Neural Machine Translation With Imbalanced Classes[*]

THAMME GOWDA, Dept. of Computer Science, University of Southern California, Los Angeles, CA, USA

Neural machine translation (NMT) models are commonly viewed as encoder-decoder networks; such a view is useful for implementation, however, inadequate for theoretical analysis. In this work, we cast NMT as a *classification* task in an *autoregressive* setting and analyze the limitations of both classification and autoregression components. Balanced class distribution during training is known to improve the classifier performance, whereas imbalanced class distribution is known to induce unwelcome frequency-based biases on classes. Since the Zipfian nature of languages causes imbalanced classes, we explore its effect and find that frequency-based class biases exist in NMT models. Using our proposed abstraction of NMT as classifier and autoregressor components, we analyze the effect of various vocabulary sizes on the end performance of NMT on multiple languages with many data sizes, and reveal an explanation for *why* certain vocabulary sizes are better than others. Furthermore, recently, model-based MT metrics trained on segment-level human judgments have emerged as an attractive replacement to traditional corpus-level evaluation metrics. However, these metrics are costly, their decisions are inherently non-transparent and they appear to reflect unwelcome biases. We evaluate NMT using the well known classifier evaluation metrics that are easy to compute, transparent, and hyperparameter-free, and find that MacroF$_1$, a metric commonly used for evaluating classifiers on imbalanced test sets, is competitive on direct human assessment and outperforms others as a performance indicator of a downstream task. [1]

CCS Concepts: • **Computing methodologies → Natural language processing**; **Natural language generation**; **Machine translation**; **Machine learning**.

Additional Key Words and Phrases: Zipfian long tail, Rare phenomenon learning, Imbalanced learning

ACM Reference Format:
Thamme Gowda. 2020. Neural Machine Translation With Imbalanced Classes. 1, 1 (February 2020), 24 pages.

1 INTRODUCTION

Natural language processing (NLP) tasks such as sentiment analysis [32, 61] and spam detection are modeled as classification tasks, where instances are independently labeled. Tasks such as part-of-speech tagging [60] and named entity recognition [56] are examples of structured classification tasks, where instance classification is decomposed into a sequence of per-token contextualized labels. We can similarly cast neural machine translation (NMT), an example of a natural language generation (NLG) task, as a form of structured classification, where an instance label (a translation) is generated as a sequence of contextualized labels, here by an autoregressor (see Section 2).

Since the parameters of modern machine learning (ML) classification models are estimated from training data, whatever biases exist in the training data will affect model performance. Among those biases, *class imbalance* is a topic of our interest. Class imbalance is said to exist when one or more classes are not of approximately equal frequency in data. The effect of class imbalance

[*] This work has been published at *Findings of the Association for Computational Linguistics: EMNLP 2020* with title "Finding the optimal vocabulary size for neural machine translation."

[1] Tools, configurations, system outputs, and analyses are at https://github.com/thammegowda/005-nmt-imbalance and https://github.com/thammegowda/007-mt-eval-macro

Author's address: Thamme Gowda, tg@isi.edu, tnarayan@usc.edu, Dept. of Computer Science, University of Southern California, Los Angeles, CA, USA, Los Angeles, CA, 90089.

has been extensively studied in several domains where classifiers are used (see Section 8.3). With neural networks, the imbalanced learning problem is mostly targeted to computer vision tasks; NLP tasks are under-explored [21].

Word types in natural language models resemble a Zipfian distribution, i.e. in any natural language corpus, we observe that a type's rank is roughly inversely proportional to its frequency. Thus, a few types are extremely frequent, while most of the rest lie on the long tail of infrequency. Zipfian distributions cause two problems in classifier-based NLG systems:

(1) **Unseen Vocabulary:** Any hidden data set may contain types not seen in the finite set used for training. A sequence drawn from a Zipfian distribution is likely to have a large number of rare types, and these are likely to have not been seen in training.
(2) **Imbalanced Classes:** There are a few extremely frequent types and many infrequent types, causing an extreme imbalance. Such an imbalance, in other domains where classifiers are used, has been known to cause undesired biases and severe performance degradation [21].

The use of *subwords*, that is, decomposition of word types into pieces , such as the widely used Byte Pair Encoding (BPE) [48] addresses the open-ended vocabulary problem by ultimately allowing a word to be represented as a sequence of characters if necessary. BPE has a single hyperparameter named *merge operations* that governs the vocabulary size. The effect of this hyperparameter is not well understood. In practice, it is either chosen arbitrarily or via trial-and-error [46].

Regarding the problem of imbalanced classes, Steedman [53] states that "the machine learning techniques that we rely on are actually very bad at inducing systems for which the crucial information is in rare events." However, to the best of our knowledge, this problem has not yet been directly addressed in the NLG setting.

Model-based metrics for evaluating MT such as BLEURT [47], ESIM [33], and YiSi [28] have recently attracted attention due to their superior correlation with human judgments [31]. However, Bleu [39] remains the most widely used corpus-level MT metric. It correlates reasonably well with human judgments, and moreover is easy to understand and cheap to calculate, requiring only reference translations in the target language. By contrast, model-based metrics require tuning on thousands of examples of human evaluation for every new target language or domain [47]. Model-based metric scores are also opaque and can hide undesirable biases, as can be seen in Table 1.

Reference:	You must be a doctor.	
Hypothesis:	______ must be a doctor.	
	He	-0.735
	Joe	-0.975
	Sue	-1.043
	She	-1.100
Reference:	It is the greatest country in the world.	
Hypothesis:	______ is the greatest country in the world.	
	France	-0.022
	America	-0.060
	Russia	-0.161
	Canada	-0.309

Table 1. A demonstration of BLEURT's internal biases; model-free metrics like BLEU would consider each of the errors above to be equally wrong.

The source of model-based metrics' (e.g. BLEURT) correlative superiority over model-free metrics (e.g. BLEU) appears to be the former's ability to focus evaluation on *adequacy*, while the latter are overly focused on *fluency*. BLEU and most other generation metrics consider each output *token* equally. Since natural language is dominated by a few high-count types, an MT model that concentrates on getting its *if*s, *and*s and *but*s right will benefit from BLEU in the long run more than one that gets its *xylophone*s, *peripatetic*s, and *defenestrate*s right. Can we derive a metric with the discriminating power of BLEURT that does not share its bias or expense and is as interpretable as BLEU?

As it turns out, the metric may already exist and be in common use. The areas concerned with classification such as information extraction have long used both *micro averaging*, which treats each token equally, and *macro averaging*, which instead treats each *type* equally, when evaluating. The latter in particular is useful on imbalanced test sets to avoid results dominated by overly frequent types. In this work we take a classification-based approach to evaluating machine translation in order to obtain an easy-to-calculate metric that focuses on adequacy as much as BLEURT but does not have the expensive overhead, opacity, or bias of model-based methods.

The contributions of this paper are as follows: We offer a simplified abstraction of NMT architectures by re-envisioning them as two high-level components: a multi-class *classifier* and an *autoregressor* (Section 2). We describe some of the desired settings for the classifier (Section 2.1) and autoregressor (Section 2.2) components. In Section 2.3, we describe how vocabulary size choice relates to the desired settings for the two components. Our experimental setup is described in Section 3, followed by an analysis of results in Section 4 that offers an explanation with evidence for *why* some vocabulary sizes are better than others. Building upon our view of NMT as multi-class classifier, Section 5 describes evaluation of NMT using standard classifier evaluation metrics such as precision, recall, and F-measure. Section 6 uncovers the impact of class imbalance, particularly frequency based discrimination and how it affects precision and recall of classes.[2] Section 7 justifies macro-averaged F-measure as a legitimate MT evaluation metric.

2 CLASSIFIER BASED NLG

Machine translation is commonly defined as the task of transforming sequences from the form $x = x_1x_2x_3...x_m$ to $y = y_1y_2y_3...y_n$, where x is in source language X and y is in target language Y. There are many variations of NMT architectures (Section 8.1), however, all share the common objective of maximizing $\prod_{t=1}^{n} P(y_t|y_{<t}, x_{1:m})$ for pairs $(x_{1:m}, y_{1:n})$ sampled from a parallel dataset. NMT architectures are commonly viewed as encoder-decoder networks. We instead re-envision the NMT architecture as two higher level components: an autoregressor (R) and a token classifier (C), as shown in Figure 1.

Autoregressor R, [6] being the most complex component of the NMT model, has many implementations based on various neural network architectures: recurrent neural networks (RNN) such as long short-term memory (LSTM) and gated recurrent unit (GRU), convolutional neural networks (CNN), and Transformer (Section 8.1). At time step t, R transforms the input context $y_{<t}, x_{1:m}$ into hidden state vector $h_t = R(y_{<t}, x_{1:m})$.

Classifier C is the same across all architectures. It maps h_t to a distribution $P(y_j|h_t) \forall y_j \in V_Y$, where V_Y is the vocabulary of Y. In machine learning, input to classifiers such as C is generally described as features that are either hand-engineered or automatically extracted. In our high-level view of NMT architectures, R is a neural network that serves as an automatic feature extractor for C.

[2] In this work, 'type' and 'class' are used interchangeably.

