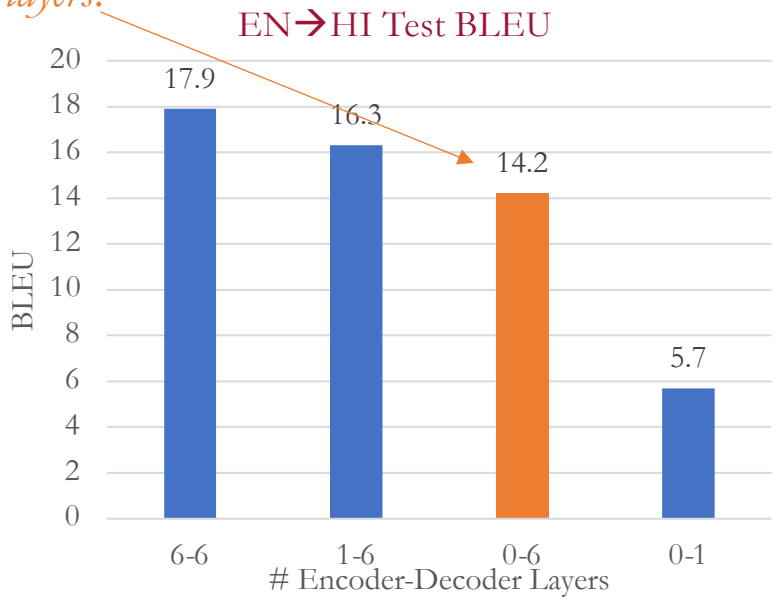
Transformer Ablation

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)



No encoder layers!



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

https://github.com/thammegowda/015-nmt-ablation

Imbalanced Classification Learning Methods

- Sampling methods:

  (1) Up-sampling /over-sampling (2) Down-sampling / under-sampling,
  (3) Synthetic Minority Over-sampling (SMOTE)

  ➔ *Not straight forward in MT; word types are imbalanced, but sentences are to be sampled*

- Weighted loss functions:
  - Weighted cross entropy
  - Focal-loss
  - Effective number of samples

  ➔ *Adaptive learning methods, e.g., ADAM, and label smoothing achieve a similar effect*

- Byte Pair Encoding (BPE)

  Balancing classes via splitting and merging in a sequence

  ➔ *Tuning the vocabulary size improves class balance* ✅

Classifier (C)

- BPE merge ops modify class distribution
- Goal: balanced class distribution during training
  (Maximum entropy principle)
- Balance = Uniform distribution
- Imbalance = Divergence from balance

1. Using simplified Earth Mover Distance

$$D = \frac{1}{2}\sum_{i=1}^{K}\left|p_i - \frac{1}{K}\right|$$

$0 \leq D \leq 1$ for a distribution of K classes

2. Sufficient training examples, $F_{95\%}$
   - $F_{95\%}$ defined as least frequency in the 95th % of most frequent classes
   - Least frequent 5% classes excluded as noise

AutoRegressor (R)

- BPE merge ops vary sequence length
- Autoregressor's predictions are based on its past prediction(s)
  - As well as Classifier's predictions
- Shorter sequences are easy, longer sequences are hard
  - Both C and R are approximations, having non-zero probability of errors; errors compound on long seqs
- Mean sequence length $\mu = \frac{1}{N} \sum_{i}^{N} |y^{(i)}|$
  where $y^{(i)}$ is a target sequence in a parallel corpus of N sequences

Frequency-based Bias on Class Performance

- $\mathcal{R}$: Ranking of test set classes based on *training set frequency*
  - Classes are BPE sub-words

- Pearson Correlation Coefficient

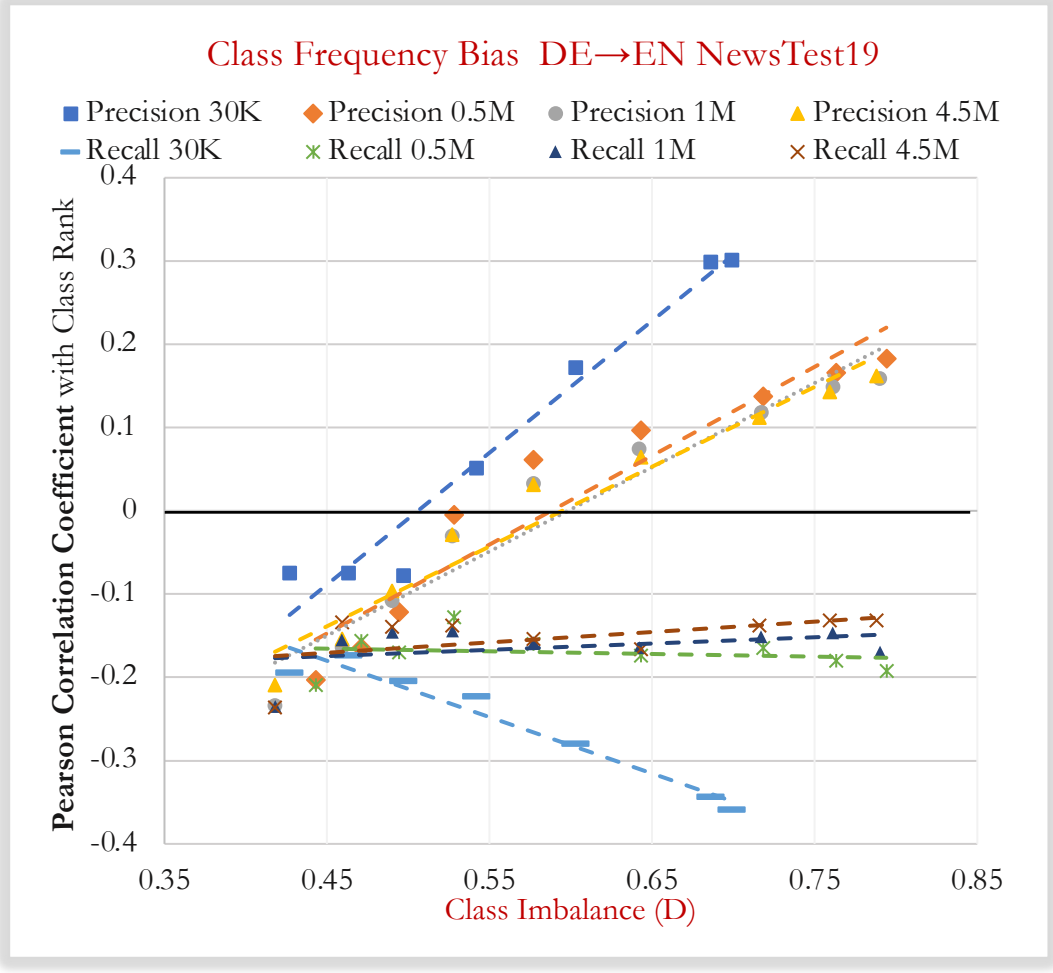
1. Rank vs Precision ($\rho_{\mathcal{R},P}$)
   - $\rho_{\mathcal{R},P}$ is positive at high D

