Macro-Average: Rare Types Are Important Too

Anonymous NAACL-HLT 2021 submission

Abstract

While traditional corpus-level evaluation metrics for machine translation correlate well with fluency, they struggle to reflect adequacy. Model-based MT metrics trained to correlate well with 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, MacroF$_1$, and study its applicability to NLG evaluation. We find that MacroF$_1$ is competitive on direct assessment, and outperforms others in indicating downstream cross-lingual information retrieval task performance. Further, we show that MacroF$_1$ can be used to effectively compare supervised and unsupervised neural machine translation, and reveal significant qualitative differences in the methods' outputs.[1]

1 Introduction

Model-based metrics for evaluating machine translation such as BLEURT (?), ESIM (?), and YiSi (?) have recently attracted attention due to their superior correlation with human judgements (?). However, Bleu (?) remains the most widely used corpus-level MT metric. It correlates reasonably well with human judgements, 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 (?). 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. 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

[1] Tools and analysis are available at https://github.com/HELD-FOR-ANONYMITY

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 MACROF$_1$ as a legitimate approach to evaluation (Section 2). We show that MACROF$_1$ is competitive with other popular methods at tracking human judgements in translation (Section 3.2). We offer an additional justification of MACROF$_1$ as a performance indicator on adequacy-focused downstream tasks such as cross-lingual information retrieval (Section 3.3). Finally, we demonstrate that MACROF$_1$ 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. **?** 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, $T = \{(x^{(i)}, h^{(i)}, y^{(i)}) | i = 1, 2, 3...m\}$ 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 all $h^{(i)}$, all $y^{(i)}$, their intersection, and their union, respectively. For each class $c \in V$,

$$\text{PREDS}(c) = \sum_{i=1}^{m} C(c, h^{(i)})$$

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

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

where $C(c, a)$ counts the number of tokens of type c in sequence a (**?**). 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:[2]

$$P_c = \frac{\text{MATCH}(c)}{\text{PREDS}(c)}; \quad R_c = \frac{\text{MATCH}(c)}{\text{REFS}(c)}$$

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

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.[3] We scale MACROF$_\beta$ and MICROF$_\beta$ values to percentile, similar to BLEU, for the sake of easier readability.

3 Justification for MACROF$_1$

In the following sections, we verify and justify the utility of MACROF$_1$ while also offering a comparison with the alternatives such as MICROF$_1$, BLEU, CHRF$_1$, and BLEURT.[4] We use Kendall's rank correlation coefficient, τ, to compute the association between metrics and human judgements. Correlations with p-vales smaller than $\alpha = 0.05$ are considered to be statistically significant.

3.1 Data-to-Text: WebNLG

Name	Fluency & Grammar	Semantics
BLEU	×.444	×.500
CHRF$_1$	×.278	.778
MACROF$_1$	×.222	.722
MICROF$_1$	×.333	.611
BLEURTmean	×.444	.833
BLEURTmedian	.611	.667

Table 2: WebNLG data-to-text task: Kendall's τ between system-level MT metric scores and human judgements. Fluency and grammar display same correlations to all the methods. Values that are *not* significant at $\alpha = 0.05$ are indicated by ×.

We use the 2017 WebNLG Challenge dataset (**??**) [5] to analyze the differences between micro-

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

[3] We use $k = 1$.

[4] BLEU and CHRF$_1$ scores reported in this work are computed with SACREBLEU; see the Appendix for details. BLEURT scores are from the *base* model **?**. 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

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

and macro- averaging. WebNLG is a task of generating English text for sets of triples extracted from DBPedia. Human annotations are available to a sample of 223 records each from nine NLG systems. The human judgements provided as three linguistic aspects – fluency, grammar, and semantics – enable us to do a fine grained analysis of autoeval metrics. We compute Kendall's τ between autoeval metrics and human judgements, which are reported in Table 2.

As seen in Table 2, the autoeval metrics exhibit much variance in agreements with human judgements. For instance, BLEURTmedian is the best indicator of fluency and grammar, however BLEURTmean scores the best on semantics. BLEURT, being a *model-based* measure that is directly trained on human judgements, scores relatively higher than others (though the decision of whether to use mean or median is unsettled). Considering BLEU as a baseline, and excluding the trained BLEURT measures, CHRF$_1$ scores high 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 where as 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 indicators 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.

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 (**???**).[6] We first compute scores from each MT metric, and then calculate the correlation τ with human judgements.

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 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.[7] CHRF$_1$ has the strongest mean and median agreement with human judgements 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 3.1, 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 statistical MT when fluency was a harder objective. Recent advancements in neural MT models such as Transformers (**?**) 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 (**?**). 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.

