Abstract

While traditional corpus-level evaluation metrics for machine translation (MT) correlate well with fluency, they struggle to reflect adequacy. Model-based MT metrics trained on segment-level human judgments have emerged as an attractive replacement due to strong correlation results. These models, however, require potentially expensive re-training for new domains and languages. Furthermore, their decisions are inherently non-transparent and appear to reflect unwelcome biases. We explore the simple type-based classifier metric, MACRO F₁, and study its applicability to MT evaluation. We find that MACRO F₁ is competitive on direct assessment, and outperforms others in indicating downstream cross-lingual information retrieval task performance. Further, we show that MACRO F₁ can be used to effectively compare supervised and unsupervised neural machine translation, and reveal significant qualitative differences in the methods’ outputs.

1 Introduction

Model-based metrics for evaluating machine translation such as BLEURT (Sellam et al., 2020), ESIM (Mathur et al., 2019), and YiSi (Lo, 2019) have recently attracted attention due to their superior correlation with human judgments (Ma et al., 2019). However, BLEU (Papineni et al., 2002) 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.

(Sellam et al., 2020). 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, ands and buts right will benefit from BLEU in the long run more than one that gets its xylophones, peripatetics, and defenestrates 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. Information extraction and other areas concerned with classification 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 when seeking 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.

Our contributions are as follows: We consider MT as a classification task, and thus admit MACROF₁ as a legitimate approach to evaluation (Section 2). We show that MACROF₁ is competitive with other popular methods at tracking human judgments in translation (Section 3.2). We offer an additional justification of MACROF₁ as a performance indicator on adequacy-focused downstream tasks such as cross-lingual information retrieval (Section 3.3). Finally, we demonstrate that MACROF₁ is just as good as the expensive BLEURT at discriminating between structurally different MT approaches in a way BLEU cannot, especially regarding the adequacy of generated text, and provide a novel approach to qualitative analysis of the effect of metrics choice on quantitative evaluation (Section 4).

2 NMT as Classification

Neural machine translation (NMT) models are often viewed as pairs of encoder-decoder networks. Viewing NMT as such is useful in practice for implementation; however, such a view is inadequate for theoretical analysis. Gowda and May (2020) provide 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 classifier-based evaluation metrics.

Consider a test corpus, \(T = \{(x^{(i)}, h^{(i)}, y^{(i)}) | i = 1, 2, 3, \ldots m\}\) where \(x^{(i)}, h^{(i)}, \) and \(y^{(i)}\) are source, system hypothesis, and reference translation, respectively. Let \(x = \{x^{(i)}\}_i\) 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 \(c \in V\),

\[
PRED(c) = \sum_{i=1}^{m} C(c, h^{(i)})
\]

\[
REFS(c) = \sum_{i=1}^{m} C(c, y^{(i)})
\]

\[
MATCH(c) = \sum_{i=1}^{m} \min\{C(c, h^{(i)}), C(c, y^{(i)})\}
\]

where \(C(c, a)\) counts the number of tokens of type \(c\) in sequence \(a\) (Papineni et al., 2002). For each class \(c \in V_{h\cap y}\), precision \((P_c)\), recall \((R_c)\), and \(F_\beta\) measure \((F_{\beta,c})\) are computed as follows:

\[
P_c = \frac{\text{MATCH}(c)}{\text{PREDs}(c)}; \quad R_c = \frac{\text{MATCH}(c)}{\text{REFS}(c)}
\]

\[
F_{\beta,c} = (1 + \beta^2) \frac{P_c \times R_c}{\beta^2 \times P_c + R_c}
\]

The macro-average consolidates individual performance by averaging by type, while the micro-average averages by token:

\[
\text{MACROF}_\beta = \frac{\sum_{c \in V} F_{\beta,c}}{|V|}
\]

\[
\text{MICROF}_\beta = \frac{\sum_{c \in V} f(c) \times F_{\beta,c}}{\sum_{c \in V} f(c)}
\]

where \(f(c) = \text{REFS}(c) + k\) for smoothing factor \(k\).

We scale MACROF₁ and MICROF₁ values to percentiles, similar to BLEU, for the sake of easier readability.

3 Justification for MACROF₁

In the following sections, we verify and justify the utility of MACROF₁ while also offering a comparison with popular alternatives such as MICROF₁, BLEU, CHRF₁, and BLEURT. We use Kendall’s rank correlation coefficient, \(\tau\), to compute the association between metrics and human judgments. Correlations with \(p\)-values smaller than \(\alpha = 0.05\) are considered to be statistically significant.

3.1 Data-to-Text: WebNLG

We use the 2017 WebNLG Challenge dataset (Gardent et al., 2017; Shimorina, 2018) 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

\footnote{We consider \(F_{\beta,c}\) for \(c \notin V_{h\cap y}\) to be 0.}

\footnote{We use \(k = 1\). When \(k \to \infty\), \(\text{MICROF}_\beta \to \text{MACROF}_1\).}

\footnote{BLEU and CHRF₁ scores reported in this work are computed with SACREBLEU; see the Appendix for details. BLEURT scores are from the base model (Sellam et al., 2020). We consider two varieties of averaging to obtain a corpus-level metric from the segment-level BLEURT: mean and median of segment-level scores per corpus.}

\footnote{https://gitlab.com/webnlg/webnlg-human-evaluation}
judgments provided have three linguistic aspects—fluency, grammar, and semantics\(^6\)—which enable us to perform a fine grained analysis of our metrics. We compute Kendall’s \(\tau\) between metrics and human judgments, which are reported in Table 2.

As seen in Table 2, the metrics exhibit much variance in agreements with human judgments. For instance, BLEURT\text{median} is the best indicator of fluency and grammar, however BLEURT\text{mean} 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, CHRF\(_1\) does well on semantics but poorly on fluency and grammar compared to BLEU. Not surprisingly, both MACRO\(_F\) and MACRO\(_F\)\(_1\), which rely solely on unigrams, are poor indicators of fluency and grammar compared to BLEU, however MACRO\(_F\)\(_1\) clearly a better indicator of semantics than BLEU. The discrepancy between MACRO\(_F\) and MACRO\(_F\)\(_1\) 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: MACRO\(_F\)\(_1\) is a strong indicator of semantic adequacy, however, it is a poor indicator of fluency. We recommend using either MACRO\(_F\) or CHRF\(_1\) when semantic adequacy and not fluency is a desired goal.

### 3.2 Machine Translation: WMT Metrics

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 (Bojar et al., 2017; Ma et al., 2018, 2019).\(^7\) We first compute scores from each MT metric, and then calculate the correlation \(\tau\) with human judgments.

