Macro-Average: Rare Types Are Important Too

Thamme Gowda
Information Sciences Institute
University of Southern California
tg@isi.edu

Weiqiu You
Dept of Computer and Information Science
University of Pennsylvania
weiqiuy@seas.upenn.edu

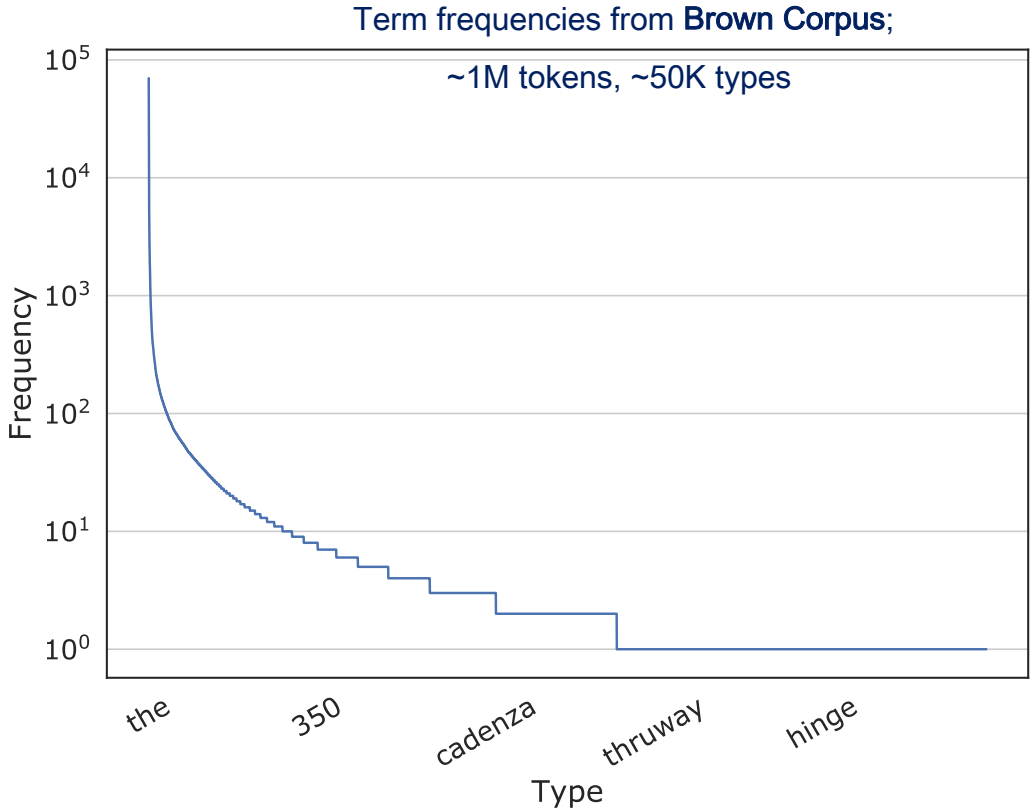
Constantine Lignos
Michtom School of Computer Science
Brandeis University
lignos@brandeis.edu

Jonathan May
Information Sciences Institute
University of Southern California
jonmay@isi.edu