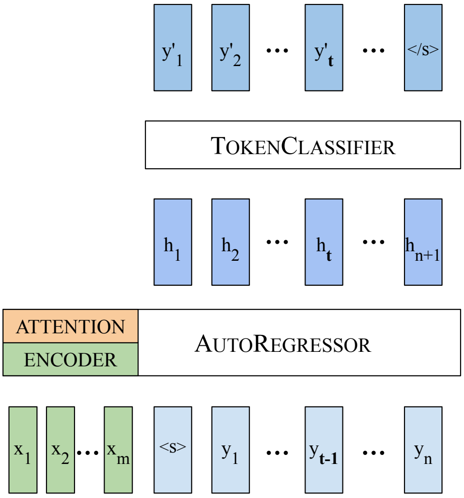


Fig. 1. The NMT model re-envisioned as a token classifier with an autoregressive feature extractor.

2.1 Balanced Classes for Token Classifier

Untreated, class imbalance leads to bias based on class frequencies. Specifically, classification learning algorithms focus on frequent classes while paying relatively less importance to infrequent classes. Frequency-based bias leads to poor recall of infrequent classes [21].

When a model is used in a *domain mismatch* scenario, i.e. where a test set's distribution does not match the training set's distribution, model performance generally degrades. It is not surprising that frequency-biased classifiers show particular degradation in domain mismatch scenarios, as types that were infrequent in the training distribution and were ignored by the learning algorithm may appear with high frequency in the new domain. Koehn and Knowles [24] showed empirical evidence of poor generalization of NMT to out-of-domain datasets.

In other classification tasks, where each instance is classified independently, methods such as up-sampling infrequent classes and down-sampling frequent classes are used. In NMT, since classification is done within the context of sequences, it is possible to accomplish the objective of balancing by altering sequence lengths. This can be done by choosing the level of subword segmentation [48].

Quantification of Zipfian Imbalance: We use two statistics to quantify the imbalance of a training distribution:

The first statistic relies on a measure of **Divergence** (D) from a balanced (uniform) distribution. We use a simplified version of Earth Mover Distance, in which the total cost for moving a probability mass between any two classes is the sum of the total mass moved. Since any mass moved *out of* one class is moved *into* another, we divide the total per-class mass moves in half to avoid double counting. Therefore, the imbalance measure D on K class distributions where p_i is the observed probability of class i in the training data is computed as:

$$D = \frac{1}{2}\sum_{i=1}^{K}|p_i - \frac{1}{K}|; \quad 0 \leq D \leq 1$$

A lower value of D is the desired setting for C, since the lower value results from a balanced class distribution. When classes are balanced, they have approximately equal frequencies; biasing the classes based on their frequencies is unlikely.

The second statistic is **Frequency at 95th% Class Rank ($\mathcal{F}_{95\%}$)**, defined as the least frequency in the 95^{th} percentile of most frequent classes. More generally, $\mathcal{F}_{P\%}$ is a simple way of quantifying the minimum number of training examples for at least the Pth percentile of classes. The bottom $(1-P)$ percentile of classes are overlooked to avoid the noise that is inherent in the real-world natural-language datasets.

A higher value for $\mathcal{F}_{95\%}$ is the desired setting for C, as a higher value indicates the presence of many training examples per class, and ML methods are known to perform better when there are many examples for each class.

2.2 Shorter Sequences for Autoregressor

Every autoregressive model is an approximation; some may be better than others, but no model is perfect. The total error accumulated grows in proportion to the length of the sequence. These accumulated errors alter the prediction of subsequent tokens in the sequence. Even though beam search attempts to mitigate this, it does not completely resolve it. These challenges with respect to long sentences and beam size are examined by Koehn and Knowles [24].

We summarize sequence lengths using **Mean Sequence Length**, μ, computed trivially as the arithmetic mean of the lengths of *target* language sequences after encoding them: $\mu = \frac{1}{N}\sum_{i=1}^{N}|y^{(i)}|$ where $y^{(i)}$ is the ith sequence in the training corpus of N sequences. Since shorter sequences have relatively fewer places where an imperfectly approximated autoregressor model can make errors, a smaller μ is a desired setting for R.

2.3 Choosing the Vocabulary Size Systematically

BPE [48] is a greedy iterative algorithm often used to segment a vocabulary into useful *subwords*. The algorithm starts with characters as its initial vocabulary. In each iteration, it greedily selects the most frequent type bigram in the training corpus, and replaces the sequence with a newly created compound type. Once the subword vocabulary is learned, it can be applied to a corpus by greedily segmenting words with the longest available subword type. These operations have an effect on D, $\mathcal{F}_{95\%}$, and μ.

Effect of BPE on μ: BPE expands rare words into two or more subwords, lengthening a sequence (and raising μ) relative to simple white-space segmentation. BPE merges frequent-character sequences into one subword piece, shortening a sequence (and lowering μ) relative to character segmentation. Hence, the sequence length of BPE segmentation lies in between the sequence lengths obtained by white-space and character-only segmentation methods [37].

Effect of BPE on $\mathcal{F}_{95\%}$ and D: Whether BPE is viewed as a merging of frequent subwords into a relatively less frequent compound, or a splitting of rare words into relatively frequent subwords, BPE alters the class distribution by moving the probability mass of classes. Hence, by altering the class distribution, BPE also alters both $\mathcal{F}_{95\%}$ and D. The BPE hyperparameter controls the amount of probability mass moved between subwords and compounds.

Figure 2 shows the relation between number of BPE merges (i.e. the BPE hyperparameter), and both D and μ. When few BPE merge operations are performed, we observe the lowest value of D, which is a desired setting for C, but at the same point μ is large and undesired for R (Section 2). When a large number of BPE merges are performed, the effect is reversed, i.e. we observe that D is large and unfavorable to C while μ is small and favorable to R. In the following sections we describe our experiments and analysis to locate the optimal number of BPE merges that achieves the right trade-off for both C and R.

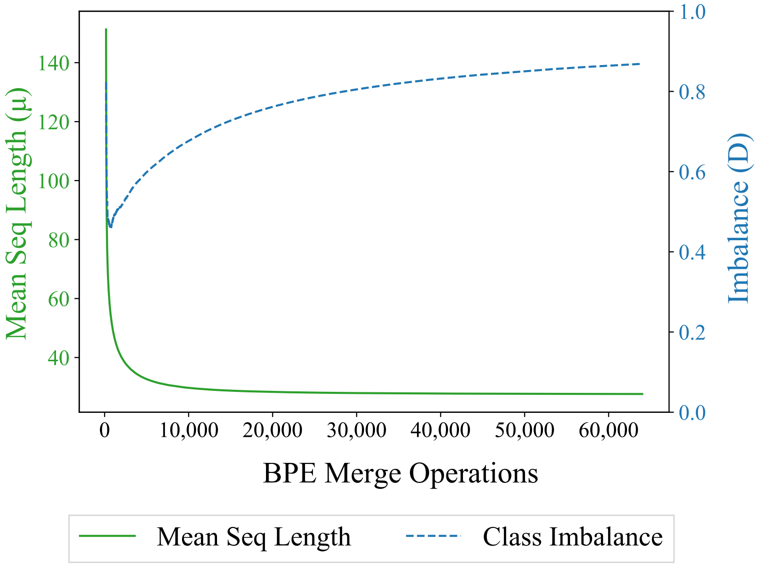


Fig. 2. Effect of BPE merge operations on mean sequence length (μ) and class imbalance (D).

3 EXPERIMENTAL SETUP

Our NMT experiments use the base Transformer model [57] on four different target languages at various training data sizes, described in the following subsections.

3.1 Datasets

We use the following four language pairs for our analysis: English→German, German→English, English→Hindi, and English→Lithuanian. To analyze the impact of different training data sizes, we randomly sub-select smaller training corpora for English↔German and English→Hindi languages. Statistics regarding the corpora used for validation, testing, and training are in Table 2. The datasets for English↔German, and English→Lithuanian are retrieved from the News Translation task of WMT2019 [3].[3] For English→Hindi, we use the IIT Bombay Hindi-English parallel corpus v1.5 [26]. English, German, and Lithuanian sentences are tokenized using SacreMoses.[4] Hindi sentences are tokenized using IndicNlpLibrary.[5] The training datasets are trivially cleaned: we exclude sentences with length in excess of five times the length of their parallel counterparts. Since the vocabulary is a crucial part of this analysis, we exclude all sentence pairs containing URLs.