As there are many language pairs and translation directions in each year, we report only the mean and median of \(\tau\), and number of wins per metric for each year in Table 3. 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.\(^8\) CHRF\(_1\) has the strongest mean and median agreement with human judgments across the years. In 2018 and 2019, both MACRO\(_F\) and MACRO\(_F\)\(_1\) mean and median agreements outperform BLEU whereas in 2017 BLEU was better than MACRO\(_F\) and MACRO\(_F\)\(_1\).

As seen in Section 3.1, MACRO\(_F\)\(_1\) weighs towards semantics whereas MACRO\(_F\) 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 statistical MT when fluency was a harder objective. Recent advancements in neural MT models such as Transformers (Vaswani et al., 2017) produce fluent outputs, and have brought us to an era where semantic adequacy is the focus.

### 3.3 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 (Grefenstette, 2012). 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

---

\(^6\)Fluency and grammar, which are elicited with nearly identical directions (Garden et al., 2017), are identically correlated.

\(^7\)http://www.statmt.org/wmt19/metrics-task.html

\(^8\)https://github.com/google-research/bleurt

---

Table 2: WebNLG data-to-text task: Kendall’s \(\tau\) 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 \(*\).

| Name   | Fluency & Grammar | Semantics |
|--------|-------------------|-----------|
| BLEU   | \(^*\) .444       | \(^*\) .500 |
| CHRF\(_1\) | .278          | .778     |
| MACRO\(_F\) | .222     | .722     |
| MACRO\(_F\)\(_1\) | .333     | .611     |
| BLEURT\text{mean} | .444 | .833     |
| BLEURT\text{median} | .611 | .667     |

---

---

---

---

---

---

---

---

---

---

---
Table 3: WMT 2017–19 Metrics task: Mean and median Kendall’s τ between MT metrics and human judgments. Correlations that are not significant at α = 0.05 are excluded from the calculation of mean, and median, and wins. See Appendix Tables 9, 10, and 11 for full details. *BLEU is pre-computed scores available in the metrics packages. In 2018 and 2019, both MACRO₁ and MICRO₁ outperform BLEU. MACRO₁ outperforms MICRO₁. CHRF₁ has strongest mean and median agreements across the years. Judging based on the number of wins, MACRO₁ has steady progress over the years, and outperforms others in 2019.

| Year | Pairs | *BLEU | BLEU | MACRO₁ | MICRO₁ | CHRF₁ |
|------|-------|-------|------|--------|--------|-------|
|      | Mean  | Median | Wins | Mean  | Median | Wins |
| 2019 | .751  | .782  | 3    | .752  | .844  | 6    | .821  | .844  | .841  | .844  | .875  | .919  |
| 2018 | .858  | .868  | 1    | .875  | .901  | 2    | .873  | .879  | .902  | .919  | .919  | .919  |
| 2017 | .752  | .758  | 5    | .713  | .733  | 2    | .714  | .728  | .791  | .791  | .791  | .791  |

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. We conduct experiments on two datasets: firstly, on data from the 2020 workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS) (Zavorin et al., 2020), and secondly, on data originally from Europarl, prepared by Lignos et al. (2019) (Europarl).

3.3.1 CLSSTS Datasets

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 is derived from Actual Term Weighted Value (ATWV) metric (Wegmann et al., 2013). We use a single CLIR system (Boschee et al., 2019) with the same IR settings for all MT models in the set, and measure Kendall’s τ between MT and IR measures. The results, in Table 4, show that MACRO₁ 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. CHRF₁ and BLEURT, which are strong contenders when generated text is directly evaluated by humans, do not indicate CLIR task performance as well as MACRO₁. As CLIR tasks require faithful meaning equivalence across the language boundary, and human translators can mistake fluent output for proper translations (Callison-Burch et al., 2007).

3.3.2 Europarl Datasets

We perform a similar analysis to Section 3.3.1 but on another cross-lingual task set up by Lignos et al. (2019) for Czech → English (CS-EN) and German → English (DE-EN), using publicly available data from the Europarl v7 corpus (Koehn, 2005). This task differs from the CLSSTS task (Section 3.3.1) in several ways. Firstly, MT metrics are computed on test sets from the news domain, whereas IR metrics are from the Europarl domain. The domains are thus intentionally mismatched between MT and IR tests. Secondly, since there are no queries specifically created for the Europarl domain, GOV2 TREC topics 701–850 are used as domain-relevant English queries. And lastly, since there are no query-document relevance human judgments for the chosen query and document sets, the documents retrieved by BM25 (Jones et al., 2000) on the English set for each query are treated as relevant documents for computing the performance of the CS-EN and DE-EN CLIR setup. As a result, IR metrics that rely on boolean query-document relevance judgments as ground truth are less informative, and we use Rank-Based Overlap (RBO; p = 0.98) (Webber et al., 2010) as our IR metric.

We perform our analysis on the same experiments as Lignos et al. (2019). NMT models for CS-EN and DE-EN translation are trained using a convolutional NMT architecture (Gehring et al., 2017) implemented in the FAIRSeq (Ott et al.,
with strong parallel corpora have made significant progress recently (Artetxe et al., 2018; Lample et al., 2018a,b; Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020). In some cases, UNMT yields a BLEU score that is comparable with strong supervised neural machine translation (SNMT) systems. In this section we leverage

| Domain | IR Score | BLEU | MACROF<sub>1</sub> | MICROF<sub>1</sub> | CHR<sub>F<sub>1</sub> | BLEURT<sub>mean</sub> | BLEURT<sub>median</sub> |
|--------|---------|------|---------------------|------------------|---------|------------------|------------------|
| LT-EN  | In      | AQWV | .429               | * .363           | .508    | * .385           | .451             | .420             |
|        |         | MAP  | .495               | .429             | .575    | .451             | .473             | .486             |
|        | In+Ext  | AQWV | * .345            | .527             | .491    | .491             | .491             | .477             |
|        |         | MAP  | * .273            | * .455           | * .418  | * .418           | * .418           | * .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 | * .455             | * .550           | * .527  | * .382           | * .418           | * .418           |
|        |         | MAP  | * .491             | * .661           | * .564  | * .491           | * .527           | * .527           |
|        | In+ext  | AQWV | * .257             | * .500           | * .330  | * .404           | * .367           | * .367           |
|        |         | MAP  | * .183             | * .426           | * .257  | * .330           | * .294           | * .294           |

Table 5: Europarl CLIR task: Kendall’s τ between MT and IR 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 * are not significant at α = 0.05.