Presented at NAACL 2021

Zipfian Distribution

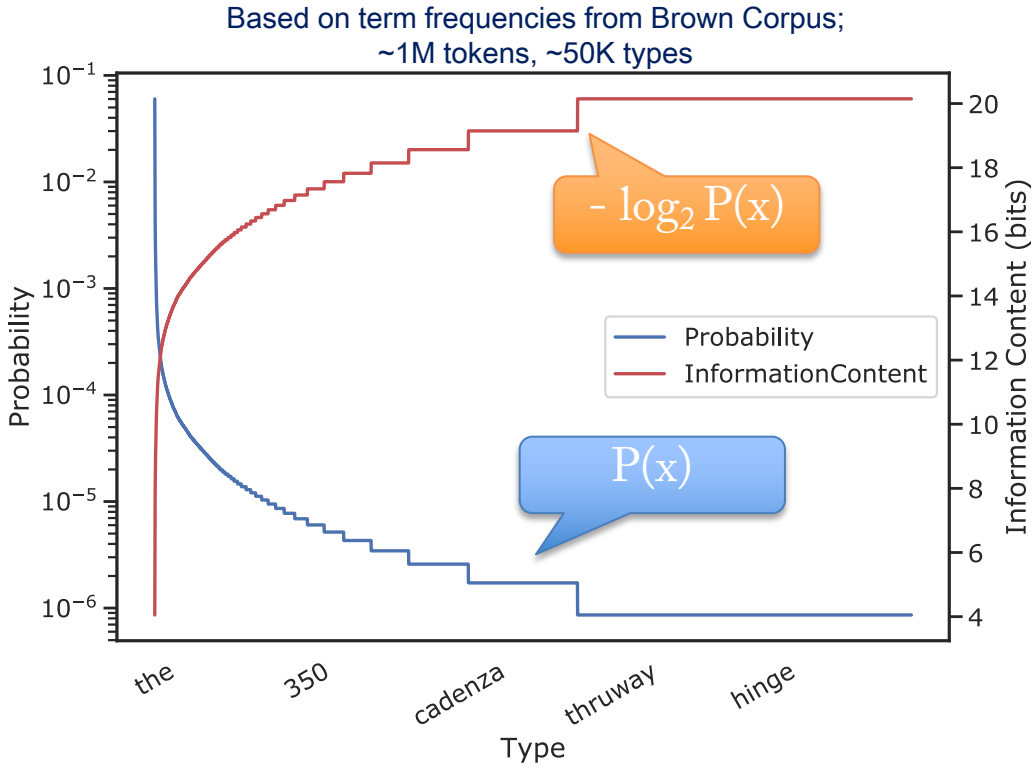
- Imbalanced types



Motivation

- Imbalanced types
- Machine learning techniques suffer from class imbalance
  - Favor the frequent types out-of-the-box
- Rare types are important

  ["Last words", Steedman, 2008]
- "Extra care" is needed to favor the rare types
  - During modeling (recent: Gowda & May, 2020)
  - During evaluation (this work)



Overview

1. Evaluate MT on imbalanced data using *MacroF1*
2. Justification for MacroF1:
    1. Direct human assessment
    2. Downstream task: cross-lingual IR
3. Qualitative differences between Supervised and Unsupervised NMT:
   Disagreement between BLEU and MacroF1

Evaluating MT as a Multi-class Classifier *on Imbalanced Test Sets*

NMT as Classifier

- MT : $(x_1 x_2 x_3 \ldots x_m) \rightarrow (y_1 y_2 y_3 \ldots y_n)$
- NMT: $y_{1:n}$ = Decoder(Encoder($x_{1:m}$))
  - Maximize $\prod_{t=1}^{n} P(y_t \mid y_{<t}, x_{1:m}; \theta)$
  - Maximize $\prod_{t=1}^{n} P(y_t \mid f(y_{<t}, x_{1:m}; \theta_1);\ \theta_2)$

NMT: Autoregressor + Classifier *[Gowda and May, 2020]*

1. Autoregressor: $\mathrm{h_t} = f(y_{0:<t}, x_{1:m})$ $h_t \in R^d$ is continuous
2. Classifier : $P(y_t \mid h_t)$ y_t is discrete.

   |Vocabulary(Y)| > 2 ⇒ Multi-class classifier

Class Performance

- 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***
- Preds(c) $= \sum_{i=1}^{m} C(c, h^{(i)})$; Refs(c) $= \sum_{i=1}^{m} C(c, y^{(i)})$
- Match(c) $= \sum_{i=1}^{m} \min\{C(c, h^{(i)}), C(c, y^{(i)})\}$ *[BLEU, Papineni et al 2002]*
- Precision, $P_c = \frac{Match(c)}{Preds(c)}$; Recall, $R_c = \frac{Match(c)}{Refs(c)}$
- F-measure per class c: $F_{\beta;c} = (1+\beta)^2 \frac{P_c \times R_c}{\beta^2 \times P_c + R_c}$

Multiclass Classifier Performance

Overall performance = an average of individual class performances

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

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

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

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

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

- *We use* $k = 1$; Note: if $k \to \infty, \text{M}icroF_\beta \to \text{Macro}F_\beta$
- *We use* $\beta = 1$ we scale [0.0, 1.0] to [0, 100], just like BLEU