3.2 Hyperparameters

Our model is a 6 layer Transformer encoder-decoder that has 8 attention heads, 512 hidden vector units, and a feed forward intermediate size of 2048, with GELU activation. We use label smoothing

[3] http://www.statmt.org/wmt19/translation-task.html
[4] github.com/alvations/sacremoses
[5] github.com/anoopkunchukuttan/indic_nlp_library

Languages	Training	Sentences	EN Toks	XX Toks	Validation	Test
DE→EN EN→DE	Europarl v10 WMT13CommonCrawl NewsCommentary v14	30K	0.8M	0.8M	NewsTest18	NewsTest19
		0.5M	12.9M	12.2M		
		1M	25.7M	24.3M		
		4.5M	116M	109.8M		
EN→HI	IITB Training	0.5M	8M	8.6M	IITB Dev	IITB Test
		1.3M	21M	22.5M		
EN→LT	Europarl v10	0.6M	17M	13.4M	NewsDev19	NewsTest19

Table 2. Training, validation, and testing datsets, along with sentence and token counts in training sets. We generally refer to dataset's sentence size in this work.

at 0.1, and a dropout rate of 0.1. We use the Adam optimizer [23] with a controlled learning rate that warms up for 16K steps followed by the decay rate recommended for training Transformer models [40]. To improve performance at different data sizes we set the mini-batch size to 6K tokens for the 30K-sentence datasets, 12K tokens for 0.5M-sentence datasets, and 24K for the remaining larger datasets [40]. All models are trained until no improvement in validation loss is observed, with a patience of 10 validations, each done at 1,000 update steps apart. Our model is implemented using PyTorch and run on NVIDIA P100 and V100 GPUs. To reduce padding tokens per batch, mini-batches are made of sentences having similar lengths [57]. We trim longer sequences to a maximum of 512 tokens after BPE. To decode, we average the last 10 checkpoints, and use a beam size of 4 with length penalty of 0.6, similar to Vaswani et al. [57].

Since the vocabulary size hyperparameter is the focus of this analysis, we use a range of vocabulary sizes that include character vocabulary and BPE operations that yield vocabulary sizes between 500 and 64K types. A common practice, as seen in Vaswani et al. [57]'s setup, is to jointly learn BPE for both source and target languages, which facilitates three-way weight sharing between the encoder's input, the decoder's input, and the output (i.e. classifier's class) embeddings [44]. However, to facilitate fine-grained analysis of vocabulary sizes and their effect on class imbalance, our models separately learn source and target vocabularies; weight sharing between the encoder's and decoder's embeddings is thus not possible. For the target language, however, we share weights between the decoder's input and the classifier's class embeddings.

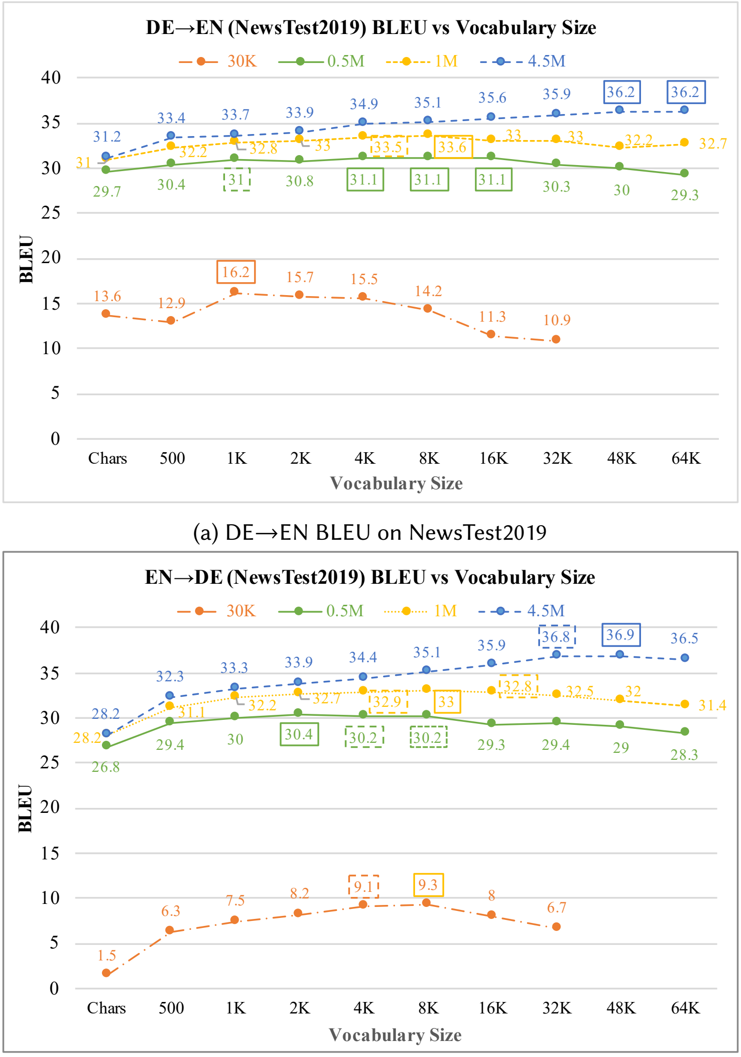
4 RESULTS AND ANALYSIS

BLEU scores for DE→EN and EN→DE experiments are reported in Figures 3a and 3b respectively. Results from EN→HI, and EN→LT are combined in Figure 4. All the reported BLEU scores are obtained using SacreBLEU [42].[6]

We make the following observations: smaller vocabulary such as characters have not produced the best BLEU for any of our language pairs or dataset sizes. A vocabulary of 32K or larger is unlikely to produce optimal results unless the data set is large e.g. the 4.5M DE↔EN sets. The BLEU curves as a function of vocabulary sizes have a shape resembling a hill. The position of the peak of the hill seems to shift towards a larger vocabulary when the datasets are large. However, there is a lot of variance in the position of the peak: one extreme is at 500 types on 0.5M EN→HI, and the other extreme is at 64K types in 4.5M DE→EN.

Although Figures 3 and 4 indicate *where* the optimal vocabulary size is for these chosen language pairs and datasets, the question of *why* the peak is where it is remains unanswered. We visualize μ, D, and $F_{95\%}$ in Figures 5 and 6 to answer that question, and report these observations:

[6] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.6



(a) DE⟶EN BLEU on NewsTest2019

(b) EN⟶DE BLEU on NewsTest2019

Fig. 3. EN↔DE NewsTest2019 BLEU as a function of vocabulary size at various training set sizes. Only the large dataset with 4.5M sentences has its best performance at a large vocabulary; all others peak at an 8K or smaller vocabulary size. Settings with the highest BLEU score are indicated by solid boundary and the ones within 0.2 BLEU to the highest BLEU are indicated by dashed boundary.

(1) Small vocabularies have a relatively larger $\mathcal{F}_{95\%}$ (favorable to classifier), yet they are suboptimal. We reason that this is due to the presence of a larger μ, which is unfavorable to the autoregressor.
(2) Larger vocabularies such as 32K and beyond have a smaller μ which favors the autoregressor, yet rarely achieved the best BLEU. We reason this is due to the presence of a lower $\mathcal{F}_{95\%}$ and a higher D being unfavorable to the classifier. Since the larger datasets have many training

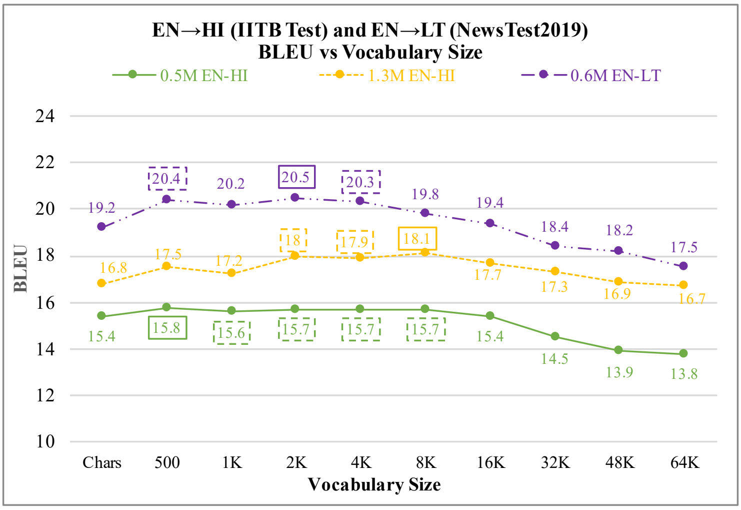


Fig. 4. BLEU on EN→HI IITB Test and EN→LT NewsTest2019 as a function of vocabulary size. Settings with the highest BLEU score are indicated by solid boundary and the ones within 0.2 BLEU to the highest BLEU are indicated by dashed boundary. These language pairs observed the best BLEU scores in the range of 500 to 8K vocabulary size.

examples for each class, as indicated by a generally larger $\mathcal{F}_{95\%}$, we conclude that bigger vocabularies tend to yield optimal results compared to smaller datasets in the same language.

(3) On small (30K) to medium (1.3M) data sizes, the vocabulary size of 8K seems to find a good trade-off between μ and D, as well as between μ and $\mathcal{F}_{95\%}$.

There is a *simple heuristic* to locate the peak: the near-optimal vocabulary size is where sentence length μ is small, while $\mathcal{F}_{95\%}$ is approximately 100 or higher. BLEU scores are often lower at larger vocabulary sizes—where μ is (favorably) low but D is (unfavorably) high (Figures 5 and 6). This calls for a further investigation that is reported in the following section.

5 EVALUATING MT AS CLASSIFICATION

Section 2 provides a high-level view of NMT as two fundamental ML components: an autoregressor and a classifier. Specifically, NMT is viewed as a multi-class classifier that operates on representations from an autoregressor. We may thus consider the well known evaluation metrics such as precision, recall, and F-measure.