   ⇨ Frequent classes have relatively poor precision

2. Rank vs Recall ($\rho_{\mathcal{R},R}$)
   - $\rho_{\mathcal{R},R}$ is negative at high D

   ⇨ Rare classes have poor recall

- Takeaway: precision is improved with lower D, but recall of rare classes is still problematic



Review of MT Metrics

- [Papineni et al. 2002[1]] $\text{BLEU} = [\prod_{n=1}^{4} P_n)]^{\frac{1}{4}} . BP$

  P_n is n-gram precision of tokens. BP is brevity penalty

  $\text{BP} = \min\left(1, \exp\left(1 - \frac{r}{c}\right)\right)$

- [Popović, 2015[2]] $\text{ChrF}_\beta = (1 + \beta^2)\frac{\text{ChrP} \times \text{ChrR}}{\beta^2 \times \text{ChrP} + \text{ChrR}}$

  Character n-grams for up to 6-grams

- [Sellam et al. 2020[3]] BLEURT

  BERT, a transformer language model finetuned to predict human judgements scores on WMT

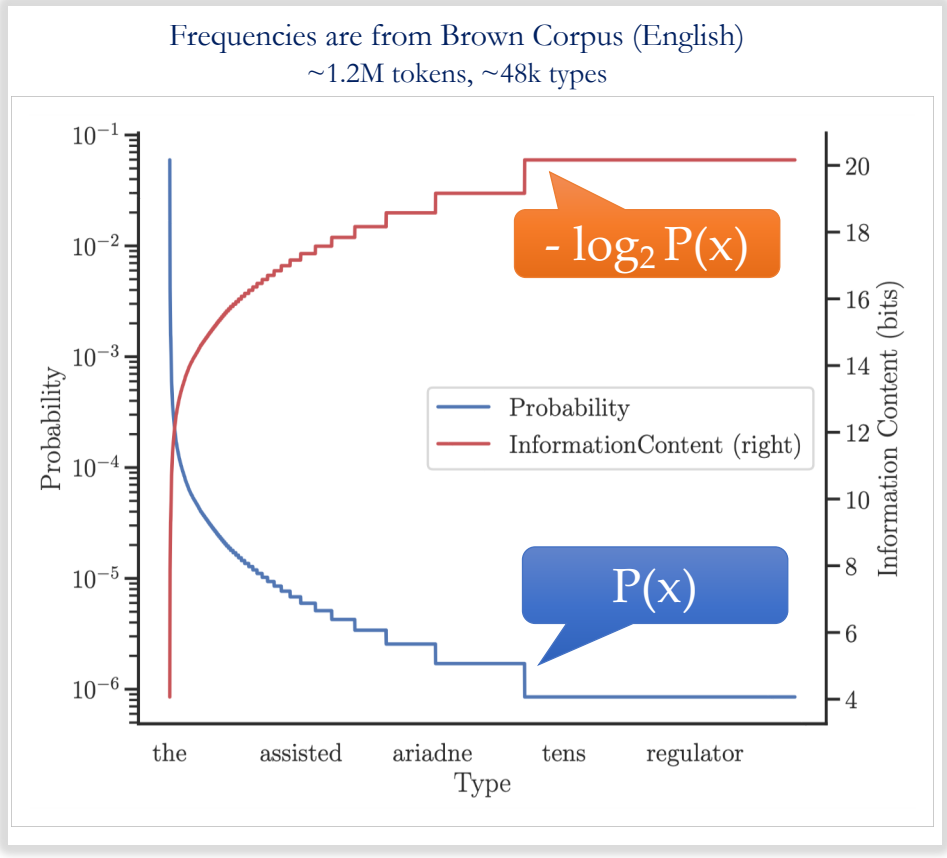
Reference:	You must be a doctor.	
Hypothesis:	_____ must be a doctor.	
	He	-0.735
	Joe	-0.975
	Sue	-1.043
	She	-1.100