Table 6 shows performance for these three language pairs using a variety of metrics. Despite comparable scores in BLEU and only minor differences in MICROF<sub>1</sub> and CHR<sub>F<sub>1</sub></sub>, SNMT models have consistently higher MACROF<sub>1</sub> and BLEURT than the UNMT models for all six translation directions.

In the following section, we use a pairwise maximum difference discriminator approach to compare corpus-level metrics BLEU and MACROF<sub>1</sub> on a segment level. Qualitatively, we take a closer look at the behavior of the two metrics when comparing a translation with altered meaning to a translation with differing word choices using the metric.

Unsupervised neural machine translation (UNMT) systems trained on massive monolingual data without parallel corpora have made significant progress recently (Artetxe et al., 2018; Lample et al., 2018a,b; Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020). In some cases, UNMT yields a BLEU score that is comparable with strong supervised neural machine translation (SNMT) systems. In this section we leverage

11 MACROF<sub>1</sub> to investigate differences in the translations from UNMT and SNMT systems that have similar BLEU.

We compare UNMT and SNMT for English ↔ German (EN-DE, DE-EN), English ↔ French (EN-FR, FR-EN), and English ↔ Romanian (EN-RO, RO-EN). All our UNMT models are based on XLM (Conneau and Lample, 2019), pretrained by Yang (2020). We choose SNMT models with similar BLEU on common test sets by either selecting from systems submitted to previous WMT News Translation shared tasks (Bojar et al., 2014, 2016) or by building such systems. Specific SNMT models chosen are in the Appendix (Table 12).

We were unable to find EN-DE and DE-EN systems with comparable BLEU in WMT submissions so we built standard Transformer-base (Vaswani et al., 2017) models for these using appropriate quantity of training data to reach the desired BLEU performance. We report EN-RO results with diacritic removed to match the output of UNMT.
|          | BLEU | MACROF₁ | MICROF₁ | CHRF₁ | BLEURTmean | BLEURTmedian |
|----------|------|---------|---------|-------|------------|--------------|
| DE-EN    | 32.7 | 33.9    | 4.9     | 58.7  | 57.9       | 0.8          |
| EN-DE    | 24.0 | 24.0    | 0.0     | 47.7  | 48.1       | -0.4         |
| FR-EN    | 31.1 | 31.2    | 0.1     | 60.5  | 58.3       | 2.2          |
| EN-FR    | 25.6 | 27.1    | -1.5    | 53.0  | 52.3       | 0.7          |
| RO-EN    | 30.8 | 29.6    | 1.2     | 59.8  | 56.5       | 3.3          |
| EN-RO    | 31.2 | 31.0    | 0.2     | 55.4  | 53.4       | 2.0          |

Table 6: For each language direction, UNMT (UN) models have similar BLEU to SNMT (SN) models, and CHRF₁ and MICROF₁ have small differences. However, MACROF₁ scores differ significantly, consistently in favor of SNMT. Both corpus-level interpretations of BLEURT support the trend reflected by MACROF₁, but the value differences are difficult to interpret.

| δMACROF₁ | Fav Analysis | δBLEU | Fav Analysis |
|----------|---------------|-------|---------------|
| .071     | S: synonym; U: untranslation, noun | .048 | S: word order; U: word order, untranslation, ending |
| .064     | S: synonym; U: untranslation | .046 | S: spelling variation; U: synonym, word order, punctuation |
| -.055    | U: no issues; S: untranslation | .044 | S: extra determiner; U: paraphrase, synonym, number, untranslation |
| .052     | S: synonym; U: untranslation, noun | .042 | S: synonym; U: synonym, punctuation, extra adverb |
| -.045    | U: no issues; S: untranslation | -.039 | U: no issues; S: noun, verb |
| .044     | S: synonym, word order; U: subject, truncation, word order | -.037 | U: no issues; S: punctuation |
| .044     | S: synonym, tense; U: untranslation | -.034 | U: no issues; S: symbol |
| .043     | S: inflection, word order; U: number | -.032 | U: no issues; S: adjective, noun |
| -.041    | U: adjective, verb; S: omitted verb, untranslation | -.032 | U: untranslation; S: tense, word order, meaning, active/passive voice |
| .041     | S: time, word order; U: time, nouns | -.031 | U: untranslation; S: word order, synonym, extraconj |

Table 7: Analysis of the ten DE-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MACROF₁ (left) and BLEU (right). Fav is the favored system by metrics. The complete text of the sentences is in the Appendix, Tables 15 and 16.

4.1 Pairwise Maximum Difference Discriminator

We consider cases where a metric has a strong opinion of one translation system over another, and analyze whether the opinion is well justified. In order to obtain this analysis, we employ a pairwise segment-level discriminator from within a corpus-level metric, which we call favoritism.

We extend the definition of T from Section 2 to \( T = \{ x, h_S, h_U, y \} \) where each of \( h_S \) and \( h_U \) is a separate system’s hypothesis set for \( x \). 13 Let \( M \) be a corpus-level measure such that \( M(h, y) \in \mathbb{R} \) and a higher value implies better translation quality. \( M(h^{(-i)}, y^{(-i)}) \) is the corpus-level score obtained by excluding \( h^{(i)} \) and \( y^{(i)} \) from \( h \) and \( y \), respectively. We define the benefit of segment \( i \), \( \delta_M(i; h) \):

\[
\delta_M(i; h) = M(h, y) - M(h^{(-i)}, y^{(-i)})
\]

If \( \delta_M(i; h) > 0 \), then \( i \) is beneficial to \( h \) with respect to \( M \), as the inclusion of \( h^{(i)} \) increases the corpus-level score. We define the favoritism of \( M \) toward \( i \) as \( \delta_M(i; h_S, h_U) \):

\[
\delta_M(i; h_S, h_U) = \delta_M(i; h_S) - \delta_M(i; h_U)
\]

If \( \delta_M(i; h_S, h_U) > 0 \) then \( M \) favors the translation of \( x^{(i)} \) by system \( S \) over that in system \( U \).

Table 7 reflects the results of a manual examination of the ten sentences in the DE-EN test set with greatest magnitude favoritism; complete results are in the Appendix, Tables 15 and 16. Meaning-altering changes such as ‘untranslation’, (wrong) ‘time’, and (wrong) ‘translation’ are marked in italics, while changes that do not fundamentally alter the meaning, such as ‘synonym,’ (different) ‘inflection,’ and (different) ‘word order’ are marked in plain text. 14

The results indicate that MACROF₁ generally favors SNMT, and with good reasons, as the favored translation does not generally alter sentence meaning, while the disfavored translation does. On

13The subscripts represent SNMT and UNMT in this case, though the definition is general.