[6] http://www.statmt.org/wmt19/metrics-task.html

[7] https://github.com/google-research/bleurt

Year	Pairs		$\star$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 3: WMT 2017–19 Metrics task: Mean and median Kendall's τ between MT metrics and human judgements. Correlations that are not significant at $\alpha = 0.05$ are excluded from the calculation of mean, and median, and wins. See Appendix Tables 9, 10, and 11 for full details. $\star$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.

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) (**?**), and secondly, on data originally from Europarl, prepared by **?** (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 judgements. 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[8] is derived from Actual Term Weighted Value (ATWV) metric (**?**).

We use a single CLIR system with the same IR settings for all MT models in the set,[9] and measure Kendall's τ between MT and IR measures. The results, in Table 4, 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 (**?**).

3.3.2 Europarl Datasets

We perform a similar analysis to Section 3.3.1 but on another cross-lingual task set up by **?** for Czech → English (CS-EN) and German → English (DE-EN), using publicly available data from the Europarl v7 corpus (**?**). 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 701850 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 (**?**) 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 judgements as ground truth are less informative, and we use Rank-Based Overlap (RBO; $p = 0.98$) (**?**) as our IR metric.

We perform our analysis on the same experiments as **?**.[10] NMT models for CS-EN and DE-EN translation are trained using a convolutional NMT architecture (**?**) implemented in the FAIRSeq (**?**) toolkit. For each of CS-EN and DE-EN, a total of 16 NMT models that are based on different quantities of training data and BPE hyperparameter values are used. The results in Table 5

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

[9] Details of IR and MT models anonymized

[10] https://github.com/ConstantineLignos/mt-clir-emnlp-2019

	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 4: 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 $\alpha = 0.05$.

	BLEU	MACROF$_1$	MICROF$_1$	CHRF$_1$	$\overline{\text{BT}}$	$\widetilde{\text{BT}}$
CS-EN	.850	.867	.850	.850	**.900**	.867
DE-EN	.900	.900	.900	.912	**.917**	.900

Table 5: Europarl CLIR task: Kendall's τ between MT metrics and RBO. $\overline{\text{BT}}$ and $\widetilde{\text{BT}}$ are short for BLEURTmean and BLEURTmedian. All correlations are significant at $\alpha = 0.05$.

show that BLEURT has highest correlation in both cases. Apart from the trained BLEURTmedian metric, MACROF$_1$ scores higher than the others on CS-EN, and is competitive on CS-EN. MACROF$_1$ is not the metric with highest IR task correlation in this setting, unlike in Section 3.3.1, however it is competitive with BLEU and CHRF$_1$, and thus a safe choice as a downstream task performance indicator.

4 Spotting Qualitative Differences between Supervised and Unsupervised NMT with MACROF$_1$

Unsupervised neural machine translation (UNMT) systems trained on massive monolingual data without parallel corpora have made significant progress recently (**???????**). In some cases, UNMT yields a BLEU score that is comparable with strong[11] supervised neural machine translation (SNMT) systems. In this section we leverage MACROF$_1$ 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 (**?**), pretrained by **?**. We choose SNMT models with similar BLEU on common test sets by either selecting from systems submitted to previous WMT News Translation shared tasks (**??**) or by building such systems.[12] Specific SNMT models chosen are in the Appendix (Table 12).

Table 6 shows performance for these language pairs using a variety of metrics. Despite comparable BLEU and only minor differences in MICROF$_1$ and CHRF$_1$, SNMT models have consistently higher MACROF$_1$ and BLEURT than the UNMT models for all six translation directions.

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,

[11] though not, generally, the strongest

[12] We were unable to find EN-DE and DE-EN systems with comparable BLEU in WMT submissions so we built standard Transformer-base (**?**) models for these using appropriate training data to reach the desired BLEU performance. We report EN-RO results with diacritic removed to match the output of UNMT.

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

	BLEU			MACROF$_1$			MICROF$_1$			CHRF$_1$			BLEURTmean			BLEURTmedian		
	SN	UN	Δ	SN	UN	Δ	SN	UN	Δ	SN	UN	Δ	SN	UN	Δ	SN	UN	Δ
DE-EN	32.7	33.9	-1.2	38.5	33.6	4.9	58.7	57.9	0.8	59.9	58.0	1.9	.211	-.026	.24	.285	.067	.22
EN-DE	24.0	24.0	0.0	24.0	23.5	0.5	47.7	48.1	-0.4	53.3	52.0	1.3	-.134	-.204	.07	-.112	-.197	.09
FR-EN	31.1	31.2	-0.1	41.6	33.6	8.0	60.5	58.3	2.2	59.1	57.3	1.8	.182	.066	.17	.243	.154	.09
EN-FR	25.6	27.1	-1.5	31.9	27.3	4.6	53.0	52.3	0.7	56.0	57.7	-1.7	.104	.042	.06	.096	.063	.03
RO-EN	30.8	29.6	1.2	40.3	33.0	7.3	59.8	56.5	3.3	58.0	54.7	3.3	.004	-.058	.06	.045	-.004	.04
EN-RO	31.2	31.0	0.2	34.6	31.0	3.6	55.4	53.4	2.0	59.3	56.7	2.6	.030	-.046	.08	.027	-.038	.07

Table 6: For each language direction, UNMT (UN) models have similar BLEU to SNMT (SN) models, and CHRF$_1$ and MICROF$_1$ have small differences. However, MACROF$_1$ scores differ significantly, consistently in favor of SNMT. Both corpus-level interpretations of BLEURT support the trend reflected by MACROF$_1$, but the value differences are difficult to interpret.