Model based metrics (e.g., BLUERT) have unwanted biases

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

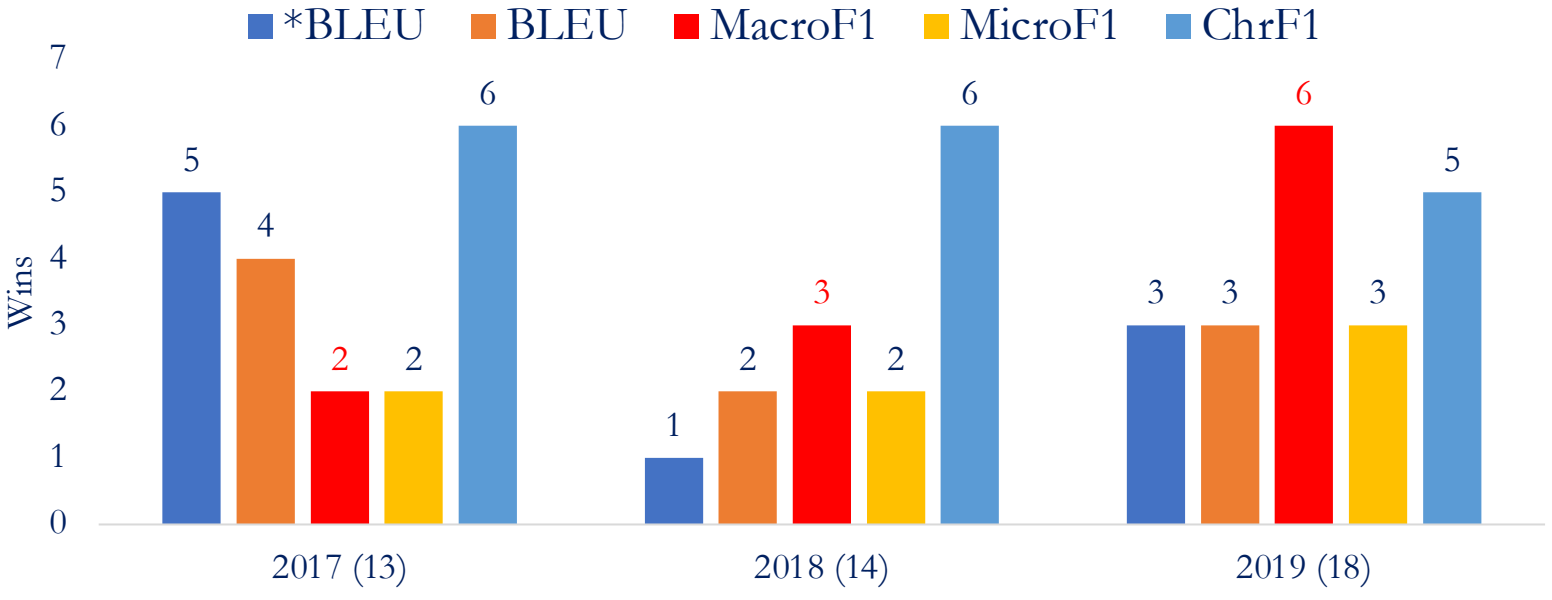
Motivation / Problem

- Natural language datasets have imbalanced word types
  - Rare types have more information content
- Classification evaluation metrics
  - Micro: treat each instance equally
  - Macro: treat each class equally
  - Use macro metrics on imbalanced sets
- Micro metrics on imbalanced sets give a false sense of model performance
  - E.g., cancer detection with 1:99 imbalance; majority label classifier gets 99% overall accuracy, but that's not a useful metric
  - NLP is similar



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



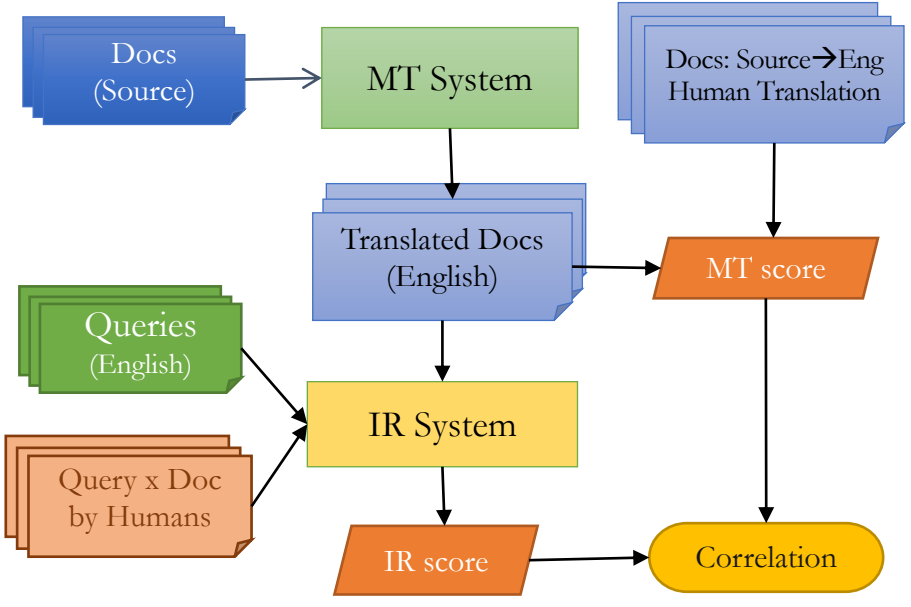
MacroF1 has more wins in the recent year -- when systems are mostly fluent, adequacy is a key discriminator

CLIR Task: Pipeline

CLSSTS 2020 / IARPA MATERIAL

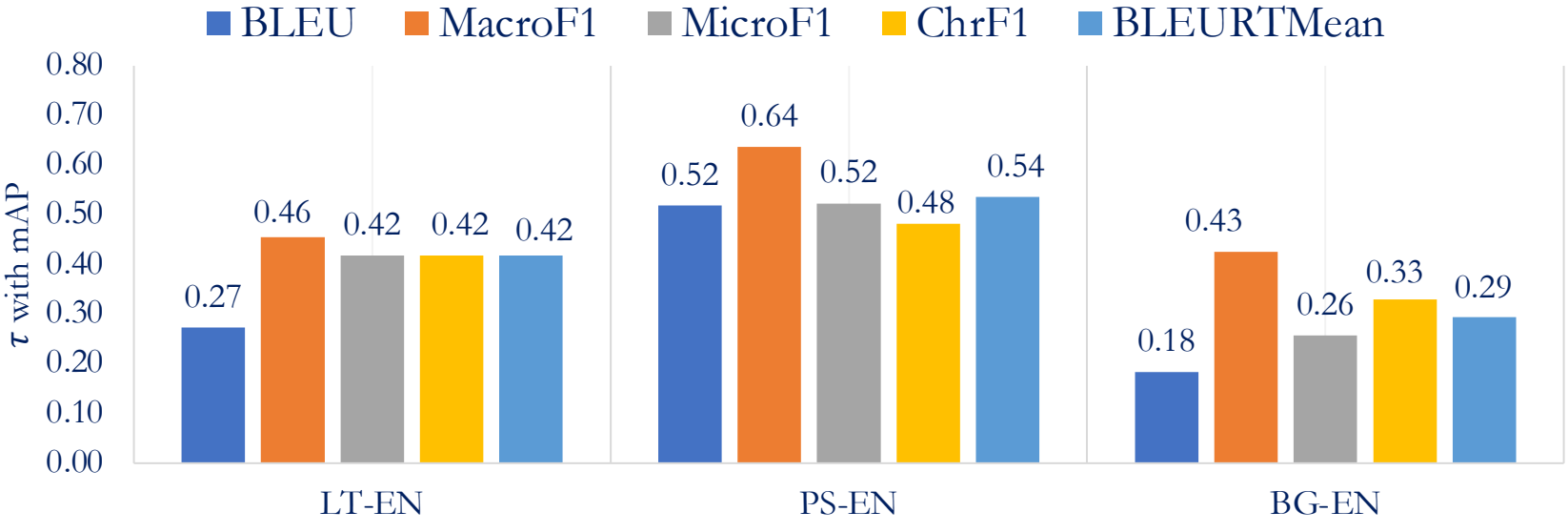
IR task with queries and docs in different languages