14Some changes, such as ‘word order’ may change meaning; these are italicized or not on a case-by-case basis.
the other hand, for the ten most favored sentences according to BLEU, four do not contain meaning-altering divergences in the disfavored translation. Importantly, none of the sentences with greatest favoritism according to MACROF₁, all of which having meaning altering changes in the disfavored alternatives, appears in the list for BLEU. This indicates relatively bad judgment on the part of BLEU. One case of good judgment from MACROF₁ and bad judgment from BLEU regarding truncation is shown in Table 8.

From our qualitative examinations, MACROF₁ is better than BLEU at discriminating against untranslated and truncations in UNMT. The case is similar for FR-EN and RO-EN, except that RO-EN has more untranslated sentences for both SNMT and UNMT, possibly due to the smaller training data. Complete tables and annotated sentences are in the Appendix, in Section C.

5 Related Work

5.1 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 (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), and CHRF₁ (Popović, 2015). Model-based metrics have a significant number of parameters and, sometimes, external resources that must be set prior to use. These include METEOR (Banerjee and Lavie, 2005), BLEURT (Sellam et al., 2020), YiSi (Lo, 2019), ESIM (Mathur et al., 2019), and BEER (Stanojević and Sima’an, 2014). 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. (2020) have recently evaluated the utility of popular metrics and recommend the use of either CHRFF₁ or a model-based metric instead of BLEU. We compare our MACROF₁ and MICROF₁ metrics with BLEU, CHRF₁, and BLEURT (Sellam et al., 2020). While Mathur et al. (2020) use Pearson’s correlation coefficient ($r$) to quantify the correlation between automatic evaluation metrics and human judgements, we instead use Kendall’s rank coefficient ($\tau$), since $\tau$ is more robust to outliers than $r$ (Croux and Dehon, 2010).

5.2 Rare Words are Important

That natural language word types roughly follow a Zipfian distribution is a well known phenomenon (Zipf, 1949; Powers, 1998). The frequent types are mainly so-called “stop words,” function words, and other low-information types, while most content words are infrequent types. To counter this natural frequency-based imbalance, statistics such as inverted document frequency (IDF) are commonly used to weight the input words in applications such as information retrieval (Jones, 1972). In NLP tasks such as MT, where words are the output of a classifier, there has been scant effort to address the imbalance. Doddington (2002) 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 (2002), 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 (Banerjee and Lavie, 2005) and CHRF (Popović, 2015) include both precision and recall, similar to our methods. However, neither of these measures try to address the natural imbalance of class distribution. BEER (Stanojević and Sima’an, 2014) and METEOR (Denkowski and Lavie, 2011) make an explicit distinction between function and content words; such a distinction inherently captures frequency differences since 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.

5.3 F-measure as an Evaluation Metric

F-measure (Rijsbergen, 1979; Chinchor, 1992) is extensively used as an evaluation metric in classification tasks such as part-of-speech tagging, named entity recognition, and sentiment analysis (Derczynski, 2016). Viewing MT as a multi-class classifier is a relatively new paradigm (Gowda and May, 2020), and evaluating MT solely as a multi-class classifier as proposed in this work is not an established practice. However, we find that the $F_1$ measure is sometimes used for various analyses when BLEU and others are inadequate: The comparemt tool (Neubig et al., 2019) supports comparison of MT models based on $F_1$ measure of individual types. Gowda and May (2020) use $F_1$ of individual types to uncover frequency-based bias in MT models. Sennrich et al. (2016) use corpus-level unigram $F_1$ in addition to BLEU and CHRF, however, corpus-level $F_1$ is computed as MICRO$F_1$. 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 MACRO$F_1$ for MT evaluation.

6 Discussion and Conclusion

We have evaluated NLG in general and MT specifically as a multi-class classifier, and illustrated the differences between micro- and macro- averages using MICRO$F_1$ and MACRO$F_1$ as examples (Section 2). MACRO$F_1$ captures semantic adequacy better than MICRO$F_1$ (Section 3.1). BLEU, being a micro-averaged measure, served well in an era when generating fluent text was at least as difficult as generating adequate text. Since we are now in an era in which fluency is taken for granted and semantic adequacy is a key discriminating factor, macro-averaged measures such as MACRO$F_1$ are better at judging the generation quality of MT models (Section 3.2). 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 MACRO$F_1$ on downstream CLIR tasks (Section 3.3.1). Unlike BLEURT, which is also adequacy-oriented, MACRO$F_1$ 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 effects. 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. For instance, we have used MACRO$F_1$ to compare supervised and unsupervised NMT models at the same operating point measured in BLEU, and determined that supervised models have better adequacy than the current unsupervised models (Section 4).

Macro-average is a useful technique for addressing the importance of the long tail of language, and MACRO$F_1$ is our first step in that direction; we anticipate the development of more advanced macro-averaged metrics that take advantage of higher-order and character n-grams in the future.

7 Ethical Consideration

Since many ML models including NMT are themselves opaque and known to possess data-induced biases (Prates et al., 2019), using opaque and biased evaluation metrics in concurrence makes it even harder to discover and address the flaws in modeling. Hence, we have raised concerns about the opaque nature of the current model-based evaluation metrics, and demonstrated examples displaying unwelcome biases in evaluation. We advocate the use of the MACRO$F_1$ metric, as it is easily interpretable and offers the explanation of score as a composition of individual type performances. In addition, MACRO$F_1$ treats all types equally, and has no parameters that are directly or indirectly estimated from data sets. Unlike MACRO$F_1$, MICRO$F_1$ and other implicitly or explicitly micro-averaged metrics assign lower importance to rare concepts and their associated rare types. The use of micro-averaged metrics in real world evaluation could lead to marginalization of rare types.
**Failure Modes:** The proposed MACROF₁ metric is not the best measure of fluency of text. Hence we suggest caution while using MACROF₁ to draw fluency related decisions. MACROF₁ is inherently concerned with words, and assumes the output language is easily segmentable into word tokens. Using MACROF₁ to evaluate translation into alphabetical languages such as Thai, Lao, and Khmer, that do not use white space to segment words, requires an effective tokenizer. Absent this the method may be ineffective; we have not tested it on languages beyond those listed in Section B.