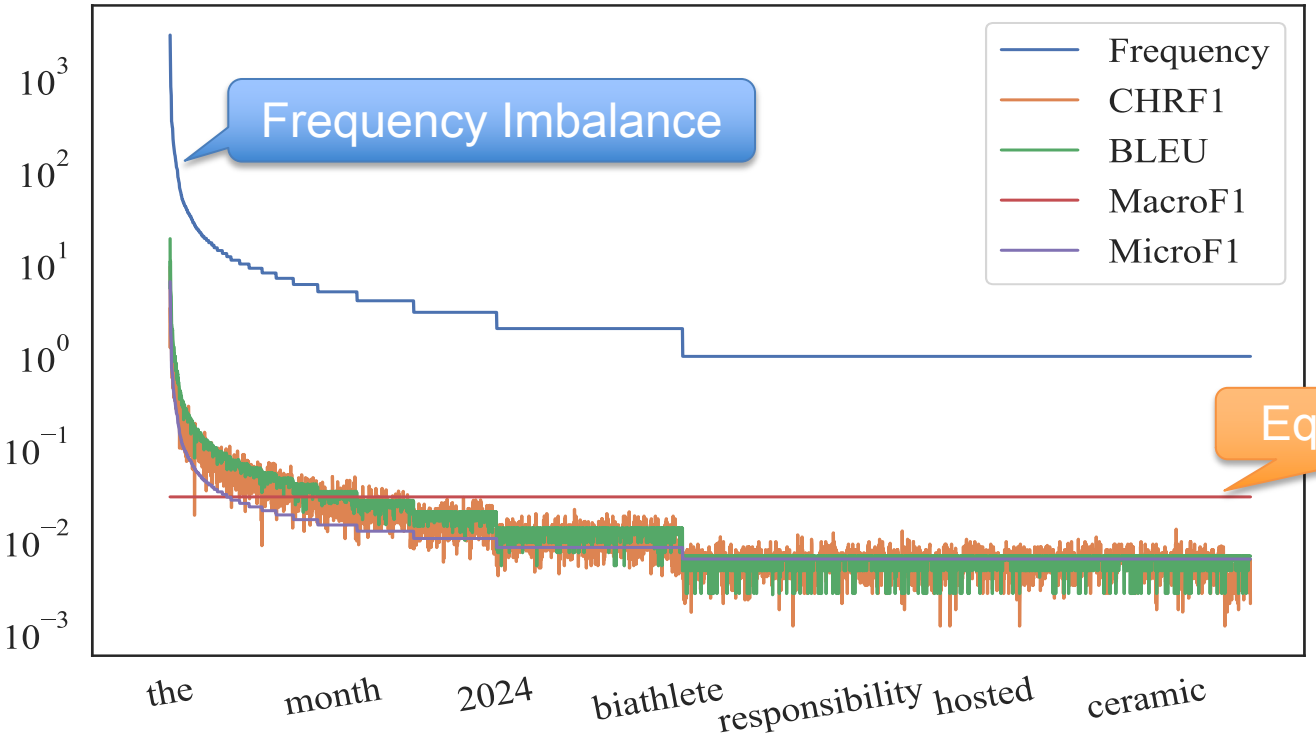
Justification for MacroF1 as an MT evaluation metric

➔Compare MacroF1 with

- BLEU and ChrF1 – popular options
- 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)*