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
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

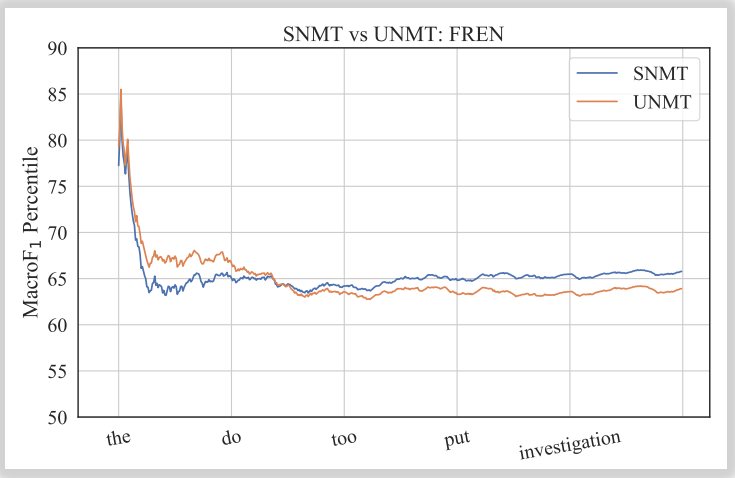
- IR task with queries and docs in different languages
- Translate source docs to target language, and match queries with docs
- MT metric having strong correlation with IR metric (e.g., mAP) is more useful

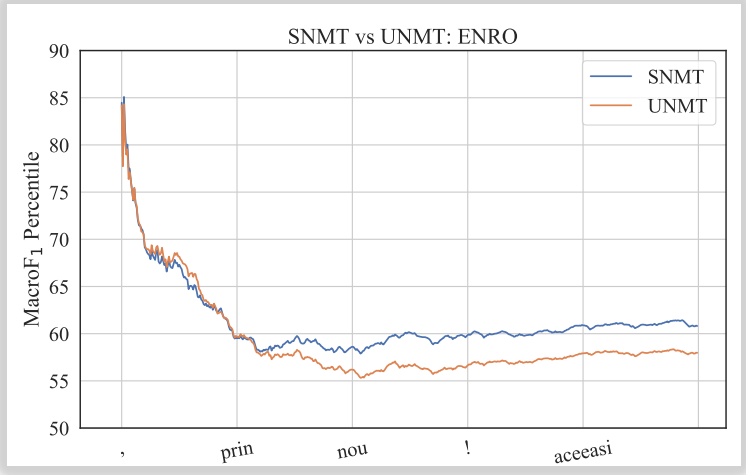


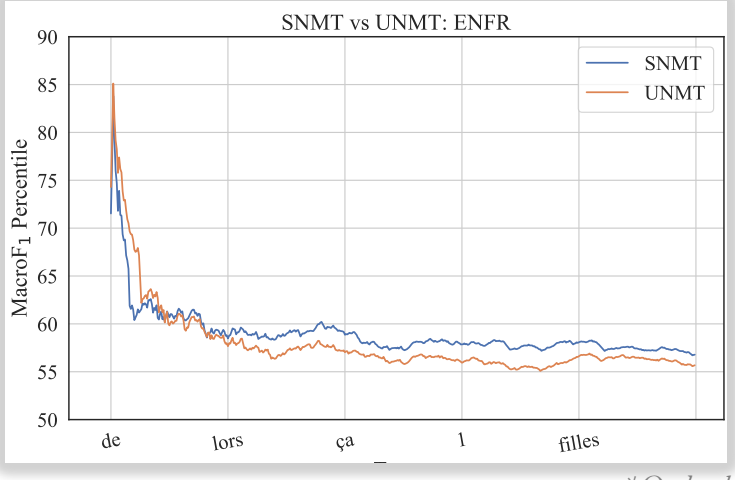
MacroF1 is the strongest indicator of downstream IR task performance

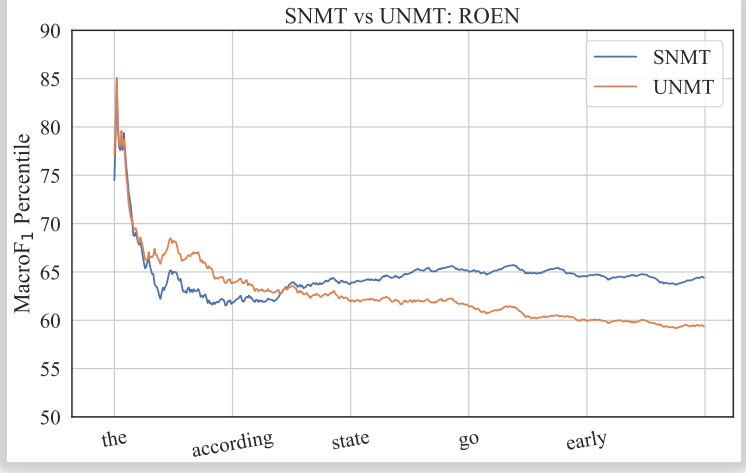
MacroF1 Difference Between SNMT and UNMT