**Reproducibility:** Our implementation of MACROF₁ and MICROF₁ has the same user experience as BLEU as implemented in SACREBLEU; signatures are provided in Section A. In addition, our implementation is computationally efficient, and has the same (minimal) software and hardware requirements as BLEU. All data for MT and NLG human correlation studies is publicly available and documented. Data for reproducing the IR experiments in Section 3.3.2 is also publicly available and documented. The data for reproducing the IR experiments in Section 3.3.1 is only available to participants in the CLSSTS shared task.

**Climate Impact:** Our proposed metrics are on par with BLEU and such model-free methods, which consume significantly less energy than most model-based evaluation metrics.

**Acknowledgements**

The authors thank Shantanu Agarwal, Joel Barry, and Scott Miller for their help with CLSSTS CLIR experiments, and Daniel Cohen for discussions on IR evaluation methods. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

**References**

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised 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.

Santanjeev 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, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In Proceedings of the Second Conference on Machine Translation, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In Proceedings of the First Conference on Machine Translation, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Elizabeth Boschee, Joel Barry, Jayadev Billa, Marjorie Freedman, Thamme Gowda, Constantine Lignon, 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, pages 19–24, Florence, Italy. Association for Computational Linguistics.

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, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.
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, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

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 (Demonstrations), pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-nt: 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), pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations.

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, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

David M. W. Powers. 1998. Applications and explanations of Zipf’s law. In New Methods in Language Processing and Computational Natural Language Learning.

Marcelo O.R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications, pages 1–19.

C. J. Van Rijsbergen. 1979. Information Retrieval, 2nd edition. Butterworth-Heinemann, USA.

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, pages 7881–7892, Online. Association for Computational Linguistics.

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), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anastasia Shimorina. 2018. Human vs automatic metrics: on the importance of correlation design. CoRR, abs/1805.11474.

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, volume 200. Cambridge, MA.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In Proceedings of Machine Learning Research, volume 97, pages 5926–5936, Long Beach, California, USA. PMLR.

Miloš Stanojević and Khalil Sima’an. 2014. BEER: BBetter evaluation as ranking. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, 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, pages 5998–6008.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS), 28(4):1–38.

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, pages 192–197. IEEE.

Hans Yang. 2020. XLM-UNMT-Models. https://github.com/Hansxsource/XLM-UNMT-Models.

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 (CLSSSTS2020), pages 7–13, Marseille, France. European Language Resources Association.

George Kingsley Zipf. 1949. Human behaviour and the principle of least-effort. Cambridge MA edn. Addison-Wesley.
A Metrics Reproducibility

BLEU scores reported in this work are computed with the SACREBLEU library and have signature BLEU-case.mixed-lang.<xxx>-<yy>+numrefs.1+smooth.exp.tok.<TOK>+version.1.4.13, where <TOK> is zh for Chinese, and 13a for all other languages. MACRO1 and MICRO1 use the same tokenizer as BLEU. CHR1 is also obtained using SACREBLEU and has signature chr1-lang.<xxx>-<yy>+numchars.6+space.false+version.1.4.13. BLUERT scores are from the base model of Sellam et al. (2020), which is fine-tuned on WMT Metrics ratings data from 2015-2018. The BLEURT model is retrieved from https://storage.googleapis.com/bleurt-oss/bleurt-base-128.zip.

MACRO1 and MICRO1 are computed using our fork of SACREBLEU as:

```
sacrebleu $REF -m macrof microf < $HYP.
```

B Agreement with WMT Human Judgments

Tables 9, 10, and 11 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 ¨ are not significant at α = 0.05, and hence corresponding rows are excluded from the calculation of mean, median, and standard deviation.

Since MACRO1 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 (Ma et al., 2019), 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, MACRO1 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 MACRO1. However, since the τ values for EN-ZH are relatively lower than the other language pairs, we conclude that poor correlation of MACRO1 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 | MACRO1 | MICRO1 | CHRF1 |
|-------|------|--------|--------|-------|
| 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 | .849   | .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  |
| LU-EN | .692 | .729   | .714   | .780  | .757  |
| ZH-EN | .695 | .695   | .752   | .676  | .715  |

Table 9: WMT19 Metrics task: Kendall’s τ between metrics and human judgments.

C UNMT and SNMT Models

The UNMT models follow XLM’s standard architecture and are trained with 5 million monolingual sentences for each language using a vocabulary size of 60,000. We train SNMT models for EN↔DE and select models with the most similar (or a slightly lower) BLEU as their UNMT counterparts on newstest2019. The DE-EN model selected is trained with 1 million sentences of parallel data and a vocabulary size of 64,000, and the EN-DE model selected is trained with 250,000 sentences of parallel data and a vocabulary size of 48,000. For EN↔FR and EN↔RO, we select SNMT models from submitted systems to WMT shared tasks that have similar or slightly lower BLEU scores to corresponding UNMT models, based on NewsTest2014 for EN↔FR and NewsTest2016 for EN↔RO.

Figure 1, which is a visualization of MACRO1 for SNMT and UNMT models, shows that UNMT is generally better than SNMT on frequent types, however, SNMT outperforms UNMT on the rest leading to a crossover point in MACRO1 curves. Since MACRO1 assigns relatively higher weights to infrequent types than in BLEU, SNMT gains higher MACRO1 than UNMT while both have approximately the same BLEU, as reported in Table 6.

A complete comparison of UNMT vs SNMT in different languages is in Table 12. A manual analysis of the ten sentences with the largest magnitude favoritism according to MACRO1 and BLEU in
the FR-EN and RO-EN test sets is in Table 13 and Table 14. The complete texts of these sentences, their reference translations, and the system translations (including DE-EN mentioned in Sec 4), are shown in Tables 15, 16, 17, 18, 19, and 20.