Consider a test corpus, $T = \{(x^{(i)}, h^{(i)}, y^{(i)}) | i = 1, 2, 3...m\}$ where $x^{(i)}$, $h^{(i)}$, and $y^{(i)}$ are source, system hypothesis, and reference translation, respectively. Let $x = \{x^{(i)} \forall i\}$ and similar for h and y. Let V_h, V_y, $V_{h\cap y}$, and V be the vocabulary of h, the vocabulary of y, $V_h \cap V_y$, and $V_h \cup V_y$, respectively. For each class $c \in V$,

$$\textsc{Preds}(c) = \sum_{i=1}^{m} C(c, h^{(i)}); \qquad \textsc{Refs}(c) = \sum_{i=1}^{m} C(c, y^{(i)})$$

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

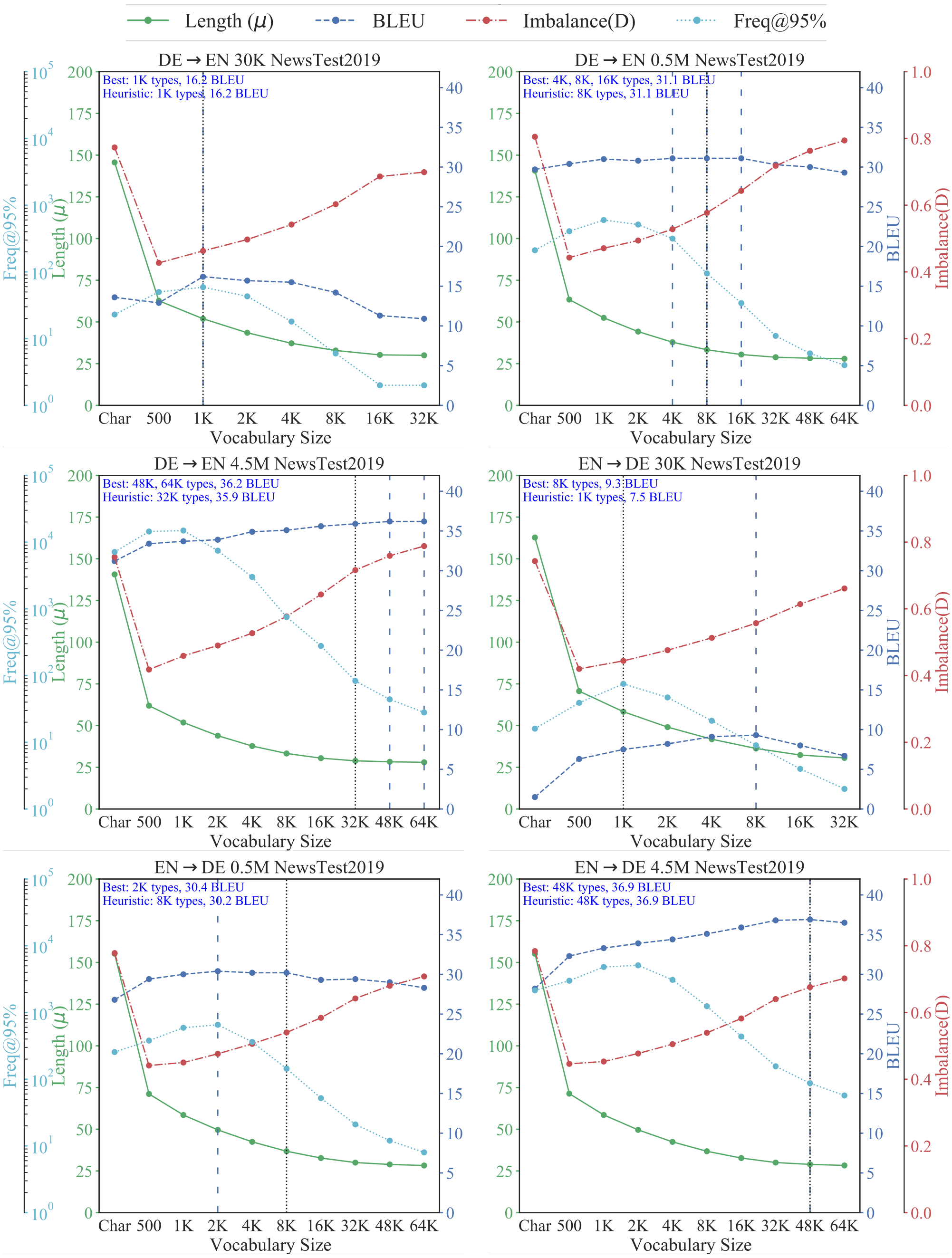


Fig. 5. Visualization of sequence length (μ) (lower is better), class imbalance (D) (lower is better), frequency of 95^{th} percentile class ($\mathcal{F}_{95\%}$) (higher is better; plotted in logarithmic scale), and test set BLEU (higher is better) on all language pairs and training data sizes. The vocabulary sizes that achieved highest BLEU are indicated with dashed vertical lines, and the vocabulary our heuristic selects is indicated by dotted vertical lines.

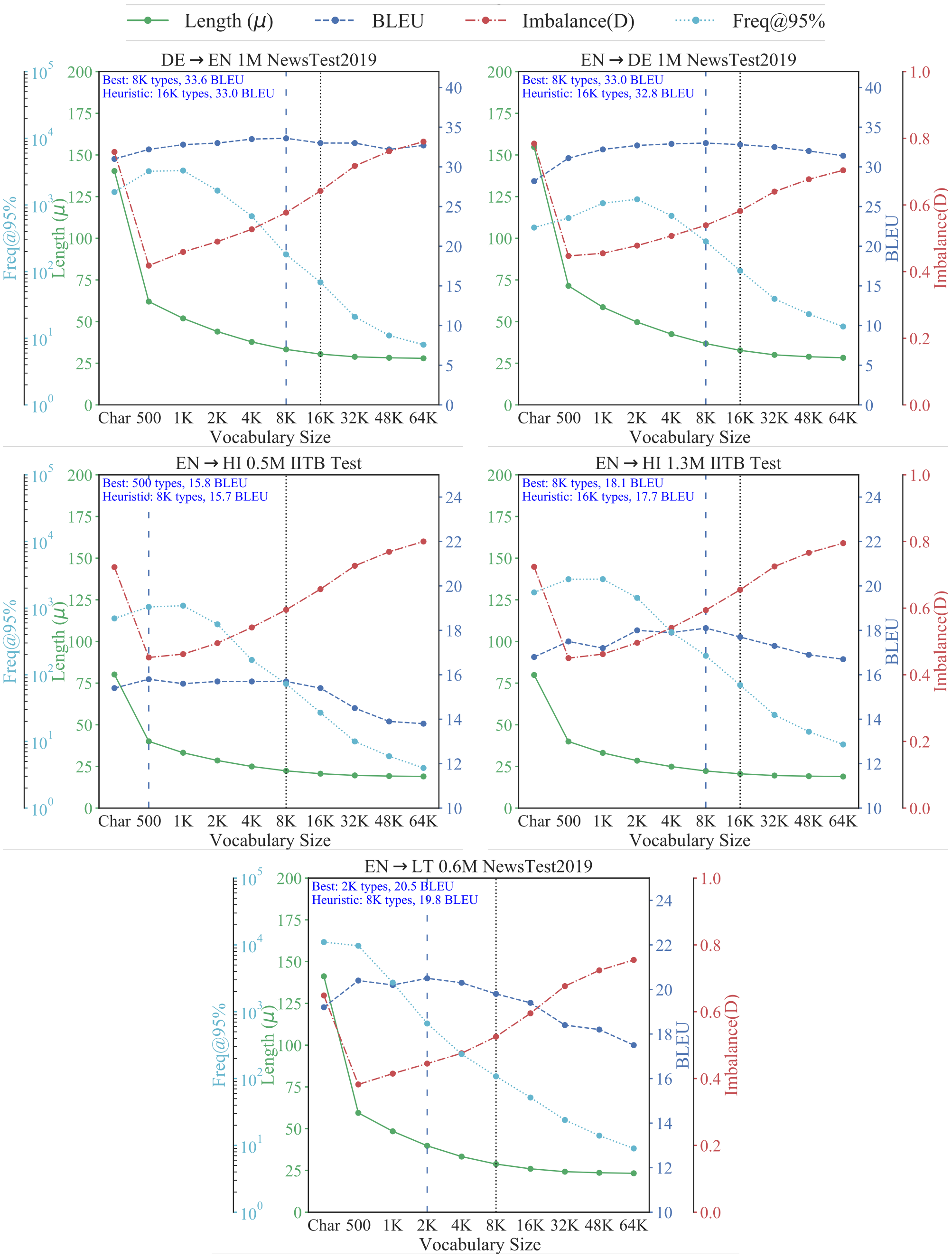


Fig. 6. Visualization of sequence length (μ) (lower is better), class imbalance (D) (lower is better), frequency of 95^{th} percentile class ($\mathcal{F}_{95\%}$) (higher is better; plotted in logarithmic scale), and test set BLEU (higher is better) on all language pairs and training data sizes. The vocabulary sizes that achieved highest BLEU are indicated with dashed vertical lines, and the vocabulary our heuristic selects is indicated by dotted vertical lines.

where $C(c, a)$ counts the number of tokens of type c in sequence a, similar to Papineni et al. [39]. For each class $c \in V_{h \cap y}$, precision (P_c), recall (R_c), and F_β measure ($F_{\beta;c}$) are computed as follows:[7]

$$P_c = \frac{\textsc{Match}(c)}{\textsc{Preds}(c)}$$

$$R_c = \frac{\textsc{Match}(c)}{\textsc{Refs}(c)}$$

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

The overall performance of a multi-classifier is obtained by averaging individual class performances. The two most popular averaging methods are: *macro-average*, which assigns equal importance to each type, and *micro-average*, which assigns equal importance to each token, as follows:

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

$$\textsc{MicroF}_\beta = \frac{\sum_{c \in V} f(c) \times F_{\beta;c}}{\sum_{c' \in V} f(c')}$$

where $f(c) = \textsc{Refs}(c) + k$ for smoothing factor $k > 0$.[8] We scale $\textsc{MacroF}_\beta$ and $\textsc{MicroF}_\beta$ values to percentile, similar to Bleu, for the sake of easier readability. Figure 7 provides a visualization of $\textsc{MacroF}_1$ and $\textsc{MicroF}_1$ as well as the popular alternatives such as Bleu and $\textsc{ChrF}_1$ in context.