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



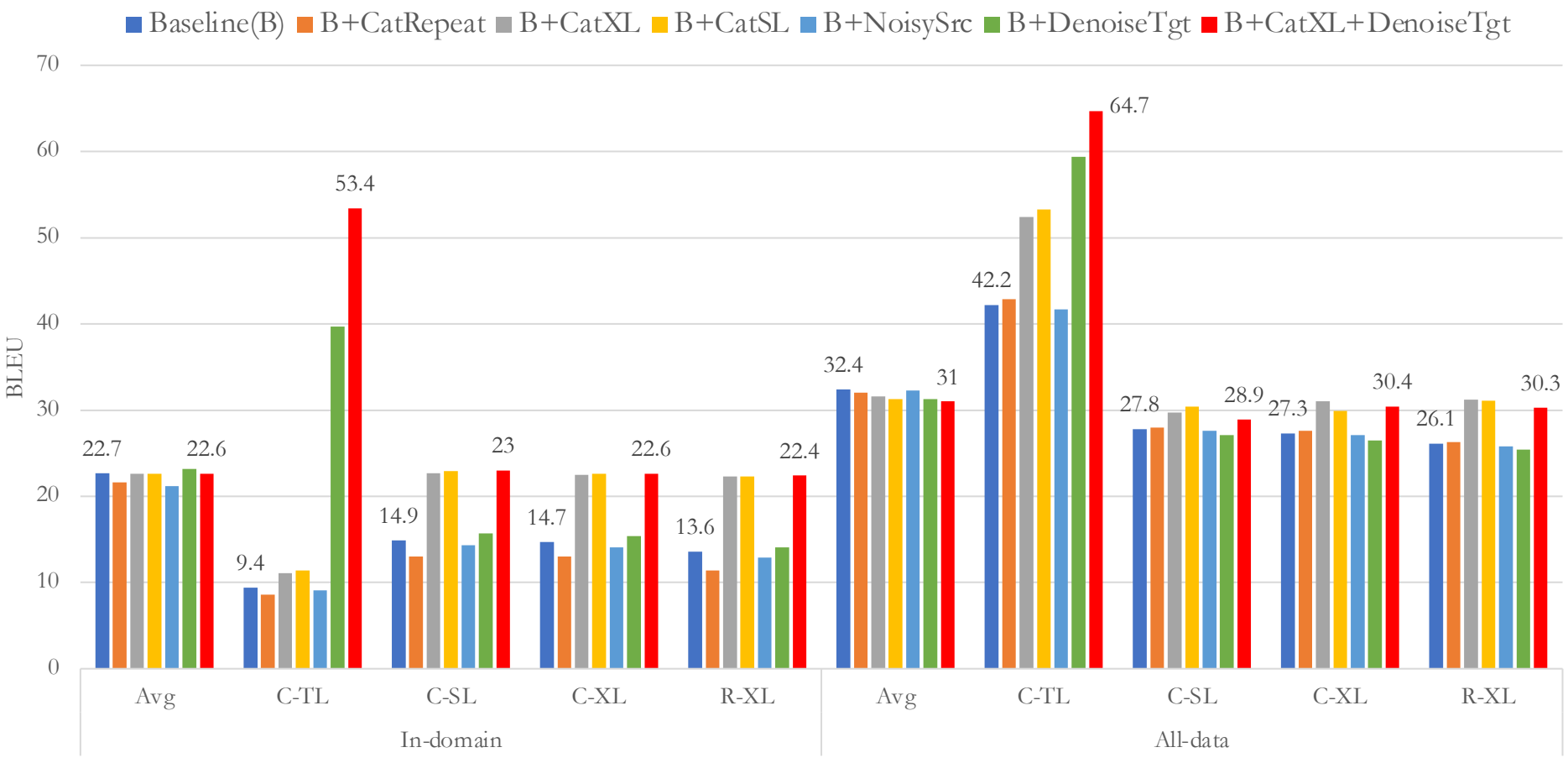






Only the top 500 types are visualized

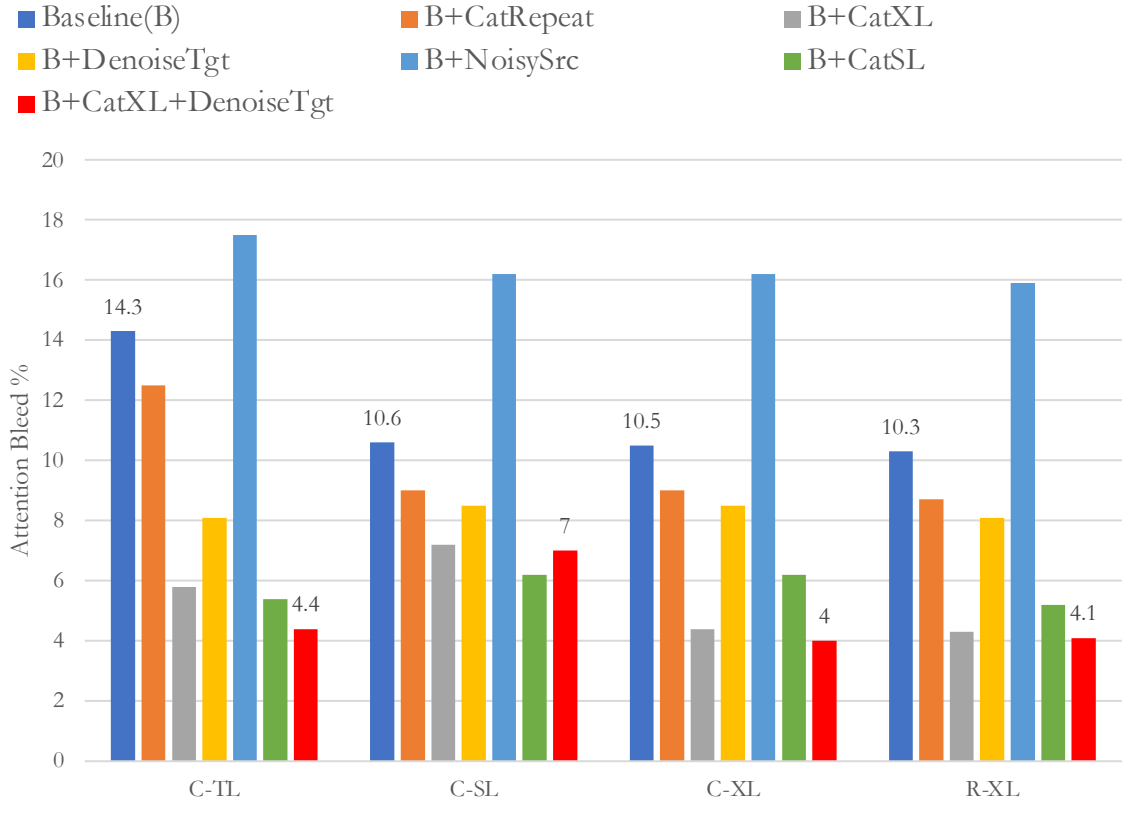
Results: BLEU



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

Attention Bleed

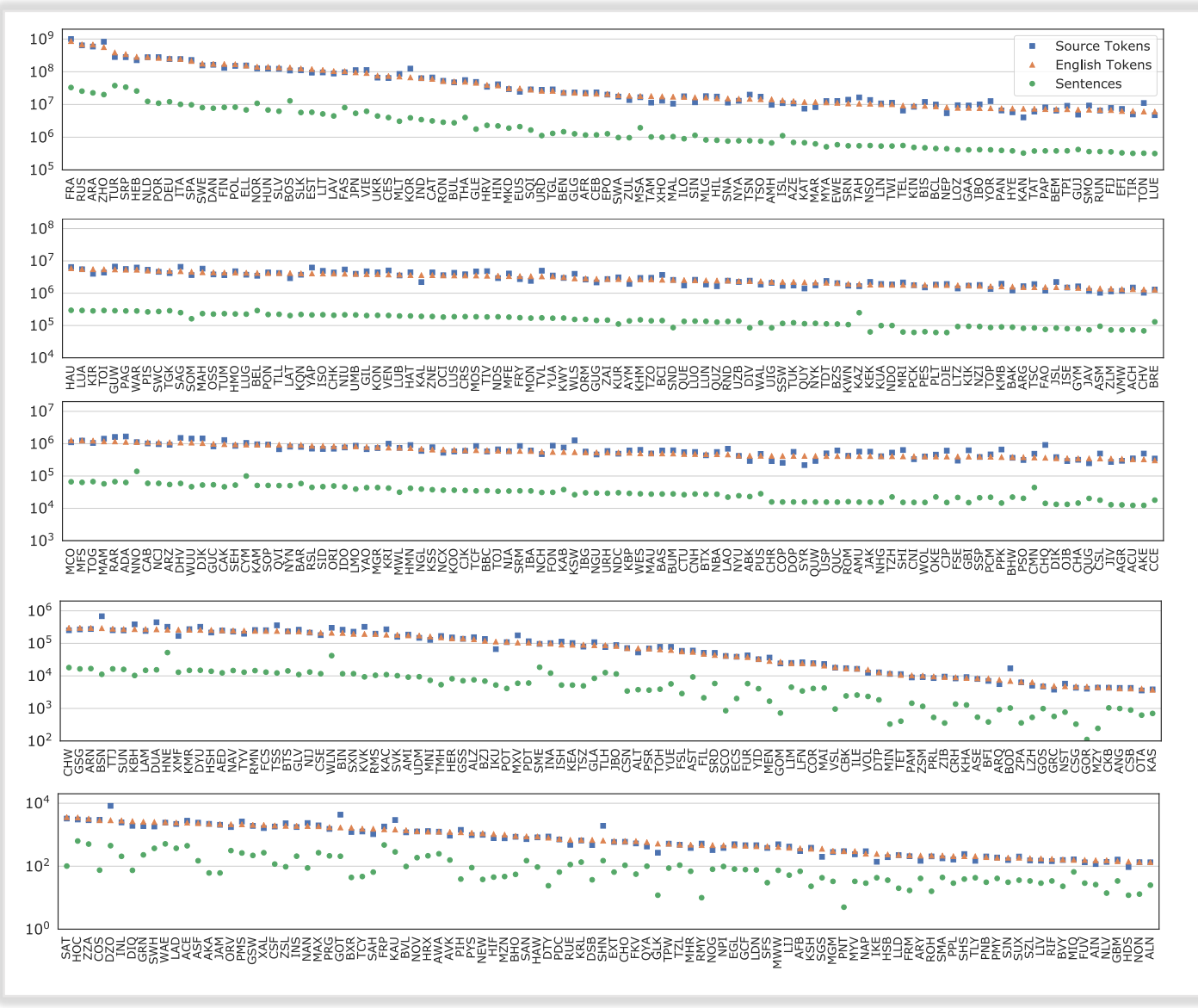
- Cross-attention mass crossing sentence boundaries in concatenated test sentences
  - Lower is better
- Average *attention bleed* across
  - All sentences
  - Transformer layers
  - Attention heads
- Models trained on augmented sentences achieve lower bleed
  ➔ Learn better attention



