Macro-average: Rare Types Are Important Too

USC Viterbi School of Engineering
Information Sciences Institute

Thamme Gowda Weiqiu You Constantine Lignos Jonathan May

SCAN ME

Introduction:

- Natural language types are imbalanced
  - A few types are frequent ⇒ less *information content*
  - Most types are rare ⇒ more *information content*
- Current eval metrics do not address imbalance
- MacroF measure is used for classifier evaluation on imbalanced test sets; we apply MacroF for machine translation evaluation
- MacroF shows differences in quality of supervised and unsupervised NMT in a way BLEU cannot

Methods

$$\text{Match}(\mathbf{c}) = \sum_{i=1}^{m} \min\{C(c, h^{(i)}), C(c, y^{(i)})\}$$

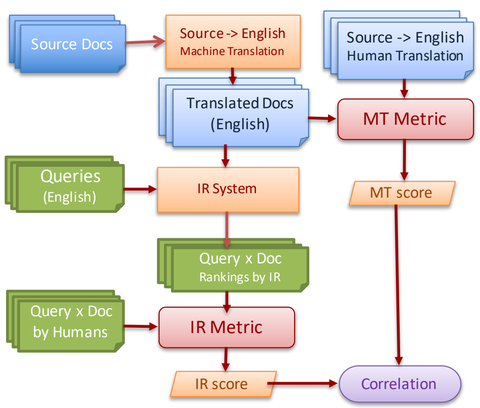
$$P_c = \frac{Match(c)}{Preds(c)} \qquad R_c = \frac{Match(c)}{Refs(c)}$$

$$F_{\beta;c} = (1+\beta)^2 \frac{P_c \times R_c}{\beta^2 \times P_c + R_c}$$

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

$$\text{Micro}F_\beta = \frac{\sum_{c \in V} f(c) \times F_{\beta;c}}{\sum_{c' \in V} f(c')} \text{ where } f(c) = Refs(c) + k;\ k \geq 1$$

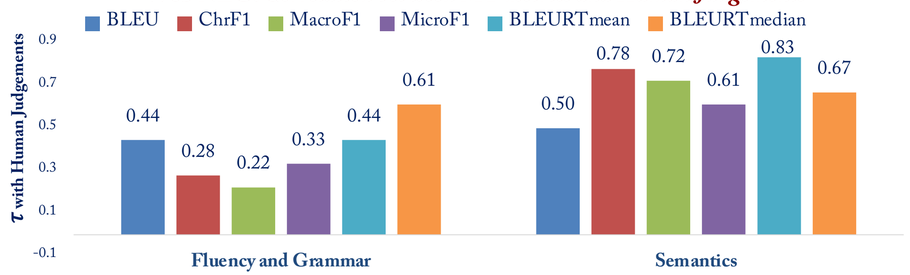
CLIR Pipeline



Justification

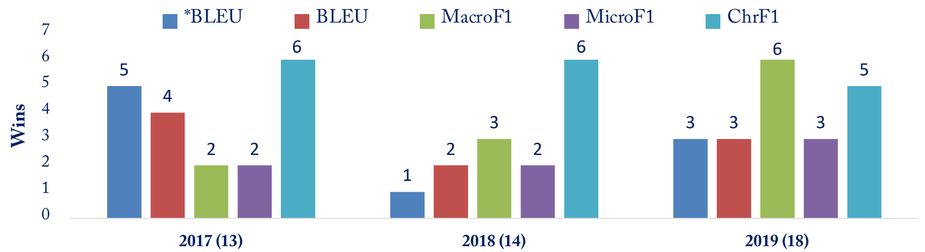
1. Direct assessment : WebNLG , WMT Metrics Task
2. Downstream task performance indication

Web NLG: ***Correlation between MT metrics and Human Judgements***



MacroF1 is a poor indicator of fluency and grammar; but a strong indicator of semantics; These results are based on English only.