δ_{MACROF_1}	Fav	Analysis	δ_{BLEU}	Fav	Analysis
0.071	S	S: synonym; U: *untranslation*, *noun*	0.048	S	S: word order; U: word order, *untranslation*, *ending*
0.064	S	S: synonym; U: *untranslation*	0.046	S	S: spelling variation; U: synonym, word order, punctuation
-0.055	U	U: no issues; S: *translation*	0.044	S	S: extra determiner; U: paraphrase, synonym, *number*, *untranslation*
0.052	S	S: synonym; U: *untranslation*, *noun*	0.042	S	S: synonym; U: synonym, punctuation, extra adverb
-0.045	U	U: no issues; S: *untranslation*	-0.039	U	U: no issues; S: *noun*, *verb*
0.044	S	S: synonym, word order; U: *subject*, *truncation*, word order	-0.037	U	U: no issues; S: punctuation
0.044	S	S: synonym, tense; U: *untranslation*	-0.034	U	U: no issues; S: symbol
0.043	S	S: inflection, word order; U: *number*	-0.032	U	U: no issues; S: *adjective*, *noun*
-0.041	U	U: *adjective*, *verb*; S: *omitted verb*, *untranslation*	-0.032	U	U: *untranslation*; S: tense, *word order*, *meaning*, active/passive voice
0.041	S	S: *time*, word order; U: *time*, *nouns*	-0.031	U	U: *untranslation*; S: word order, synonym, *extra_conj*

Table 7: Analysis of the ten DE-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MACROF$_1$ (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.

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 $\delta_M(i;h_S,h_U)$:

$$\delta_M(i;h_S,h_U) = \delta_M(i;h_S) - \delta_M(i;h_U) \quad (1)$$

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,' (wrong) 'inflection,' and (wrong) 'word order' are marked in plain text.[14]

The results indicate that MACROF$_1$ generally favors SNMT, and with good reason, as the favored translation does not generally alter sentence meaning, while the disfavored translation does. On 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 MACROF$_1$, eight of which all have meaning altering changes in the disfavored alternative and none in the favored, appears in the list for BLEU. This indicates relatively bad judgement on the part of BLEU. The case is similar for FR-EN and RO-EN, except that RO-EN has more untranslations for both SNMT and UNMT possibly due to the smaller training data. Complete tables and annotated sentences are in the Ap-

[14] Some changes, such as 'word order' may change meaning; these are italicized or not on a case-by-case basis.

6^{th}	$\delta_{\text{MacroF1}}(i, h_S, h_U)$: 0.044, $\delta_{\text{Bleu}}(i, h_S, h_U)$: -0.00087, $\delta_{BLEURT}(i, h_S, h_U)$: 0.97
Ref	Ever since I joined Labour 32 years ago as a school pupil, provoked by the Thatcher government's neglect that had left my comprehensive school classroom literally falling down, I've sought to champion better public services for those who need them most - whether as a local councillor or government minister.
SNMT	32 years ago, I joined Labour as a student because of the neglect of the Thatcher government, which had led to my classroom literally collapsed, and as a result I tried to promote better public services for those who need it most, whether as a local council or ministers.
UNMT	Last 32 years ago, as a student, because of the disdain for the Thatcher-era government, Labour joined Labour.
Problems	SNMT: synonym, word_order UNMT: *subject*, *truncation* , *word_order*

Table 8: An example of favoritism that illustrates the differences between MacroF$_1$ and Bleu. Translations of the DE-EN test sentence with sixth largest magnitude favoritism according to MacroF$_1$, along with the favoritism according to Bleu(not in the top ten). UNMT's translation does not include the second half of the sentence. MacroF$_1$ favors SNMT, but Bleufavors UNMT.

pendix, 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 (?), NIST (?), TER (?), and ChrF$_1$ (?). Model-based metrics have a significant number of parameters and, sometimes, external resources that must be set prior to use. These include METEOR (?), BLEURT (?), YiSi (?), ESIM (?), and BEER (?). We distinguish between model-based and model-free metrics because the former require significant effort and resources when adapting to a new language or domain, while the latter require only a test set with references.