Attention bleed in models trained on all-data

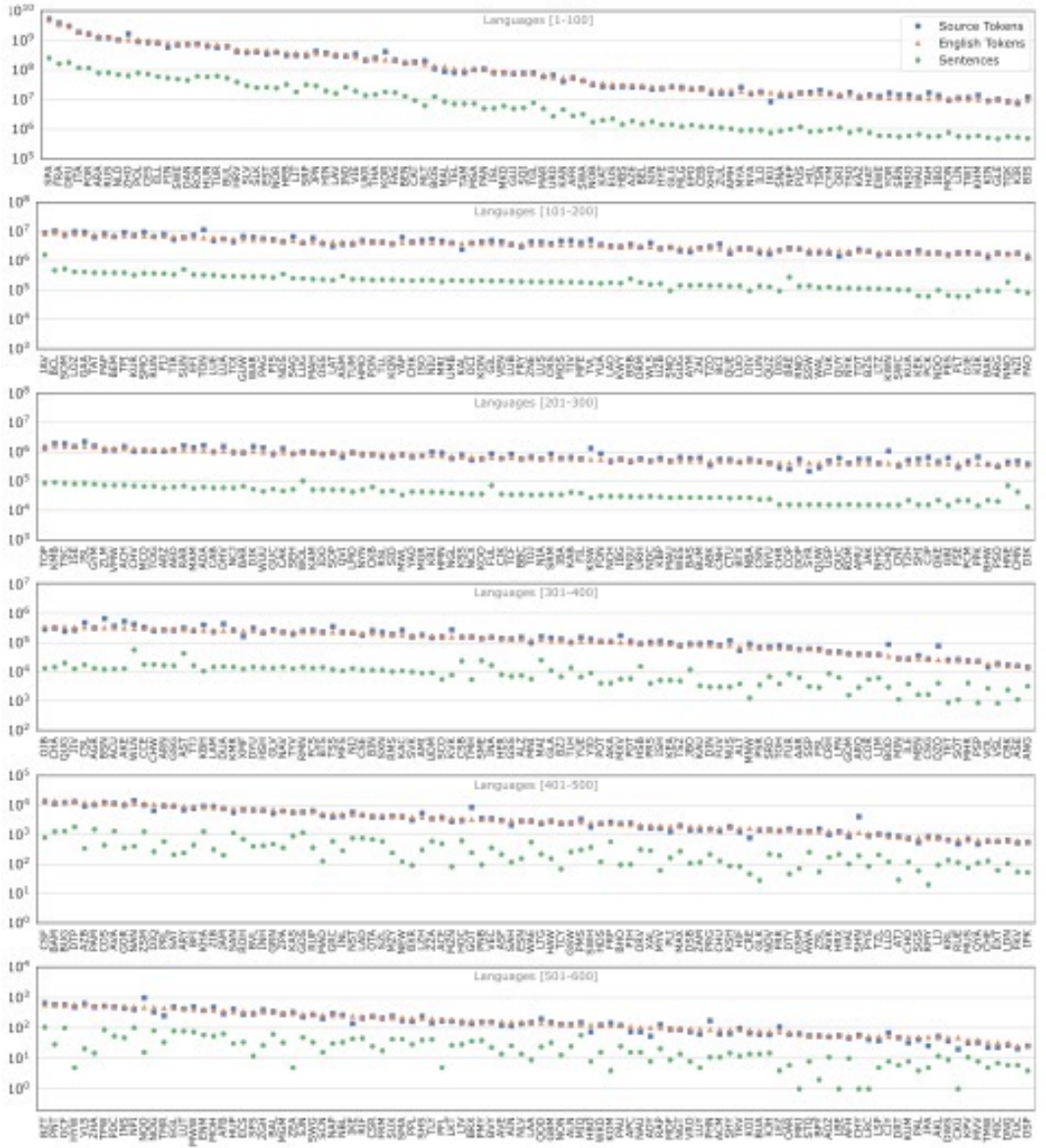
V1 Dataset Sizes: 500 Languages

* ISO 639-3 codes



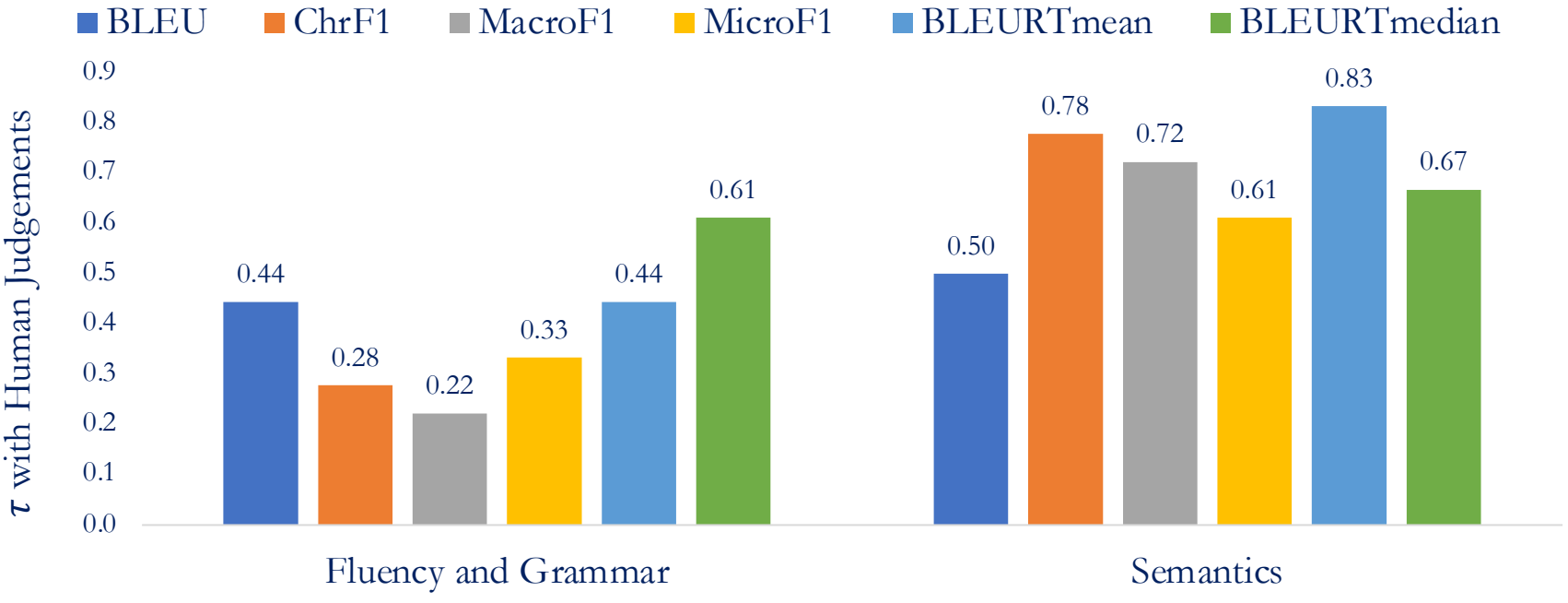
V2 Dataset Stats: 600 Languages

* ISO 639-3 codes



WebNLG Data-to-Text Evaluation

Correlation with Fluency, Grammar, and Semantics on English only



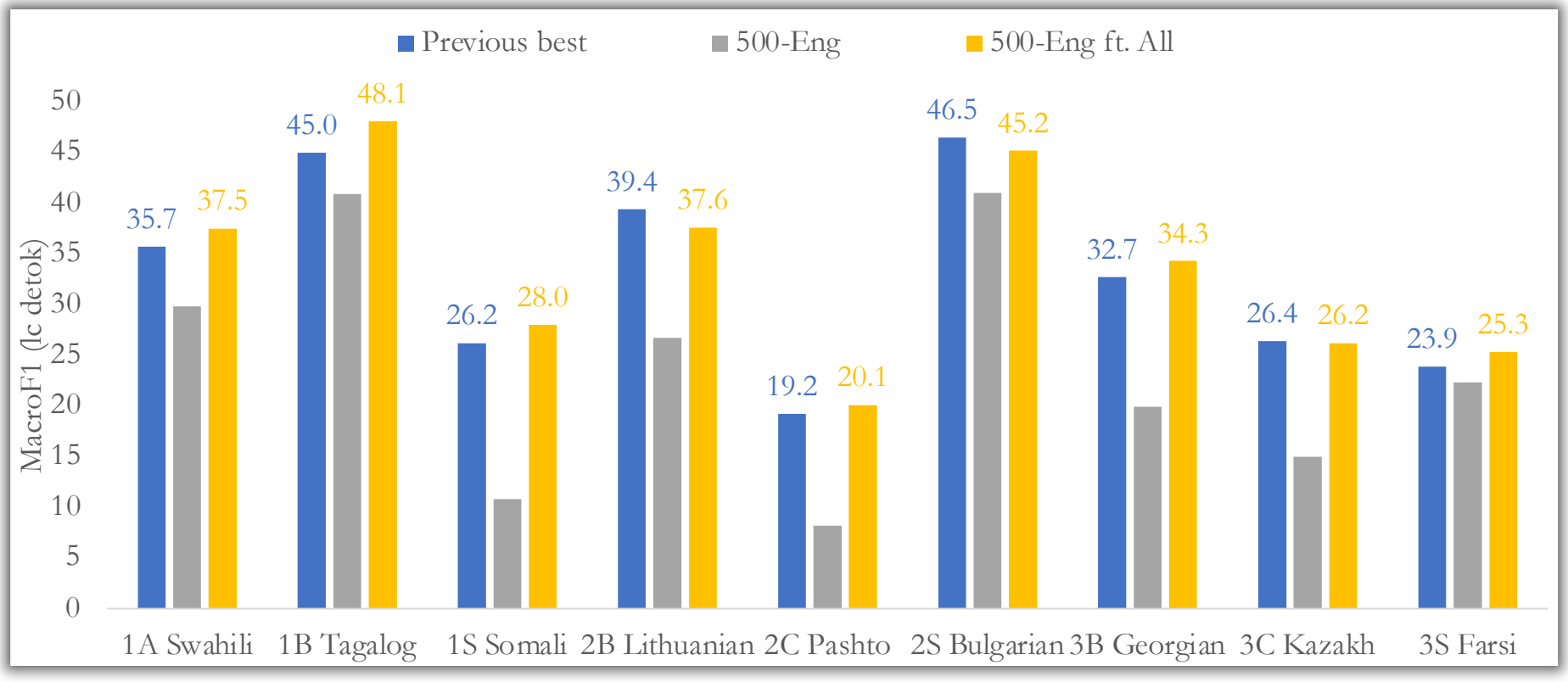
MacroF1 is a poor indicator of Fluency + Grammar, but one of the strong indicators of Semantics

Transfer Learning: Fine Tuning

- E.g., two low-resource languages
  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**

Finetuning: MacroF1
on IARPA MATERIAL Datasets (Analysis)

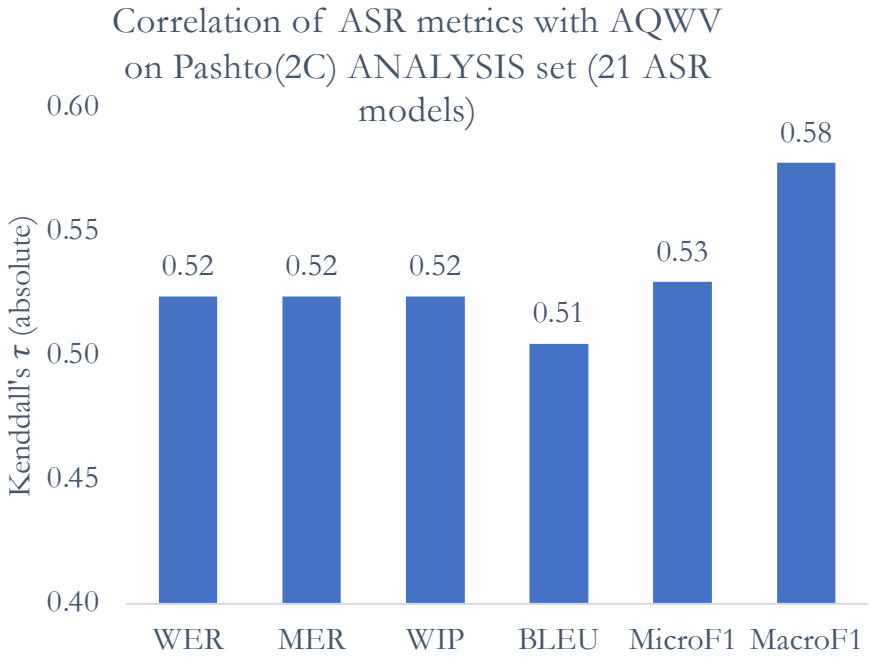


"Previous best" : the best bilingual models used in IARPA MATERIAL evaluations; separate model for each language

MacroF1 for ASR Evaluation

- Which ASR metric is the best indicator of downstream CLIR task performance?
  - Word error rate (WER)
  - Match error rare (MER)
