A Unigram Precision and Recall

For the sake of clarity, consider a test dataset T of N pairs of parallel sentences, $(x^{(i)}, y^{(i)})$ where $x^{(i)}$ and $y^{(i)}$ are the i^{th} source and reference sequences, respectively. We use single reference $y^{(i)}$ translations for this analysis. For each $x^{(i)}$, let $h^{(i)}$ be the translation hypothesis from an MT model.

Let the indicator $\mathbb{1}_k^a$ have value 1 iff type c_k exists in sequence a, where a can be either hypothesis $h^{(i)}$ or reference $y^{(i)}$. The function $count(c_k, a)$ counts the times token c_k exists in sequence a; $match(c_k, y^{(i)}, h^{(i)})$ returns the times c_k is matched between hypothesis and reference, given by $min\{count(c_k, y^{(i)}), count(c_k, h^{(i)})\}$

Let $P_k^{(i)}$ and $R_k^{(i)}$ be precision and recall of c_k on a specific record $i \in T$, given by:

$$P_k^{(i)} = \frac{match(c_k, y^{(i)}, h^{(i)})}{count(c_k, h^{(i)})}, \text{ defined iff } \mathbb{1}_k^{h^{(i)}}$$

$$R_k^{(i)} = \frac{match(c_k, y^{(i)}, h^{(i)})}{count(c_k, y^{(i)})}, \text{ defined iff } \mathbb{1}_k^{y^{(i)}}$$

Let P_k, R_k be the expected precision and recall for c_k over the whole T, given by:

$$P_k = \mathbb{E}_{i \in T}[P_k^{(i)}] = \frac{\sum_{i=1}^{N} \mathbb{1}_k^{h^{(i)}} P_k^{(i)}}{\sum_{i=1}^{N} \mathbb{1}_k^{h^{(i)}}}$$

$$R_k = \mathbb{E}_{i \in T}[R_k^{(i)}] = \frac{\sum_{i=1}^{N} \mathbb{1}_k^{y^{(i)}} R_k^{(i)}}{\sum_{i=1}^{N} \mathbb{1}_k^{y^{(i)}}}$$

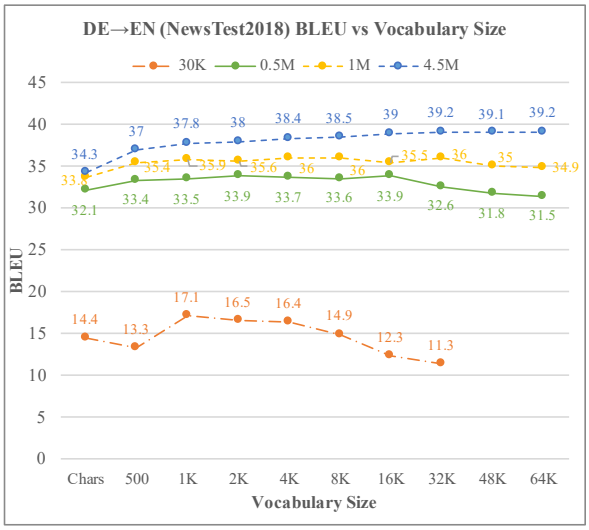
B More Visualizations

We provide BLEU scores on validation datasets in Figure 7.

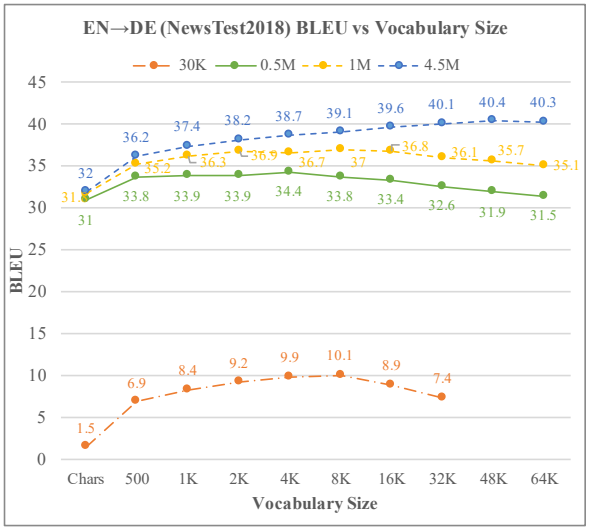
Figure 8 contains visualization of μ, D, $F_{95\%}$ for BLEU on test set for EN→DE 1M and DE→EN 1M.

Figure 9 contains visualization of frequency bias on EN→DE and EN→HI languages test sets.