6 MEASURING CLASSIFIER BIAS DUE TO IMBALANCE

In a typical classification setting with imbalanced classes, the classifier learns an undesired bias based on frequencies. A balanced class distribution debiases in this regard, leading to improvement in the precision of frequent classes as well as recall of infrequent classes. However, BLEU focuses only on the *precision* of classes; except for adding a global brevity penalty, it is ignorant of the poor recall of infrequent classes. Therefore, the BLEU scores shown in Figures 3a, 3b and 4 capture only a part of the improvements and biases. In this section we perform a detailed analysis of the impact of class balancing by considering both precision *and* recall of classes.

We accomplish this in two stages: First, we define a method to measure the bias of the model for classes based on their frequencies. Second, we track the bias in relation to vocabulary size and class imbalance, and report DE→EN and EN→DE, as these have many data points.

6.1 Frequency Based Bias

We measure frequency bias using the Pearson correlation coefficient, ρ, between class rank and class performance, where for performance measures we use precision and recall. Classes are ranked based on descending order of frequencies in the *training data* encoded with the same encoding schemes used for reported NMT experiments. With this setup, the class with rank 1, say $\mathcal{R}_1$, is the one with the highest frequency, rank 2 is the next highest, and so on. More generally, $\mathcal{R}_k$ is an index in the class rank list which has an inverse relation to class frequencies.

The Pearson correlation coefficients between class rank and precision ($\rho_{\mathcal{R},P}$), and class rank and recall ($\rho_{\mathcal{R},R}$) are reported in Figure 8. In datasets where D is high, the performance of classifier correlates with class rank. Such correlations are undesired for a classifier.

[7] We consider $F_{\beta;c}$ for $c \notin V_{h \cap y}$ to be 0.

[8] We use $k = 1$. If $k \to \infty$, then $\textsc{MicroF}_1 \to \textsc{MacroF}_1$.

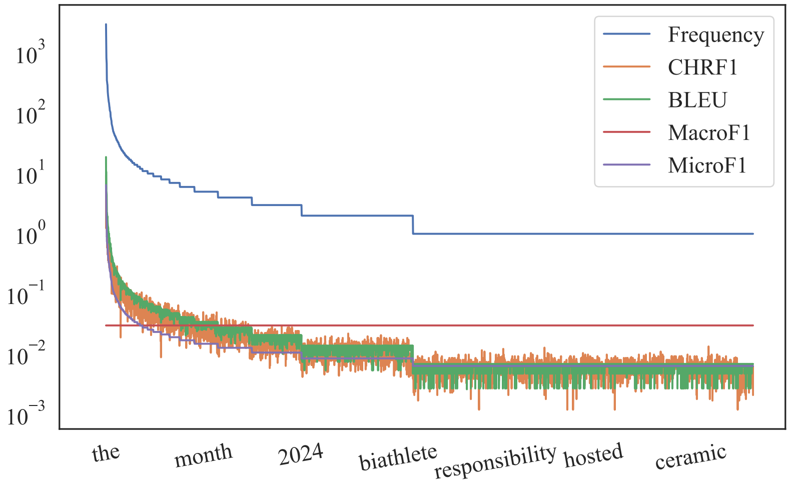


Fig. 7. MT metrics and their weight distribution across vocabulary as measured on WMT NewsTest2019 De-En corpus. The horizontal axis contains word types ranked by frequencies. The vertical axis, in logarithmic scale, represents type frequencies (raw count) and percentile of contribution to the total corpus-level score. The weight of a type is the loss in corpus-level MT score when each token of the type is replaced by an out-of-vocabulary token of similar character count. The roughness in Bleu and ChrF$_1$ lines is due to the variation in n-grams affected by a type. MicroF$_1$ uses only unigrams and scales the contribution by its frequency, whereas MacroF$_1$ is unweighted and assigns equal importance to all the types regardless of their frequencies. It is evident that the type contribution in Bleu and ChrF$_1$ is scaled according to frequencies in the similar manner as MicroF$_1$. For instance, 'the' type appears 3019 times in the chosen test corpus, and zero-recall of 'the' results in a loss up to 18.83%, 9.38%, 6.45%, and 0.03% of overall score in Bleu, ChrF$_1$, MicroF$_1$, and MacroF$_1$, respectively.

6.2 Analysis of Class Frequency Bias

An ideal classifier is one that does not discriminate classes based on their frequencies, i.e. one that exhibits no correlation between $\rho_{\mathcal{R},P}$, and $\mathbf{r}_{\mathcal{R},R}$. However, we see in Figure 8 that:

(1) $\rho_{\mathcal{R},P}$ is positive when the dataset has high D; i.e if the class rank increases (frequency decreases), precision increases in relation to it. This indicates that frequent classes have relatively less precision than infrequent classes. The bias is strongly positive on smaller datasets such as 30K DE→EN, which gradually diminishes if the training data size is increased or a vocabulary setting is chosen to reduce D.
(2) $\rho_{\mathcal{R},R}$ is negative, i.e., if the class rank increases, recall decreases in relation to it. This is an indication that infrequent classes have relatively lower recall than frequent classes.

Figure 8 shows a trend that frequency based bias measured by correlation coefficient is lower in settings that have lower D. However, since D is non-zero, there still exists non-zero correlation between recall and class rank ($\rho_{\mathcal{R},R}$), indicating the poorer recall of low-frequency classes.

7 JUSTIFICATION FOR EVALUATING MT AS CLASSIFIER

In the following sections, we verify and justify the utility of MacroF$_1$ while also offering a comparison with popular alternatives such as MicroF$_1$, Bleu [39], ChrF$_1$ [41], BLEURT [47]. Bleu and ChrF$_1$ scores are computed using SacreBleu [42]. MacroF$_1$ and MicroF$_1$ use the same tokenizer as Bleu. Since BLEURT is a segment-level measure, we consider both BLEURTmean and

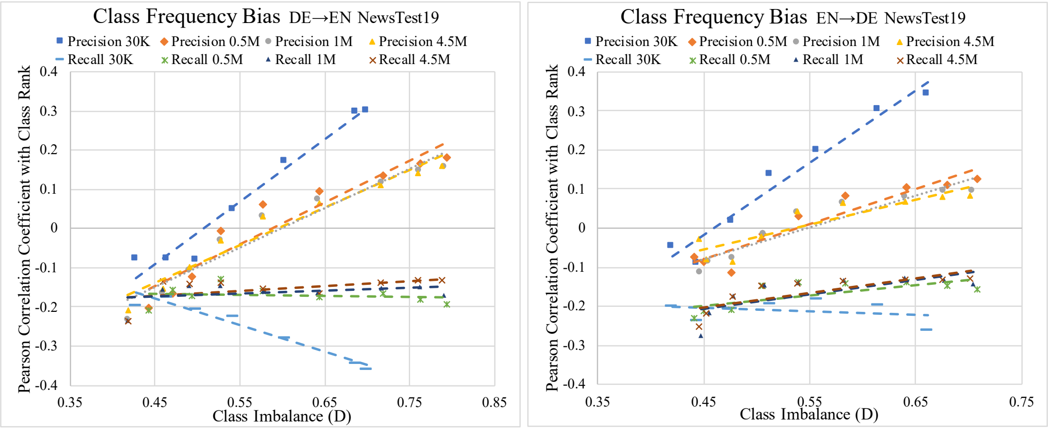


Fig. 8. Correlation analysis on DE→EN and EN→DE show that NMT models suffer from frequency based class bias, indicated by non-zero correlation of both precision and recall with class rank. Reduction in class imbalance (D), as shown by the horizontal axis, generally reduces the bias as indicated by the reduction in magnitude of correlation.

BLEURTmedian, which are mean and median of segment-level scores, as its corpus-level measures. We use Kendall's rank correlation coefficient, τ, to compute the association between metrics and human judgments. Correlations with p-values smaller than $\alpha = 0.05$ are considered to be statistically significant.