  - Word information preserved (WIP)
- Other metrics: BLEU, MicroF1, MacroF1
- Correlation between ASR metrics and AQWV
  - Using Kendall's rank coefficient, τ

➔MacroF1 is a strong indicator, on Pashto (2C)
(with over 99% confidence)



Classifier Evaluation

- Test set, $T = \{ (h^{(i)}, y^{(i)}) \mid i = 1,2,3, \ldots m \}$ of *(hypothesis, reference)*
- Classes are the word types, after tokenization
  - We use the same tokenizer as BLEU, as implemented in SacreBLEU
- $C(c, a)$ counts the number of tokens of type c in sequence a
- $\text{Preds}(c) = \sum_{i=1}^{m} C(c, h^{(i)})$; $\text{Refs}(c) = \sum_{i=1}^{m} C(c, y^{(i)})$
- $\text{Match}(c) = \sum_{i=1}^{m} \min\{C(c, h^{(i)}), C(c, y^{(i)})\}$ *[BLEU, Papineni et al 2002]*
- Precision, $P_c = \frac{\text{Match}(c)}{\text{Preds}(c)}$; Recall, $R_c = \frac{\text{Match}(c)}{\text{Refs}(c)}$
- F-measure per class c: $F_{\beta;c} = (1+\beta)^2 \frac{P_c \times R_c}{\beta^2 \times P_c + R_c}$