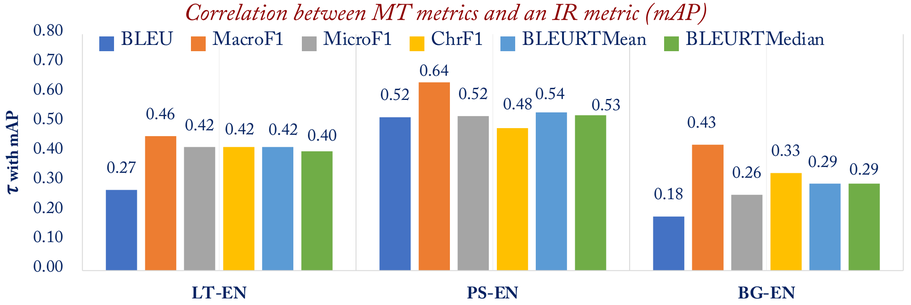
WMT Metrics Task : ***Correlation between MT metrics and Human Judgements***



MacroF1 has highest number of wins in the recent year when most systems are fluent, and adequacy is the key discriminating factor.
**BLEU is precomputed values in metrics package. MacroF1, MicroF1 share the same tokenizer as BLEU, obtained using SacreBLEU*

Cross-lingual Information Retrieval Task: CLSSTS 2020

Correlation between MT metrics and an IR metric (mAP)

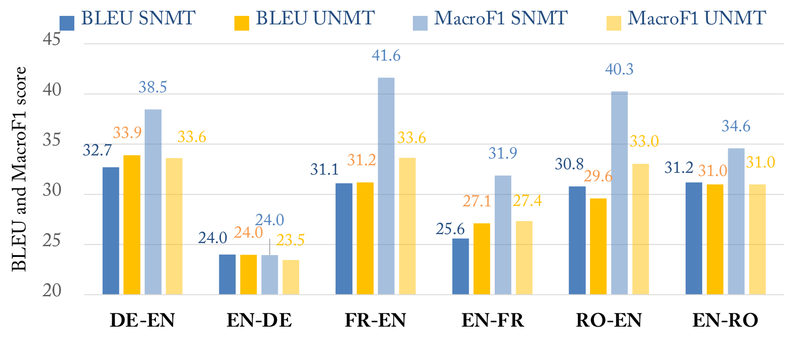


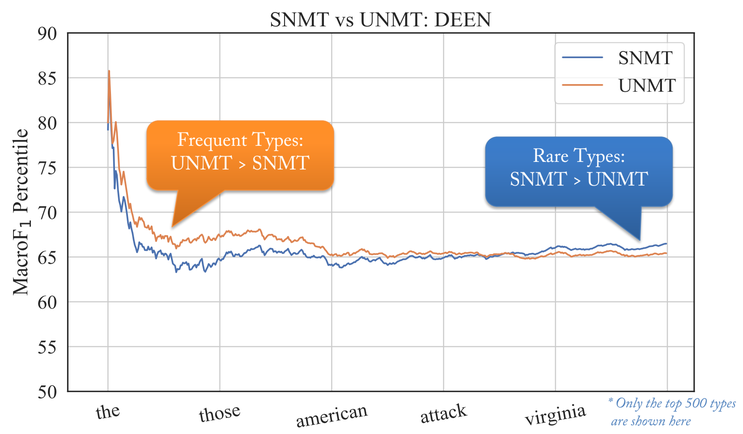
MacroF1 outperforms others while indicating downstream task performance.

Usage

```
$ pip install git+https://github.com:isi-nlp/sacrebleu.git
$ sacrebleu $REF -m macrof microf  < $HYP.detok
$ sacrebleu $REF -m macrof  --report report.txt < $HYP.detok
```

SNMT vs UNMT Quality Diff





Manual Analysis:
Untranslation and truncation are heavily penalized by MacroF1 than BLEU

Conclusion

- MacroF1 is a strong indicator of semantics; competitive on direct assessments and outperforms others on a downstream CLIR task
- MacroF1 is easily computable and interpretable, does not appear to have uncontrollable biases resulting from data (unlike model-based metrics)
- Macro-averaged evaluation is a useful technique for addressing the importance of long tail of language