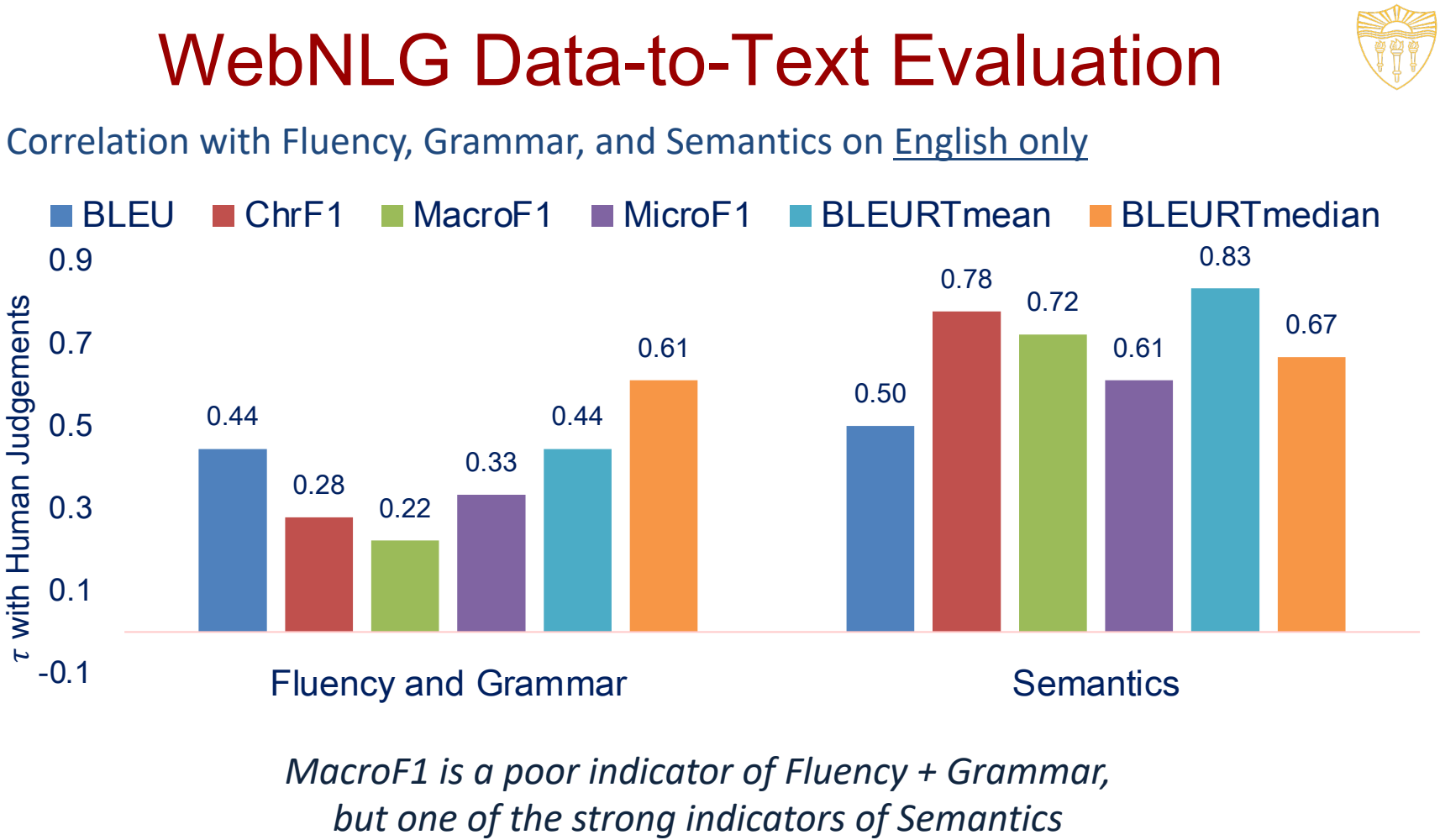
MacroF1 vs Others



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

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

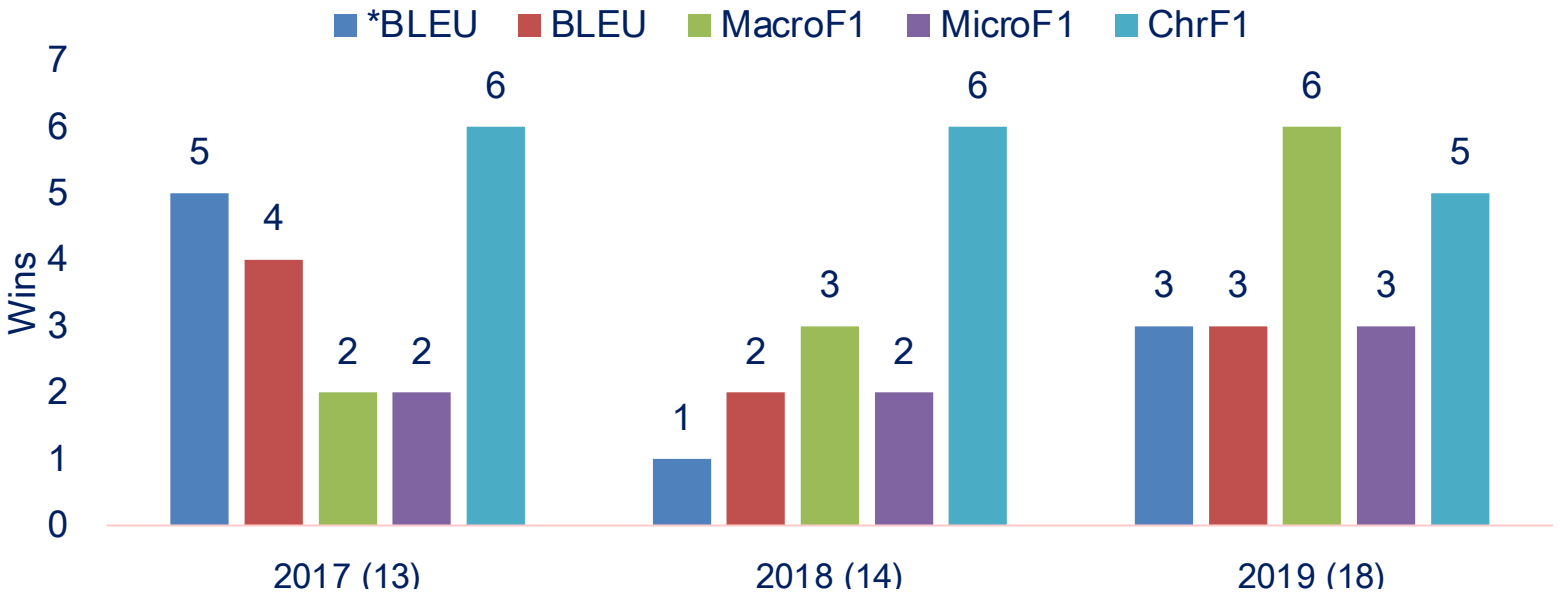
WebNLG Data-to-Text Evaluation



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