Table 10: WMT18 Metrics task: Kendall’s τ between metrics and human judgments.

| Translation        | BLEU | BLEU | MacroF1 | MicroF1 | ChrF1 |
|--------------------|------|------|---------|---------|-------|
| 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-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  |
| Wins               | 5    | 4    | 2       | 2       | 6     |
Table 13: Analysis of the ten FR-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MACROF₁ (left) and BLEU (right). Fav is the favored system by metrics. Actual examples are shown in Appendix Tables 17 and 18.

| \( \delta_{\text{MACROF}} \) | Fav Analysis | \( \delta_{\text{BLEU}} \) | Fav Analysis |
|-----------------|----------------|-----------------|----------------|
| 0.044 S         | S: synonym; U: untranslation, synonym | -0.026 U | U: synonym; S: omitted adv; word order |
| -0.038 U        | U: no issues; S: synonym | 0.025 S | S: no issues; U: determiner; word order |
| 0.035 S         | S: synonym; U: untranslation, synonym | 0.024 S | S: no issues; U: repetition; form |
| -0.034 U        | U: no issues; S: synonym; word order | 0.021 S | S: verb, synonym; U: untranslation, noun, time, synonym |
| -0.034 U        | U: synonym; S: word order, verb_ref | -0.021 S | S: synonym; U: synonym |
| -0.033 U        | U: no issues; S: synonym | -0.021 U | U: omitted NER; S: synonym, word order |
| 0.033 S         | S: word order; U: untranslation, NER, word order | -0.021 U | U: untranslation, S: verb, word order |
| 0.032 S         | S: synonym; U: number, omitted noun, untranslation, verb | -0.021 U | U: synonym, S: extra preposition, synonym, word order |
| 0.030 S         | S: adj; U: untranslation | 0.021 S | S: no issues; U: NER |
| 0.030 S         | S: noun, synonym; U: noun, synonym | -0.020 U | U: synonym, S: synonym, word order |

Table 14: Analysis of the ten RO-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MACROF₁ (left) and BLEU (right). Fav is the favored system by metrics. Actual examples are shown in Appendix Tables 19 and 20.

| \( \delta_{\text{MACROF}} \) | Fav Analysis | \( \delta_{\text{BLEU}} \) | Fav Analysis |
|-----------------|----------------|----------------|----------------|
| 0.131 S         | S: word order; U: repetition, word order | 0.114 S | S: word order; U: repetition, word order |
| 0.063 S         | S: noun, word order; U: repetition, untranslation, noun | 0.089 S | S: no issues; U: omitted noun, omitted time, NER |
| 0.062 S         | S: extra, untranslation; U: untranslation, copy | -0.072 U | U: country, untranslation; S: noun, word order |
| -0.052 U        | U: untranslation x 3, synonym; S: untranslation, synonym | -0.045 U | U: synonym; S: synonym, word order |
| -0.052 U        | U: untranslation, NER, synonym; S: NER, synonym | -0.041 U | U: untranslation; S: word order, subject |
| -0.052 U        | U: extra, S: untranslation | -0.040 U | U: no issues; S: number, omitted preposition |
| -0.050 U        | U: adv; S: incoherent, adv | 0.039 S | S: extra, untranslation; U: untranslation, copy |
| -0.050 U        | U: active/passive voice, name; S: name | 0.036 S | S: no issues; U: extra verb |
| -0.049 U        | U: untranslation; S: untranslation, word order | -0.035 U | U: repetition, untranslation, S: verb, synonym, word order |
| 0.048 S         | S: no issues; U: NER | 0.034 S | S: synonym; U: untranslation |
Es wird davon ausgegangen, dass sie über eine einzigartige Kanone, eine Reihe von Flugabwehr- und Schiffsabwehrkanonen sowie einige Stealth-Technologien verfügen, wie z. B. reduzierte Radar-, Infrarot- und akustische Signaturen.

A group of masked pro-independence separatists, who were kept away by the Bereitschaftspolizei, headed them with evocative paint and poured purple force paint and created dark Staubstations in the streets normally clogged by tourists.

A group of masked pro-separatists held hostage by the riot police brought them lacking eggs and ignited powder paint and produced dark clouds in the streets that were usually crowded by tourists.

It is understood they have a powerful canon, a number of fluke and ship fire systems and some stealth-controlled technologies, such as reduced radar, infrared and acoustic signatures.

A group of masked pro-separatists held hostage by the riot police brought them lacking eggs and ignited powder paint and produced dark clouds in the streets that were usually crowded by tourists.

For Syria and pave the way to elections.

It is understood they have a powerful canon, a number of fluke and ship fire systems and some stealth-controlled technologies, such as reduced radar, infrared and acoustic signatures.

A group of masked pro-separatists held hostage by the riot police brought them lacking eggs and ignited powder paint and produced dark clouds in the streets that were usually crowded by tourists.
**Table 16:** Top 10 segments by $|\delta_{BL}\rangle$. 