7.1 Data-to-Text: WebNLG

Name	Fluency & Grammar	Semantics
Bleu	$^{\times}$.444	$^{\times}$.500
ChrF$_1$	$^{\times}$.278	.778
MacroF$_1$	$^{\times}$.222	.722
MicroF$_1$	$^{\times}$.333	.611
BLEURTmean	$^{\times}$.444	.833
BLEURTmedian	.611	.667

Table 3. WebNLG data-to-text task: Kendall's τ between system-level MT metric scores and human judgments. Fluency and grammar are correlated identically by all metrics. Values that are *not* significant at $\alpha = 0.05$ are indicated by $^{\times}$.

We use the 2017 WebNLG Challenge dataset [15, 50][9] to analyze the differences between micro- and macro- averaging. WebNLG is a task of generating English text for sets of triples extracted from DBPedia. Human annotations are available for a sample of 223 records each from nine NLG systems. The human judgments provided have three linguistic aspects—fluency, grammar, and semantics[10]—which enable us to perform a fine grained analysis of our metrics. We compute Kendall's τ between metrics and human judgments, which are reported in Table 3.

As seen in Table 3, the metrics exhibit much variance in agreements with human judgments. For instance, BLEURTmedian is the best indicator of fluency and grammar, however BLEURTmean

[9] https://gitlab.com/webnlg/webnlg-human-evaluation

[10] Fluency and grammar, which are elicited with nearly identical directions [15], are identically correlated.

is best on semantics. BLEURT, being a *model-based* measure that is directly trained on human judgments, scores relatively higher than others. Considering the model-free metrics, chrf1 does well on semantics but poorly on fluency and grammar compared to Bleu. Not surprisingly, both MicroF$_1$ and MacroF$_1$, which rely solely on unigrams, are poor indicators of fluency and grammar compared to Bleu, however MacroF$_1$ is clearly a better indicator of semantics than Bleu. The discrepancy between MicroF$_1$ and MacroF$_1$ regarding their agreement with fluency, grammar, and semantics is expected: micro-averaging pays more attention to function words (as they are frequent types) that contribute to fluency and grammar whereas macro-averaging pays relatively more attention to the content words that contribute to semantic adequacy.

The take away from this analysis is as follows: MacroF$_1$ is a strong indicator of semantic adequacy, however, it is a poor indicator of fluency. We recommend using either MacroF$_1$ or ChrF$_1$ when semantic adequacy and not fluency is a desired goal.

7.2 Machine Translation: WMT Metrics

Year	Pairs		★Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
2019	18	Mean	.751	.771	.821	.818	.841
		Median	.782	.752	.844	.844	.875
		Wins	3	3	**6**	3	5
2018	14	Mean	.858	.857	.875	.873	.902
		Median	.868	.868	.901	.879	.919
		Wins	1	2	3	2	**6**
2017	13	Mean	.752	.713	.714	.742	.804
		Median	.758	.733	.735	.728	.791
		Wins	5	4	2	2	**6**

Table 4. WMT 2017–19 Metrics task: Mean and median Kendall's τ between MT metrics and human judgments. Correlations that are not significant at $\alpha = 0.05$ are excluded from the calculation of mean, and median, and wins. See Appendix Tables 6, 7, and 8 for full details. ★Bleu is pre-computed scores available in the metrics packages. In 2018 and 2019, both MacroF$_1$ and MicroF$_1$ outperform Bleu, MacroF$_1$ outperforms MicroF$_1$. ChrF$_1$ has strongest mean and median agreements across the years. Judging based on the number of wins, MacroF$_1$ has steady progress over the years, and outperforms others in 2019.

In this section, we verify how well the metrics agree with human judgments using Workshop on Machine Translation (WMT) metrics task datasets for 2017–2019 [4, 30, 31].[11] We first compute scores from each MT metric, and then calculate the correlation τ with human judgments.

As there are many language pairs and translation directions in each year, we report only the mean and median of τ, and number of wins per metric for each year in Table 4. We have excluded BLEURT from comparison in this section since the BLEURT models are fine-tuned on the same datasets on which we are evaluating the other methods.[12] ChrF$_1$ has the strongest mean and median agreement with human judgments across the years. In 2018 and 2019, both MacroF$_1$ and MicroF$_1$ mean and median agreements outperform Bleu whereas in 2017 Bleu was better than MacroF$_1$ and MicroF$_1$.

As seen in Section 7.1, MacroF$_1$ weighs towards semantics whereas MicroF$_1$ and Bleu weigh towards fluency and grammar. This indicates that recent MT systems are mostly fluent, and adequacy is the key discriminating factor amongst them. Bleu served well in the early era of

[11] http://www.statmt.org/wmt19/metrics-task.html
[12] https://github.com/google-research/bleurt

statistical MT when fluency was a harder objective. Recent advancements in neural MT models such as Transformers [57] produce fluent outputs, and have brought us to an era where semantic adequacy is the focus.

7.3 Cross-Lingual Information Retrieval

In this section, we determine correlation between MT metrics and downstream cross-lingual information retrieval (CLIR) tasks. CLIR is a kind of information retrieval (IR) task in which documents in one language are retrieved given queries in another [17]. A practical solution to CLIR is to translate source documents into the query language using an MT model, then use a monolingual IR system to match queries with translated documents. Correlation between MT and IR metrics is accomplished in the following steps:

(1) Build a set of MT models and measure their performance using MT metrics.
(2) Using each MT model in the set, translate all source documents to the target language, build an IR model, and measure IR performance on translated documents.
(3) For each MT metric, find the correlation between the set of MT scores and their corresponding set of IR scores. The MT metric that has a stronger correlation with the IR metric(s) is more useful than the ones with weaker correlations.
(4) Repeat the above steps on many languages to verify the generalizability of findings.

An essential resource of this analysis is a dataset with human annotations for computing MT and IR performances.

	Domain	IR Score	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$	BLEURTmean	BLEURTmedian
LT-EN	In	AQWV	.429	$^{\times}$.363	**.508**	$^{\times}$.385	.451	.420
		MAP	.495	.429	**.575**	.451	.473	.486
	In+Ext	AQWV	$^{\times}$.345	**.527**	.491	.491	.491	.477
		MAP	$^{\times}$.273	$^{\times}$**.455**	$^{\times}$.418	$^{\times}$.418	$^{\times}$.418	$^{\times}$.404
PS-EN	In	AQWV	.559	**.653**	.574	.581	.584	.581
		MAP	.493	**.632**	.487	.494	.558	.554
	In+Ext	AQWV	.589	**.682**	.593	.583	.581	.571
		MAP	.519	**.637**	.523	.482	.536	.526
BG-EN	In	AQWV	$^{\times}$.455	**.550**	.527	$^{\times}$.382	$^{\times}$.418	.418
		MAP	.491	**.661**	.564	.491	.527	.527
	In+ext	AQWV	$^{\times}$.257	**.500**	$^{\times}$.330	$^{\times}$.404	$^{\times}$.367	$^{\times}$.367
		MAP	$^{\times}$.183	$^{\times}$**.426**	$^{\times}$.257	$^{\times}$.330	$^{\times}$.294	$^{\times}$.294

Table 5. CLSSTS CLIR task: Kendall's τ between IR and MT metrics under study. The rows with Domain=In are where MT and IR scores are computed on the same set of documents, whereas Domain=In+Ext are where IR scores are computed on a larger set of documents that is a superset of segments on which MT scores are computed. **Bold** values are the best correlations achieved in a row-wise setting; values with $^{\times}$ are *not* significant at $\alpha = 0.05$.

An essential resource of this analysis is a dataset with human annotations for computing MT and IR performances. We conduct experiments on data from the 2020 workshop on *Cross-Language Search and Summarization of Text and Speech* (CLSSTS) [59]. CLSSTS datasets contain queries in English (EN), and documents in many source languages along with their human translations, as well as query-document relevance judgments. We use three source languages: Lithuanian (LT), Pashto (PS), and Bulgarian (BG). The performance of this CLIR task is evaluated using two IR measures: Actual Query Weighted Value (AQWV) and Mean Average Precision (MAP). AQWV[13] is derived from Actual Term Weighted Value (ATWV) metric [58].

[13] https://www.nist.gov/system/files/documents-/2017/10/26/aqwv_derivation.pdf

Our CLIR system is based on Boschee et al. [5], which is competitive in the workshop on CLSSTS-2020 [36]. Since the CLIR system is also treated as a blackbox, the internals of CLIR is unnecessary and beyond the scope of this work. Kendall's τ between MT and IR measures, given in Table 5, show that MacroF$_1$ is the strongest indicator of CLIR downstream task performance in five out of six settings. AQWV and MAP have a similar trend in agreement to the MT metrics. ChrF$_1$ and BLEURT, which are strong contenders when generated text is directly evaluated by humans, do not indicate CLIR task performance as well as MacroF$_1$, as CLIR tasks require faithful meaning equivalence across the language boundary, and human translators can mistake fluent output for proper translations [8].