WMT Metrics: Wins per Metric

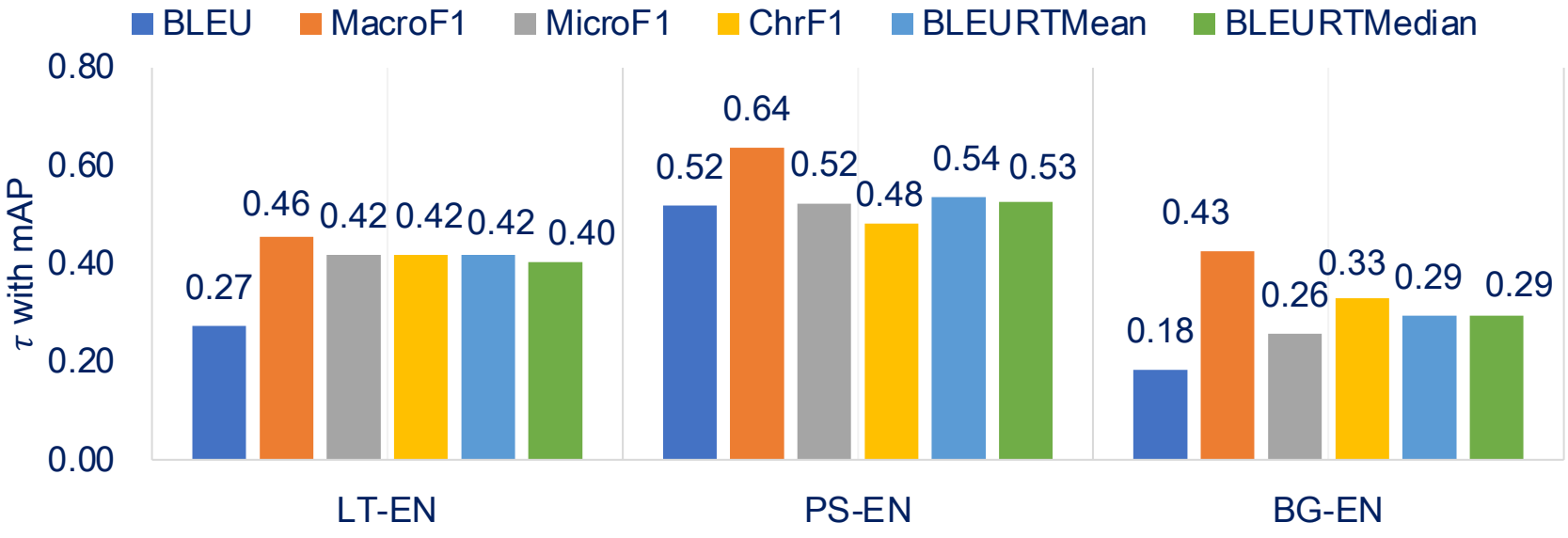
- Wins = Number of times a metric scored highest correlation with human judgements
- *BLEU is from the WMT metrics package, precomputed by task organizers
- MacroF1 and MicroF1 use the same tokenizer as BLEU, obtained using SacreBLEU