| Source | Reference | SNMT | UNMT |
|-------|-----------|------|------|
| **0.05** In der letzten Woche wurden mittlere Konzentrationen in Küstennähe und auf offener See in Pinellas County gemeldet, geringe bis hohe Konzentrationen in Manatee County, Hintergrund- bis hohe Konzentrationen auf offener See in Hillsborough County, Hintergrund- bis hohe Konzentrationen in Charlotte County, Hintergrund- bis hohe Konzentrationen in Sarasota County, Hintergrund- bis mittlere Konzentrationen in Charlotte County, Hintergrund- bis hoch Konzentrationen in Collier County. | Medium concentrations in or offshore of Pinellas County have been reported in the past week, low to high concentrations offshore of Hillsborough County, background to high concentrations in Manatee County, background to high concentrations in Sarasota County, background to medium concentrations in Charlotte County, background to medium concentrations in or offshore of Lee County, and low concentrations in Collier County. | Last week, average concentrations were reported on the coast and open seas in Pinellas County, low to high levels at open sea in Hillsborough County, background to high concentrations in Manatee County, high concentrations in coastal and open seas in Sarasota County, background to medium concentrations in Charlotte County, background to high shore and high sea levels in Lee County, and low concentrations in Collier County. | In the last week, moderate to high Konzentriflor led in Kitsa County were reported in Pinellas County, low to high Konzentriflor led levels on open water in Charlotte County, Hintergrund to high Konzentriflor led levels in Manatee County, Hintergrund to high Konzentriflor led levels in Sarasota, Charlotte County and low Konzentration in Charlotte County, Hintergrund to high Konzentriflor led levels in Lee County and on open water in Sarasota County. |
| **0.10** Moscow has repeatedly stressed that the $11$ billion Nord Stream 2 pipeline, which is set to double the existing pipeline capacity to 110 billion cubic metres, is a purely economic project. The NTS, which is responsible for the care of more than 270 historical buildings, 38 important gardens and 76,000 hectares of land around the country, takes bats very seriously. | Moscow has repeatedly stressed that the $11$ billion Nord Stream 2 pipeline, which is supposed to double the existing pipeline capacity to 110 billion cubic metres, is a purely economic project. The NTS, which is responsible for the care of more than 270 historic buildings, 38 important gardens and 76,000 hectares of land around the country, takes the bats very seriously. | Moscow has repeatedly insisted that the 11-billion pipeline, Nord Stream 2, which will double the existing Pipeline-capacity to 110 billion cubic feet, is a purely commercial project. The NTS, responsible for managing more than 270 historic buildings, 38 key gardens and 74,000 acres of land around the country, said the Fledermäuse are very important. |
| **0.20** Moskau hat wiederholt betont, dass die 11-Milliarden-Dollar-Pipeline Nord Stream 2, die die bestehende Pipeline-Kapazität auf 110 Milliar- den Kubikmeter verdoppeln soll, ein rein wirtschaftliches Projekt ist. Der NTS, der für die Betreuung von mehr als 270 historischen Gebäuden, 38 wichtigen Gärten und 76,000 Hektar Land rund um das Land verantwortlich ist, nimmt die Fledermäuse sehr ernst. | The Hill. | The Hill. | The Hill. |
| **0.30** Eine Woche nachdem eine offizielle Chinesische vierseitige Anzeige in einer US-amerikanischen Tageszeitung auf den gegenseitigen Nutzen des US-China-Handels gestellt hatte, warf der US-amerikanische Botschafter in China Peking vor, die am Wochenende durch die amerikanische Presse zu verbreiten Propaganda zu verwenden. | A week after an official Chinese four-page display in a US Chinese newspaper published a four-page ad on the mutual benefit of the US-China trade, the US ambassador to China accused of using the American press to spread propaganda. | A week after an official China's newspaper ran a four-page ad in a U.S. daily touting the mutual benefits of US-China trade, the U.S. ambassador to China accused of using the American press to spread propaganda. | A week after an official Chinese newspaper published a four-page ad on the mutual benefit of the US-China trade, the U.S. ambassador to China accused of using the American press to spread propaganda. |
| **0.40** Sie kümmern sich nicht darum, was sie verleiten, wenn sie überfahren müssen, um Macht und Kontrolle zu bekommen, ist das, was sie wollen, Macht und Kontrolle, was sie durch sie verdirben. Sie wollen, was sie haben wollen. | They don't care who they hurt, who they have to run over in order to get power and control, that's what they want, power and control, we're not going to give it to them. | They don't care who they hurt, who they have to pass over to gain power and control, that is what they want, power and control, we will not give them. | They don't care who they hurt, who they have to pass over to get power and control, that's what they want, power and control, we won't give it to them. |
| **0.50** Palin, 29, aus Wasilla, Alaska, wurde wegen des Verdachts auf häus- liche Gewalt verhaftet. Gegen ihn liegt bereits ein Bericht über häus- liche Gewalt und Widerstand bei der Festnahme vor, so eine Meldung, die am Samstag von den Alaska State Troopers veröffentlicht wurde. | Palin, 29, of Wasilla, Alaska, was arrested for domestic violence. – Against him, a report of domestic violence and resistance at arrest was already released Saturday, according to a report released Saturday by the Alaska State Troopers. | Palin, 29, of Wasilla, Alaska, was arrested for domestic violence. – Against him, a report of domestic violence and resistance at arrest was already released Saturday, according to a report released Saturday by the Alaska State Troopers. | Palin, 29, of Wasilla, Alaska, was arrested on charges of domestic violence. – Against him, a report of domestic violence and resistance at arrest was already released Saturday, according to a report released Saturday by the Alaska State Troopers. |
| **0.60** I did not hide Dr. Ford's allegations, I did not leak her story," Feinstein told the committee, the Hill reported. | I did not hide Dr. Ford's allegations, I did not leak her story," Feinstein told the committee, the Hill reported. | I did not hide Dr. Ford's allegations, I didn't leak her story," Feinstein told the committee, the Hill reported. I did not hide Dr. Ford's allegations, I didn't leak her story," Feinstein told the committee, the Hill reported. | I did not hide Dr. Ford's allegations, I didn't leak her story," Feinstein told the committee, the Hill reported. |
| **0.70** Briefings will still happen, Sanders said, but "if the press has the chance to ask the president of the United States questions directly, that's infinitely better than talking to me." | Briefings will still happen, Sanders said, but "if the press has the chance to ask the president of the United States questions directly, that's infinitely better than talking to me." | Briefings will still happen, Sanders said, but "if the press has the chance to ask the president of the United States questions directly, that's infinitely better than talking to me." | Briefings will still take place, Sanders said, but if the press has the chance to ask the president of the United States questions directly, so that is unendlich better than talking to her. |
Les violences sont de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont intersexuées, avec de formes partielles du groupe de la soixantaine de diagnostics comme des dérives du développement sexuel, un terme générique qui englobe des personnages possédant des chromosomes ou des glandes génitales (ovaires ou testicules) atypiques ou des organes sexuels anormalement développés.

Les experts disent que les personnes sont intersexuées, avec de formes partielles du groupe de la soixantaine de diagnostics comme des dérives du développement sexuel, un terme générique qui englobe des personnages possédant des chromosomes ou des glandes génitales (ovaires ou testicules) atypiques ou des organes sexuels anormalement développés.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

La violence est de plus en plus féroce en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les ministres appellent à présenter les personnes qui auraient été mordues, griffées, égratignées, ou léchées sur une zone hormis la sierra, par un chaton ou dont l’animal aurait été en contact avec ce chaton entre le 8 et le 28 octobre à contacter le 08.11.00.06.95 entre 10 heures et 18 heures à partir du 1er novembre.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

Les experts disent que les personnes sont systématiquement confrontées à faire leurs baies, malgré un changement dans la loi qui a été voté plus tôt dans l’année, interdisant aux autorités de forcer quiconque à s’incriminer lui-même.

On dit que cette violence est de plus en plus fréquentes en dépit de mesures de sécurité renforcées et d’opérations militaires d’assistance lancées depuis plus de deux ans par le gouvernement de Nouri Al Maliki, dominé par les chiites.
Le Sénat américain a approuvé

Le président Xi Jinping, qui a pris ses fonctions en mars dernier, a fait de la lutte contre la corruption une priorité nationale, estimant que le phénomène constituait une menace à l'existence même du Parti communiste.

Un peu plus tôt, sur la route menant à Bunagana, poste-frontière avec l'Ouganda, des militaires aidés de civils chargeaient un lance-roquettes multiple monté sur un camion flambant neuf des FARDC, devant assurer la relève d'un autre dispositif.

Il y a, avec la crémation, une violence faite au corps aimé, qui va être "redouté à un tas de cendres" en très peu de temps, et non plus un processus de décomposition, qui "accompagnerait les phases du deuil".

Scott Brown, le capitaine du Celtic Glasgow, a vu son appel rejeté et sera bien suspendu pour les deux prochains matchs de Ligue des champions de son club, contre l'Ajax et AC Milan.