8 RELATED WORK

8.1 NMT Architectures

Several variations of NMT models have been proposed and refined: Sutskever et al. [55] and Cho et al. [11] introduce the RNN-based encoder-decoder model. Bahdanau et al. [1] introduce the attention mechanism and Luong et al. [29] propose several variations that became essential components of many future models. RNN modules, either LSTM [19] or GRU [10], have been popular choices for composing NMT encoders and decoders. The encoder uses bidirectional information, but the decoder is unidirectional, typically left-to-right, to facilitate autoregressive generation. Gehring et al. [16] use a CNN architecture that outperforms RNN models. Vaswani et al. [57] propose the **Transformer**, whose main components are feed-forward and attention networks. Our experiments and analysis are based on Transformer NMT architecture, however our proposed abstraction of NMT as a combination of Classifier and an Autoregressor is applicable to other autoregressive NMT architectures that have the objective of maximizing $P(y_t|y_{<t}, x_{1:m})$. There are only a few models that perform non-autoregressive NMT [18, 27]. These are focused on improving the speed of inference; generation quality is currently sub-par compared to autoregressive models. These non-autoregressive models can also be viewed as token classifiers with a different kind of feature extractor, whose strengths and limitations are yet to be theoretically understood. Analyzing the non-autoregressive component, especially its performance with longer sequences, is beyond the scope of this work.

8.2 BPE Subwords

Sennrich et al. [48] introduce BPE as a simplified way to solve out-of-vocabulary (OOV) words without having to use a back-off dictionary for OOV words. They note that BPE improves the translation of not only the OOV words, but also some rare in-vocabulary words. The analysis by Morishita et al. [37] is different than ours in that they view various vocabulary sizes as hierarchical features that are used in addition to a fixed vocabulary. Salesky et al. [46] offer an efficient way to search BPE vocabulary size for NMT. Kudo [25] use BPE as a regularization technique by introducing sampling based randomness to the BPE segmentation. To the best of our knowledge, no previous work exists that analyzes BPE's effect on class imbalance.

8.3 Class Imbalance

The class imbalance problem has been extensively studied in classical machine learning [20]. In the medical domain, Mazurowski et al. [35] find that classifier performance deteriorates with even modest imbalance in the training data. Untreated class imbalance has been known to deteriorate the performance of image segmentation. Sudre et al. [54] investigate the sensitivity of various loss functions. Johnson and Khoshgoftaar [21] survey imbalance learning and report that the effort is mostly targeted to computer vision tasks. Buda et al. [7] provide a definition and quantification method for two types of class imbalance: *step imbalance* and *linear imbalance*. Since the imbalance in Zipfian distribution of classes is neither single-stepped nor linear, we use a divergence based measure to quantify the imbalance.

8.4 MT Metrics

Many metrics have been proposed for MT evaluation, which we broadly categorize into *model-free* or *model-based*. Model-free metrics compute scores based on translations but have no significant parameters or hyperparameters that must be tuned *a priori*; these include BLEU [39], NIST [14], TER [51], and CHRF$_1$ [41]. Model-based metrics have a significant number of parameters and, sometimes, external resources that must be set prior to use. These include METEOR [2], BLEURT [47], YiSi [28], ESIM [33], and BEER [52]. Model-based metrics require significant effort and resources when adapting to a new language or domain, while model-free metrics require only a test set with references. Mathur et al. [34] have recently evaluated the utility of popular metrics and recommend the use of either CHRF$_1$ or a model-based metric instead of BLEU.

8.5 Rare Words are Important

That natural language word types roughly follow a Zipfian distribution is a well known phenomenon [43, 62]. The frequent types are mainly so-called "stop words," function words, and other low-information types, while most content words are infrequent types. Even though frequent types occur several orders of magnitude more frequently than others, they carry relatively less information [49]. To counter this natural frequency-based imbalance, statistics such as inverted document frequency (IDF) are commonly used to weigh the *input* words in applications such as information retrieval [22]. In NLG tasks such as MT, where words are the *output* of a classifier, there has been scant effort to address the imbalance. Doddington [14] is the only work we know of in which the 'information' of an n-gram is used as its weight, such that rare n-grams attain relatively more importance than in BLEU. We abandon this direction for two reasons: Firstly, as noted in that work, *large amounts of data are required to estimate n-gram statistics*. Secondly, unequal weighing is a bias that is best suited to datasets where the weights are derived from, and such biases often do not generalize to other datasets. Therefore, unlike Doddington [14], we assign equal weights to all n-gram classes, and in this work we limit our scope to unigrams only.

While BLEU is a precision oriented measure, METEOR [2] and CHRF [41] include both precision and recall similar to our methods. However, neither of these measures try to address the natural imbalance of class distribution. BEER [52] and METEOR [12] make an explicit distinction between function and content words; such a distinction inherently captures the frequency differences since the function words are often frequent and content words are often infrequent types. However, doing so requires the construction of potentially expensive linguistic resources. This work does not make any explicit distinction and uses naturally occurring type counts to effect a similar result.

8.6 F-measure as an Evaluation Metric

F-measure [9, 45] is extensively used as an evaluation metric in classification tasks such as part-of-speech tagging, information extraction, named entity recognition, and sentiment analysis [13]. Viewing MT as a multi-class classifier and evaluating MT solely as a multi-class classifier as proposed in this work is not an established practice. However, F_1 measure is sometimes used for various analyses when BLEU and others are inadequate: The compare-mt tool [38] supports comparison of MT models based on F_1 measure of individual types. Sennrich et al. [48] use corpus-level *unigram* F_1 in addition to BLEU and CHRF, however, corpus-level F_1 is computed as MICROF$_1$. To the best of our knowledge, there is no previous work that clearly formulates the differences between micro- and macro- averages, and justifies the use of MACROF$_1$ for MT evaluation.

9 DISCUSSION AND CONCLUSION

Envisioning NMT as a multi-class classifier with an autoregressive feature extractor has opened several possibilities: Firstly, it has enabled us to apply the findings from classification and autoregression modeling literature to understand the strengths and weaknesses of NMT modeling. And secondly, it has enabled us to use the standard classifier evaluation metrics for MT evaluation.

Our analysis of classifier and autoregression components provides an explanation of *why* text generation using BPE vocabulary is more effective than word and character vocabularies, and *why* some BPE hyperparameters are better than others. We show that the number of BPE merges is not an arbitrary hyperparameter and that it can be tuned to address the class imbalance and sequence length problems. Our recommendation for Transformer NMT is to *use the largest possible BPE vocabulary such that at least 95% of classes have 100 or more examples in training*.

We have evaluated NLG in general and MT specifically as a multi-class classifier, and illustrated the differences between micro- and macro- averages using MicroF$_1$ and MacroF$_1$ as examples. MacroF$_1$ captures semantic adequacy better than MicroF$_1$. We have found that another popular metric, ChrF$_1$, also performs well on direct assessment, however, being an implicitly micro-averaged measure, it does not perform as well as MacroF$_1$ on downstream CLIR tasks. Unlike BLEURT, which is also adequacy-oriented, MacroF$_1$ is directly interpretable, does not require retuning on expensive human evaluations when changing language or domain, and does not appear to have uncontrollable biases resulting from data effect. It is both easy to understand and to calculate, and is inspectable, enabling fine-grained analysis at the level of individual word types. These attributes make it a useful metric for understanding and addressing the flaws of current models.

Even though some BPE vocabulary sizes indirectly reduce the class imbalance, they do not eliminate it. Even after applying BPE segmentation, the class distributions remain to possess sufficient imbalance that induces a frequency-based bias and affects the recall of rare classes. Hence more effort is needed for directly dealing with the Zipfian imbalance that is inevitable in all MT datasets. Using MacroF$_1$ as an NLG evaluation metric is our first step in acknowledging the importance of the long tail of language systems; we anticipate the development of more advanced macro-averaged metrics that take advantage of higher-order and character n-grams in the future.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). International Conference on Learning Representations, San Diego, CA, USA. http://arxiv.org/abs/1409.0473

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://www.aclweb.org/anthology/W05-0909

[3] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 1–61. http://www.aclweb.org/anthology/W19-5301

[4] Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, 489–513. https://doi.org/10.18653/v1/W17-4755

[5] Elizabeth Boschee, Joel Barry, Jayadev Billa, Marjorie Freedman, Thamme Gowda, Constantine Lignos, Chester Palen-Michel, Michael Pust, Banriskhem Kayang Khonglah, Srikanth Madikeri, Jonathan May, and Scott Miller. 2019. SARAL: A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 19–24. https://doi.org/10.18653/v1/P19-3004

[6] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control.* John Wiley & Sons.

[7] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249 – 259. https://doi.org/10.1016/j.neunet.2018.07.011

[8] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation.* Association for Computational Linguistics, Prague, Czech Republic, 136–158. https://www.aclweb.org/anthology/W07-0718

[9] Nancy Chinchor. 1992. MUC-4 Evaluation Metrics. In *Proceedings of the 4th Conference on Message Understanding (MUC4 '92).* Association for Computational Linguistics, USA, 22–29. https://doi.org/10.3115/1072064.1072067

[10] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.* Association for Computational Linguistics, Doha, Qatar, 103–111. https://doi.org/10.3115/v1/W14-4012

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[12] Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation.* Association for Computational Linguistics, Edinburgh, Scotland, 85–91. https://www.aclweb.org/anthology/W11-2107

[13] Leon Derczynski. 2016. Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA), Portorož, Slovenia, 261–266. https://www.aclweb.org/anthology/L16-1040

[14] George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT '02).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 138–145.