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

Downstream CLIR Task (CLSSTS 2020)

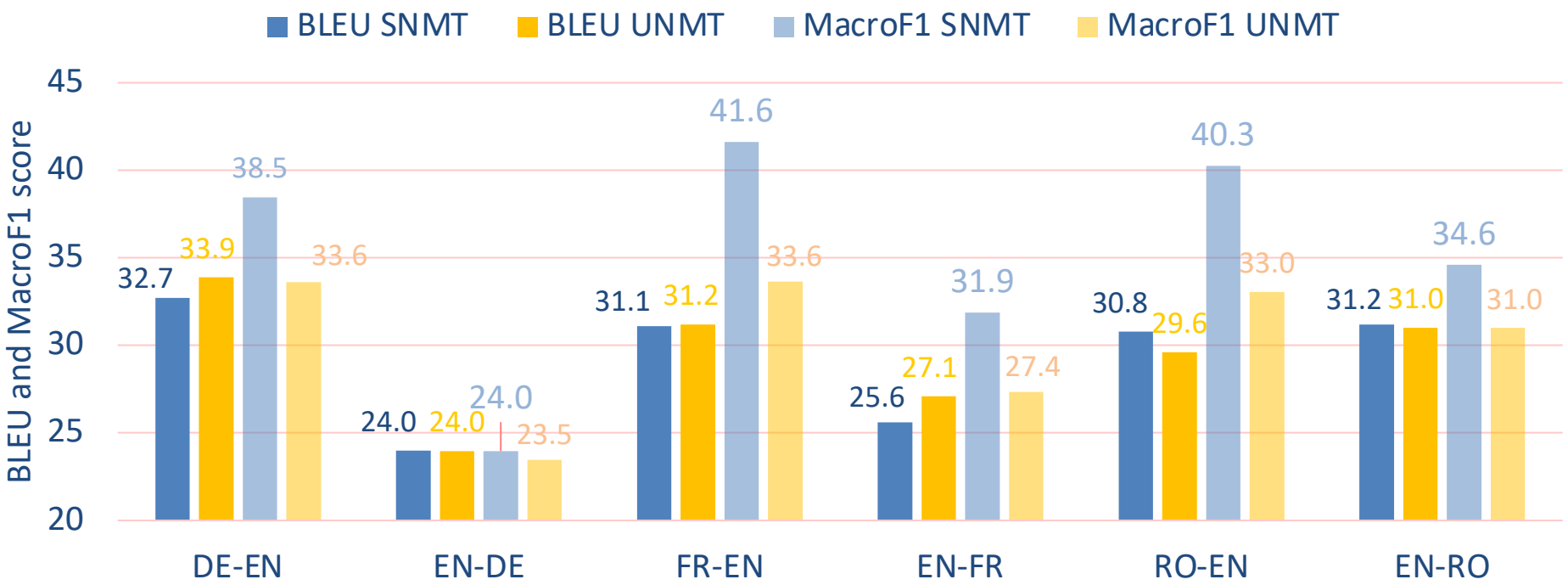
- IR task with queries and docs in different languages
- Translate source docs to target language, and match queries with docs
- MT metric having strong correlation with IR metric (e.g. mAP) is more useful