Les irréductibles du M23, soit quelques centaines de combattants, étaient retranchés à près de 2000 mètres d'altitude, sur les collines agricoles de Chanzu, Runyonyi et Mbuzi, proches de Bunagana et Jomba, deux localités situées à environ 80 km au nord de Goma, la capitale de la province du Nord-Kivu.

Il a indiqué que le nouveau tribunal des médias "sera toujours partial car il s'agit d'un prolongement du gouvernement "et que les restrictions relatives au contenu et à la publicité nuiraient à la place du Kenya dans l'économie mondiale.

Dans "Les Fous de Benghazi", il avait été le premier à révéler l'existence d'un centre de commandement secret de la CIA dans cette ville, bercé de la révolte libyenne.

Le Sénat américain a approuvé un projet pilote de 90 M$ l'année dernière qui avait été porté sur environ 10 000 voitures.

La Ligue des champions de Milan a vu son appel rejeté, et sera bien suspendu pour les deux prochains matchs de la compétition contre Ajax et AC Milan.

Il était le premier à révéler l'existence d'un centre de commandement secret de la CIA dans la ville, bercé de la révolte libyenne.

Le Congrès américain a approuvé un projet pilote de 90 M$ l'année dernière qui avait été porté sur environ 10 000 voitures.

Le Sénat américain a approuvé un projet pilote de 90 M$ l'année dernière qui avait été porté sur environ 10 000 voitures.
David Grimal has an international career as a solo violinist. During the last 20 years he held regular concerts on the major classical music stages of the world with prestigious orchestras such as the Orchestra of Paris, Orchestre Philharmonique de Radio France, Russian National Orchestra, Orchestre National de Lyon, Chamber Orchestra of Europe, Berliner Symphoniker, New Japan Philharmonic, Orchestre de l'Opera de Lyon, Salzburg Mozarteum Orchestra, Jerusalem Symphony Orchestra and Sinfonia Varsovia under the direction of conductors such as Christoph Eschenbach, Michel Pl asson, Michael Schwan nd, Peter Csaba, Hei rich Schiff, Lawrence Foster, Em manuel Krivine, Mikael Pletnev, Rafael Frubebek de Burgos and Peter Eotvos, Andris Nels ons, Christian Arming.

People interested will be able to learn how to achieve harmonious sculptural floral designs (ikebana), how to capture the soul of the surrounding elements in plastic compositions using the Japanese ink art painting (Sumie) or how to express their own individuality and creativity through the multicolored universe of Fine Arts (graphic, drawing, painting), said Sorin Mazi u, teacher of these courses.

In addition, in the coming days we will accept the works from the bridge linking the ER to the cardiology, cardiology and gas troenterology clinics to ensure the proper transfer of patients from the ER to the neighboring clinics, explained the manager of the med ical unit.

In addition, the Arabic media announced that the departure of San marthe and Itihad is imminent, coach Ladislau Boloni’s agent, Arc dia Zapornaja, says this will no longer happen.

Most are those who completed their Ph.d.0.67 and young people who have completed high school or medi cal school - 6.20%.

Most are those who completed their degree - 0.67 and young people who finished college or high school - 6.20 per cent.

They included the medley of tax shy, rascals, phone-hacking for key politicians and pseudo non-doms who own most of our great newspapers.

They included the mixture of tax revenue, foreign fraudlent tele phoning and non-persons who hold most major newspapers.

They included the mixture of unpaid tax puns, foreigners who telephone from places who own most of the key newspapers.

They included the mixture of tax revenue, phone-hacking for key politicians and pseudo non-doms who own most of our great newspapers.

They included the mixture of tax revenue, foreign fraudlent telephoning and non-persons who hold most major newspapers.
David Grimal has an international career and has wanted to do the right thing in music. He has played in various orchestras around the world with prestigious orchestras such as the Orchestre de Paris, Orchestre Philharmonique de Radio France, Russian National Orchestra, Orchestra National de Lyon, Chamber Orchestra of Europe, Berliner Philarmoniker, New Japan Philharmonic, Orchestre de l'Opera de Lyon, Salle Pleyel (Polyvalent Hall 20-21 October), Cluj (Iasi, Polyvalent Hall 22-23 October), Siibel, Sibiu (Business Centre 28-29 October) and Targu-Mureş (National Theatre, November 5).

The Fed should put the issue of financial stability first in the only place it should in major crises such as the 2008 financial market meltdown, said Adam S. Posen, a former member of the Board of the Commission for determining the interest rate within the bank of England.

When Clinton supporters are asked to answer an open question regarding why you want her to win the race, the answer is that majority holds appropriate experience (16%), followed by the fact that the time has come for a woman to be President (13%), and that she is the best candidate for this position (10%).

When Clinton supporters are asked to answer an open question regarding why you want her to win the race, the answer is that majority holds appropriate experience (16%), followed by the fact that the time has come for a woman to be President (13%), and that she is the best candidate for this position (10%).

The Fed should put financial stability first only in a major crisis, such as the 2008 financial crisis, to avoid catastrophic issues, according to the last 20 years he held regular concerts, and said that the two teams are in the world of the stage of the main classical music of the world and with prestigious orchestras such as the Orchestre de Paris, Orchestre Philharmonique de Radio France, Russian National Orchestra, Orchestra National de Lyon, Chamber Orchestra of Europe, Berliner Philarmoniker, New Japan Philharmonic, Orchestre de l'Opera de Lyon, Salle Pleyel (Polyvalent Hall 20-21 October), Cluj (Iasi, Polyvalent Hall 22-23 October), Siibel, Sibiu (Business Centre 28-29 October) and Targu-Mureş (National Theatre, November 5).

The Fed should put financial stability first only in a major crisis, such as the 2008 financial crisis, to avoid catastrophic issues, according to the last 20 years he held regular concerts, and said that the two teams are in the world of the stage of the main classical music of the world and with prestigious orchestras such as the Orchestre de Paris, Orchestre Philharmonique de Radio France, Russian National Orchestra, Orchestra National de Lyon, Chamber Orchestra of Europe, Berliner Philarmoniker, New Japan Philharmonic, Orchestre de l'Opera de Lyon, Salle Pleyel (Polyvalent Hall 20-21 October), Cluj (Iasi, Polyvalent Hall 22-23 October), Siibel, Sibiu (Business Centre 28-29 October) and Targu-Mureş (National Theatre, November 5).

The Fed should put the issue of financial stability first in the only place it should in major crises such as the 2008 financial market meltdown, said Adam S. Posen, a former member of the Board of the Commission for determining the interest rate within the bank of England.