[15] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 179–188. https://doi.org/10.18653/v1/P17-1017

[16] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR.org, 1243–1252.

[17] Gregory Grefenstette. 2012. *Cross-language information retrieval.* Vol. 2. Springer Science & Business Media.

[18] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-Autoregressive Neural Machine Translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.* OpenReview.net. https://openreview.net/forum?id=B1l8BtlCb

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6, 5 (2002), 429–449.

[21] Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (19 Mar 2019), 27. https://doi.org/10.1186/s40537-019-0192-5

[22] Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1972), 11–21.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[24] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation.* Association for Computational Linguistics, Vancouver, 28–39. https://doi.org/10.18653/v1/W17-3204

[25] Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Melbourne, Australia, 66–75. https://doi.org/10.18653/v1/P18-1007

[26] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1548

[27] Jindřich Libovický and Jindřich Helcl. 2018. End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing. Association for Computational Linguistics, Brussels, Belgium, 3016–3021. https://doi.org/10.18653/v1/D18-1336

[28] Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 507–513. https://doi.org/10.18653/v1/W19-5358

[29] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1412–1421. https://doi.org/10.18653/v1/D15-1166

[30] Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 671–688. https://doi.org/10.18653/v1/W18-6450

[31] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 62–90. http://www.aclweb.org/anthology/W19-5302

[32] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. https://www.aclweb.org/anthology/P11-1015

[33] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2799–2808. https://doi.org/10.18653/v1/P19-1269

[34] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4984–4997. https://www.aclweb.org/anthology/2020.acl-main.448

[35] Maciej A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi. 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks* 21, 2 (2008), 427 – 436. https://doi.org/10.1016/j.neunet.2007.12.031 Advances in Neural Networks Research: IJCNN '07.

[36] Kathy McKeown, Douglas W. Oard, Elizabeth, and Richard Schwartz (Eds.). 2020. *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France. https://www.aclweb.org/anthology/2020.clssts-1.0

[37] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2018. Improving Neural Machine Translation by Incorporating Hierarchical Subword Features. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 618–629. https://www.aclweb.org/anthology/C18-1052

[38] Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A Tool for Holistic Comparison of Language Generation Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 35–41. https://doi.org/10.18653/v1/N19-4007

[39] 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

[40] Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics* 110, 1 (2018), 43–70.

[41] 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

[42] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://www.aclweb.org/anthology/W18-6319

[43] David M. W. Powers. 1998. Applications and Explanations of Zipf's Law. In *New Methods in Language Processing and Computational Natural Language Learning*. https://www.aclweb.org/anthology/W98-1218

[44] Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 157–163. https://www.aclweb.org/anthology/E17-2025

[45] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann, USA.

[46] Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2018. Optimizing Segmentation Granularity for Neural Machine Translation. *CoRR* abs/1810.08641 (2018). arXiv:1810.08641 http://arxiv.org/abs/1810.08641

[47] 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

[48] 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

[49] C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[50] Anastasia Shimorina. 2018. Human vs Automatic Metrics: on the Importance of Correlation Design. *CoRR* abs/1805.11474 (2018). arXiv:1805.11474 http://arxiv.org/abs/1805.11474

[51] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, Vol. 200. Cambridge, MA.

[52] Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, 414–419. https://doi.org/10.3115/v1/W14-3354

[53] 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

[54] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu (Eds.). Springer International Publishing, Cham, 240–248.

[55] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[56] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://www.aclweb.org/anthology/W03-0419

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[58] Steven Wegmann, Arlo Faria, Adam Janin, Korbinian Riedhammer, and Nelson Morgan. 2013. The TAO of ATWV: Probing the mysteries of keyword search performance. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 192–197.

[59] 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

[60] Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, 1–19. http://www.aclweb.org/anthology/K/K17/K17-3001.pdf

[61] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 649–657. http://dl.acm.org/citation.cfm?id=2969239.2969312

[62] George Kingsley Zipf. 1949. Human behaviour and the principle of least-effort. Cambridge MA edn. *Addison-Wesley* (1949).

A AGREEMENT WITH WMT HUMAN JUDGMENTS

Tables 6, 7, and 8 provide τ between MT metrics and human judgments on WMT Metrics task 2017–2019. ★Bleu is based on pre-computed scores in WMT metrics package, whereas Bleu is based on our recalculation using SacreBleu. Values marked with $^\times$are not significant at $\alpha = 0.05$, and hence corresponding rows are excluded from the calculation of mean, median, and standard deviation.

Since MacroF$_1$ is the only metric that does not achieve statistical significance in the WMT 2019 EN-ZH setting, we carefully inspected it. Human scores for this setting are obtained without looking at the references by bilingual speakers [31], but the ZH references are found to have a large number of bracketed EN phrases, especially proper nouns that are rare types. When the text inside these brackets is not generated by an MT system, MacroF$_1$ naturally penalizes heavily due to the poor recall. Since other metrics assign lower importance to poor recall of such rare types, they achieve relatively better correlation to human scores than MacroF$_1$. However, since the τ values for EN-ZH are relatively lower than the other language pairs, we conclude that poor correlation of MacroF$_1$ in EN-ZH is due to poor quality references. Some settings did not achieve statistical significance due to a smaller sample set as there were fewer MT systems submitted, e.g. 2017 CS-EN.

	★Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
DE-CS	.855	.745	.964	.917	**.982**
DE-EN	.571	.655	.723	.695	**.742**
DE-FR	.782	.881	**.927**	.844	.915
EN-CS	.709	**.954**	.927	.927	.908
EN-DE	.540	.752	.741	.773	**.824**
EN-FI	.879	.818	.879	.848	**.923**
EN-GU	.709	.709	.600	**.734**	.709
EN-KK	.491	.527	**.685**	.636	.661
EN-LT	.879	.848	**.970**	.939	.881
EN-RU	.870	.848	**.939**	.879	.930
FI-EN	.788	.809	**.909**	.901	.875
FR-DE	**.822**	.733	.733	.764	.815
GU-EN	.782	.709	.855	.891	**.945**
KK-EN	**.891**	.844	.796	.844	.881
LT-EN	.818	**.855**	.844	**.855**	.833
RU-EN	.692	.729	.714	**.780**	.757
ZH-EN	.695	.695	**.752**	.676	.715
Median	.782	.752	.844	.844	.875
Mean	.751	.771	.821	.818	.841
SD	.124	.101	.112	.093	.095
EN-ZH	**.606**	**.606**	$^\times$.424	.595	.594
Wins	3	3	6	3	5

Table 6. WMT19 Metrics task: Kendall's τ between metrics and human judgments.

	★Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
DE-EN	.828	.845	.917	.883	**.919**
EN-DE	.778	.750	**.850**	.783	.848
EN-ET	.868	.868	.934	.906	**.949**
EN-FI	.901	.848	.901	.879	**.945**
EN-RU	.889	.889	**.944**	.889	.930
EN-ZH	.736	.729	.685	**.833**	.827
ET-EN	.884	.900	.884	.878	**.904**
FI-EN	.944	.944	.889	.915	**.957**
RU-EN	.786	.786	**.929**	.857	.869
ZH-EN	.824	**.872**	.738	.780	.820
EN-CS	**1.000**	**1.000**	.949	**1.000**	.949
Median	.868	.868	.901	.879	.919
Mean	.858	.857	.875	.873	.902
SD	.077	.080	.087	.062	.052
TR-EN	$^{\times}$.200	$^{\times}$.738	$^{\times}$.400	$^{\times}$.316	$^{\times}$.632
EN-TR	$^{\times}$.571	$^{\times}$.400	.837	$^{\times}$.571	**.849**
CS-EN	$^{\times}$.800	$^{\times}$.800	$^{\times}$.600	$^{\times}$.800	$^{\times}$.738
Wins	1	2	3	2	6

Table 7. WMT18 Metrics task: Kendall's τ between metrics and human judgments.

	★Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
DE-EN	.564	.564	.734	.661	**.744**
EN-CS	.758	.751	.767	.758	**.878**
EN-DE	.714	**.767**	.562	.593	.720
EN-FI	.667	.697	.769	.718	**.782**
EN-RU	.556	.556	**.778**	.648	.669
EN-ZH	**.911**	**.911**	.600	.854	.899
LV-EN	**.905**	.714	**.905**	**.905**	**.905**
RU-EN	.778	.611	.611	.722	**.800**
TR-EN	**.911**	.778	.674	.733	.907
ZH-EN	.758	**.780**	.736	.824	.732
Median	.758	.733	.735	.728	.791
Mean	.752	.713	.714	.742	.804
SD	.132	.110	.103	.097	.088
FI-EN	**.867**	**.867**	$^{\times}$.733	**.867**	**.867**
EN-TR	**.857**	.714	$^{\times}$.571	.643	.849
CS-EN	$^{\times}$1.000	$^{\times}$1.000	$^{\times}$.667	$^{\times}$.667	$^{\times}$.913
Wins	5	4	2	2	6

Table 8. WMT17 Metrics task: Kendall's τ between metrics and human judgments.