MacroF1 is the strongest indicator of downstream IR task performance

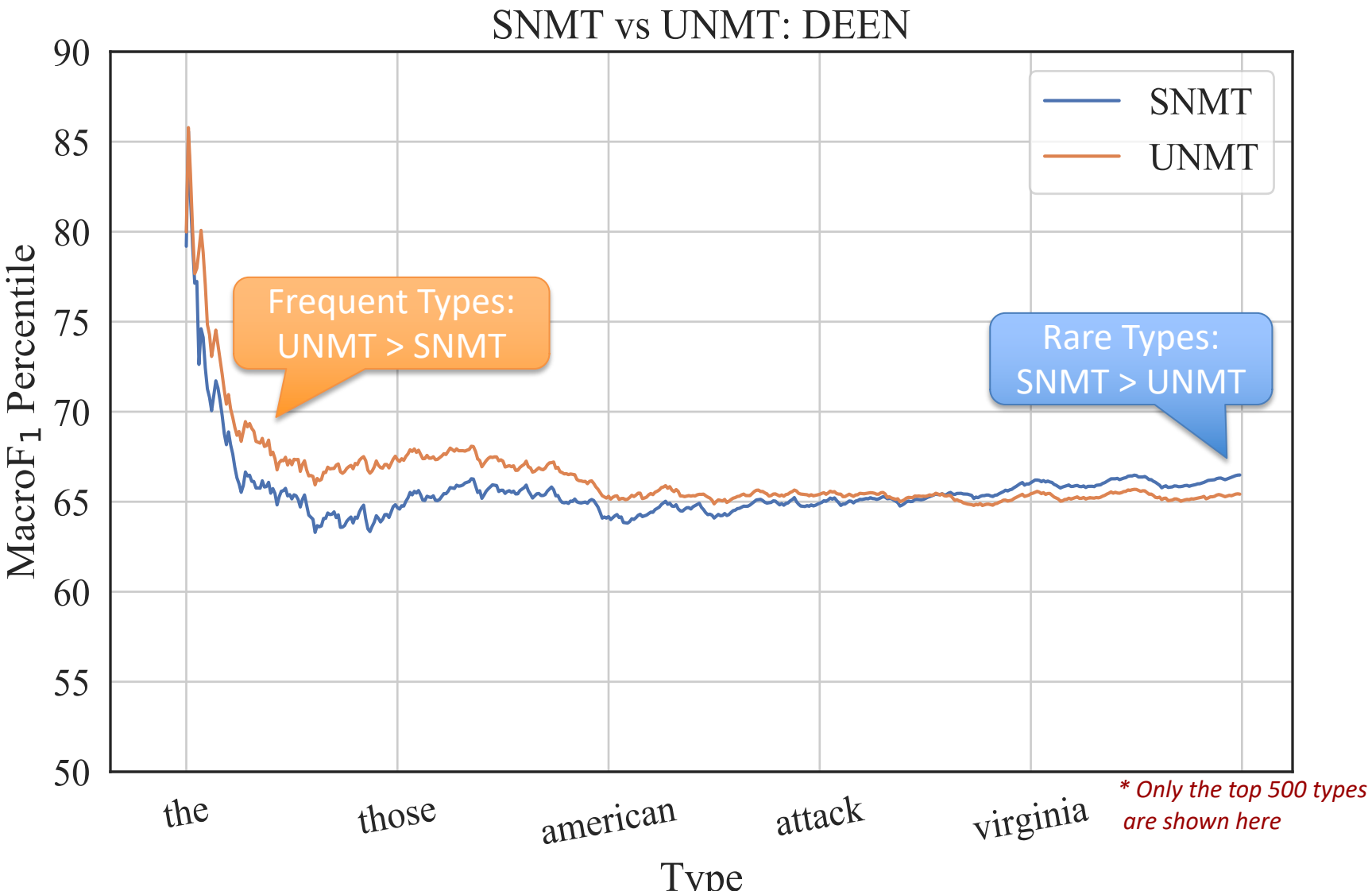
Qualitative Difference Between Supervised and Unsupervised NMT