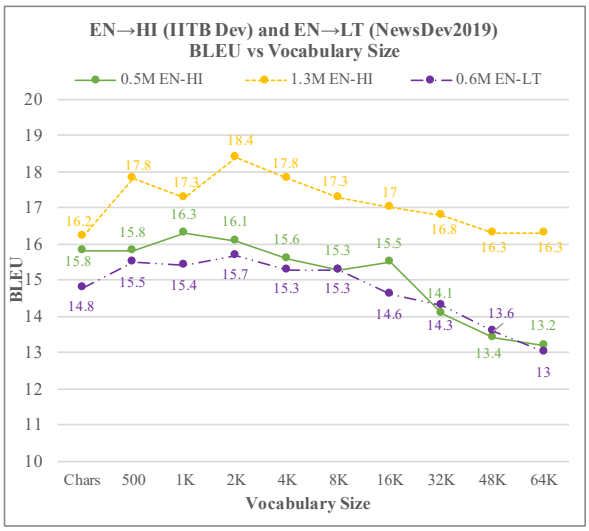
Figure 10 contains visualization of μ, D, $F_{95\%}$ for BLEU on validation sets.



(a)



(b)



(c)

Figure 7: BLEU on all validation sets as a function of vocabulary size at various training set sizes.

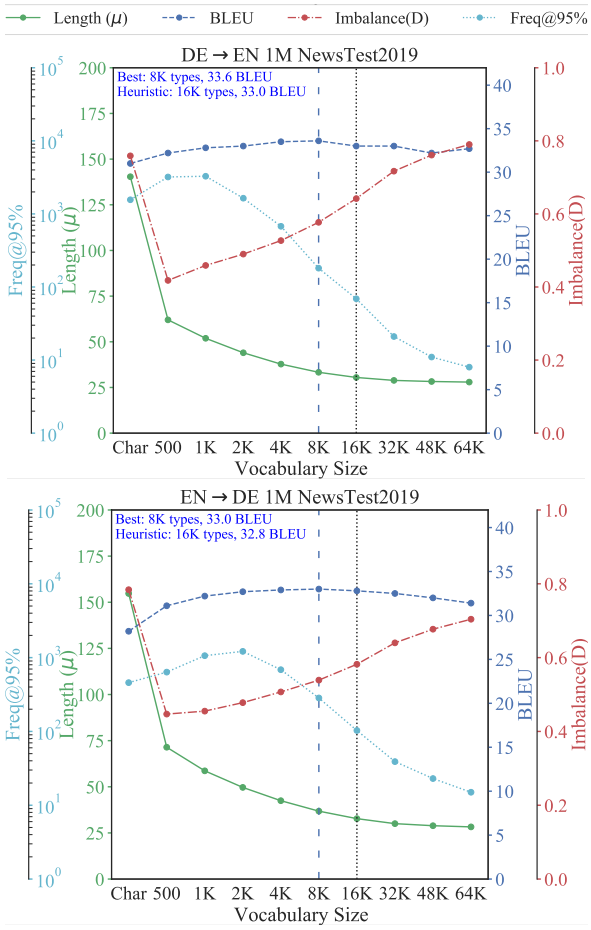


Figure 8: Visualization of sequence length (μ) (lower is better), class imbalance (D) (lower is better), frequency of 95^{th} percentile class ($F_{95\%}$) (higher is better; plotted in logarithmic scale), and test set BLEU (higher is better) on DE↔EN of 1M. This is a continuation of Figure 5.

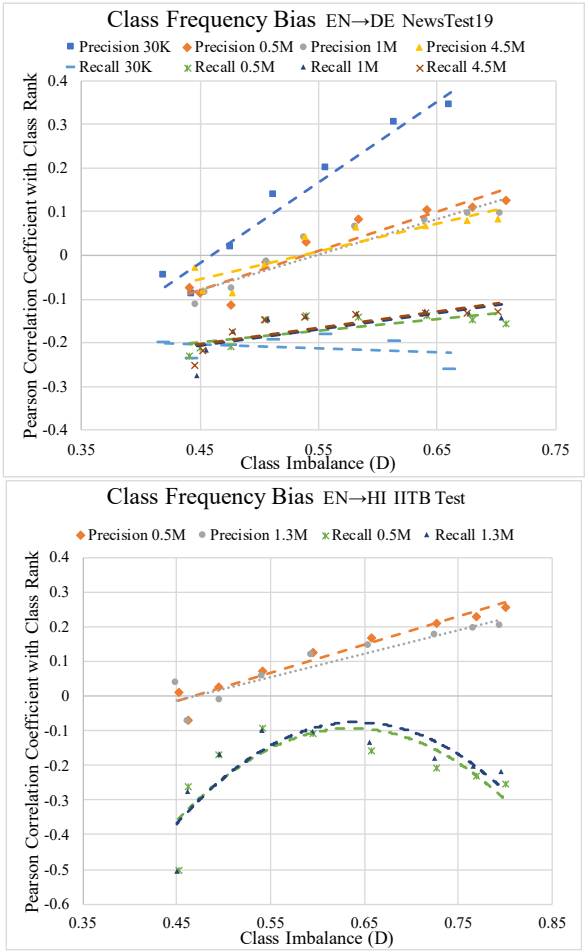


Figure 9: Correlation analysis on EN→DE, and EN→HI test sets. The non-zero correlation of precision and recall with class rank indicate the class frequency bias in NMT models.

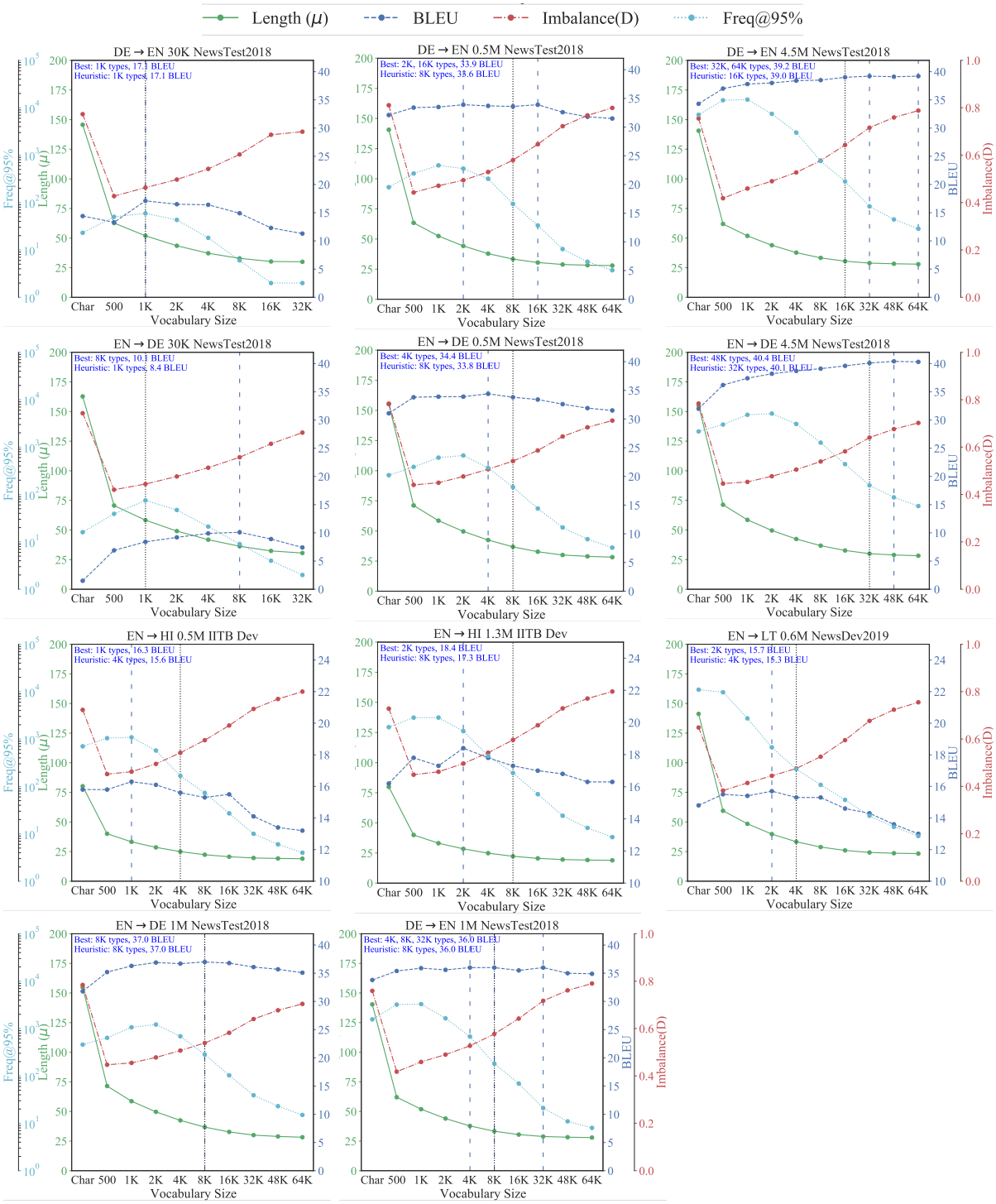


Figure 10: Visualization of sequence length (μ) (lower is better), class imbalance (D) (lower is better), frequency of 95^{th} percentile class ($F_{95\%}$) (higher is better; plotted in logarithmic scale), and *validation set* BLEU (higher is better) on all language pairs and training data sizes. The vocabulary sizes that achieved highest BLEU are indicated with dashed vertical lines.