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


SNMT vs UNMT: DEEN
90
85
80
75
70
65
60
55
50
MacroF1 Percentile
SNMT
UNMT
Frequent Types:
UNMT > SNMT
Rare Types:
SNMT > UNMT
the
those
american
attack
virginia
Type
* Only the top 500 types
are shown here

SNMT and UNMT

E.g.1: MacroF1 is more intolerant to untranslation than BLEU

	MacroF1 diff: 0.044, favors SNMT; BLEU diff: -0.00087, favors UNMT
Src	"Er ist witzig, er ist sarkastisch, witzig, er liebt es, Leute zu **unterhalten**, und er ist ein großartiger Kerl, den er im Teamraum haben kann", erklärte er.
Ref	"He's funny, he's sarcastic, witty, likes to poke fun at people, and he's a great guy to have in the team room," he explained.
SNMT	"He is funny, he is sarcastic, funny, he loves to talk people, and he is a great guy he can have in the team room," he explained
UNMT	"He's funny, he's sly, funny, he loves to **unterhalten** people, and he's a great guy he can have in the team room," he explained.
Problems	SNMT: punctuation, *wrong_verb*. UNMT: synonym, ***untranslation***

SNMT and UNMT

E.g. 2 MacroF1 is more intolerant to truncation than BLEU

	MacroF1 diff: 0.044, favors SNMT; BLEU diff: -0.00087, favors UNMT
Src	Vor 32 Jahren schloss ich mich als Schüler, wegen der Vernachlässigung der Thatcher-Regierung, Labour an. **Diese Vernachlässigung hatte dazu geführt, dass mein Klassenzimmer buchstäblich zusammengebrochen war. Infolgedessen habe ich versucht, mich für bessere öffentliche Dienstleistungen für diejenigen einzusetzen, die sie am meisten brauchen. Egal ob als Gemeinderat oder Minister.**
Ref	Ever since I joined Labour 32 years ago as a school pupil, provoked by the Thatcher government's neglect **that had left my comprehensive school classroom literally falling down, I've sought to champion better public services for those who need them most whether as a local councillor or government minister.**
SNMT	32 years ago, I joined Labour as a student because of the neglect of the Thatcher government, **which had led to my classroom literally collapsed, and as a result I tried to promote better public services for those who need it most, whether as a local council or ministers.**
UNMT	Last 32 years ago, as a student, because of the disdain for the Thatcher-era government, Labour joined Labour.
Problems	SNMT: synonym, word_order UNMT: *subject,* ***truncation****, word_order*

Summary

- ✓ MT evaluation as multiclass classifier on imbalanced test set
- ✓ Justified MacroF1 as a legitimate eval metric
  - ✓ Direct assessment: WebNLG, and WMT [2017-19]
  - ✓ Downstream task: Cross lingual information retrieval
- ✓ SNMT vs UNMT qualitative differences

References

- Mark Steedman. 2008. On Becoming a Discipline. Computational Linguistics 34, 1 (2008), 137-144. https://doi.org/10.1162/coli.2008.34.1.137 arXiv: https://doi.org/10.1162/coli.2008.34.1.137
- Gowda, Thamme, and Jonathan May. "Finding the Optimal Vocabulary Size for Neural Machine Translation." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, 1715-1725. https://doi.org/10.18653/v1/P16-1162
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **BLEU**: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311-318. https://doi.org/10.3115/1073083.1073135
- Maja Popović. 2015. **ChrF:** character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Lisbon, Portugal, 392-395. https://doi.org/10.18653/v1/W15-3049
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. **BLEURT**: Learning Robust Metrics for Text Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 7881-7892. https://www.aclweb.org/anthology/2020.acl-main.704
- Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages. In Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (**CLSSTS2020**). European Language Resources Association, Marseille, France, 7-13. https://www.aclweb.org/anthology/2020.clssts-1.2

- Fork of SacreBLEU: https://github.com/isi-nlp/sacrebleu
- Pull request is under review: https://github.com/mjpost/sacrebleu/pull/153

THANK YOU 🙏