? have recently evaluated the utility of popular metrics and recommend the use of either ChrF$_1$ or a model-based metric instead of Bleu. We compare our MacroF$_1$ and MicroF$_1$ metrics with Bleu, ChrF$_1$, and BLEURT (?).

5.2 Rare Words are Important

That natural language word types roughly follow a Zipfian distribution is a well known phenomenon (??). 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 weigh the *input* words in applications such as information retrieval (?). In NLG tasks such as MT, where words are the *output* of a classifier, there has been scant effort to address the imbalance. ? 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 ?, 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 (?) and CHRF (?) include both precision and recall similar to our methods. However, neither of these measures try to address the natural imbalance of class distribution. BEER (?) and METEOR (?) 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.

5.3 F-measure as an Evaluation Metric

F-measure (??) is extensively used as an evaluation metric in classification tasks such as part-of-speech tagging, information extraction, named entity recognition, and sentiment analysis (?). Viewing MT as a multi-class classifier is a relatively new paradigm (?), 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 compare-mt tool (?) supports comparison of MT models based on F_1 measure of individual types. **?** use F_1 of individual types to uncover frequency-based bias in MT models. **?** 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.

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 MICROF$_1$ and MACROF$_1$ as examples (Section 2). MACROF$_1$ captures semantic adequacy better than MICROF$_1$ (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 MACROF$_1$ 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 MACROF$_1$ on downstream CLIR tasks (Section 3.3.1). 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. For instance, we have used MACROF$_1$ 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 MACROF$_1$ 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 machine learning models including NMT are themselves opaque and known to possess data-induced biases (**?**), 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 MACROF$_1$ metric, as it is easily interpretable and offers the explanation of score as a composition of individual type performances. In addition, MACROF$_1$ treats all types equally, and has no parameters that are directly or indirectly estimated from data sets. Unlike MACROF$_1$, MICROF$_1$ 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 MACROF$_1$ metric is not the best measure of fluency of text. Hence we suggest caution while using MACROF$_1$ to draw fluency related decisions. MACROF$_1$ is inherently concerned with *words*, and assumes the output language is easily segmentable into word tokens. Using MACROF$_1$ 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 MACROF$_1$ and MICROF$_1$ has the same user experience as BLEU as implemented in SACREBLEU; signatures are provided in Section A. In addition, our implementation is compuationally efficient, and has the same (minimal) software and hardware requirements as BLEU. We plan to make our implementation available to the public after the anonymity period. 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 in par with BLEU and such model-free methods, which consume significantly less energy than most model-based evaluation metrics.

A Metrics Reproducibility

Bleu scores reported in this work are computed with the SacreBleu library and has signature `BLEU+case.mixed+lang.<xx>-<yy>+numrefs.1 +smooth.exp+tok.<TOK>+version.1.4.13`, where `<TOK>` is `zh` for Chinese, and `13a` for all other languages. MacroF$_1$ and MicroF$_1$ use the same tokenizer as Bleu. ChrF$_1$ is also obtained using SacreBleu and has signature `chrF1+lang.<xx>-<yy>+numchars.6 +space.false +version.1.4.13`. BLUERT scores are from the *base* model **?**, 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.

B Agreement with WMT Human Judgements

Tables 9, 10, and 11 provide τ between MT metrics and human judgements 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 (**?**), 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	0.855	0.745	0.964	0.917	**0.982**
DE-EN	0.571	0.655	0.723	0.695	**0.742**
DE-FR	0.782	0.881	**0.927**	0.844	0.915
EN-CS	0.709	**0.954**	0.927	0.927	0.908
EN-DE	0.540	0.752	0.741	0.773	**0.824**
EN-FI	0.879	0.818	0.879	0.848	**0.923**
EN-GU	0.709	0.709	0.600	**0.734**	0.709
EN-KK	0.491	0.527	**0.685**	0.636	0.661
EN-LT	0.879	0.848	**0.970**	0.939	0.881
EN-RU	0.870	0.848	**0.939**	0.879	0.930
FI-EN	0.788	0.809	**0.909**	0.901	0.875
FR-DE	**0.822**	0.733	0.733	0.764	0.815
GU-EN	0.782	0.709	0.855	0.891	**0.945**
KK-EN	**0.891**	0.844	0.796	0.844	0.881
LT-EN	0.818	**0.855**	0.844	**0.855**	0.833
RU-EN	0.692	0.729	0.714	**0.780**	0.757
ZH-EN	0.695	0.695	**0.752**	0.676	0.715
Median	0.782	0.752	0.844	0.844	0.875
Mean	0.751	0.771	0.821	0.818	0.841
SD	0.124	0.101	0.112	0.093	0.095
EN-ZH	**0.606**	**0.606**	$^\times$0.424	0.595	0.594
Wins	3	3	6	3	5

Table 9: WMT19 Metrics task: Kendall's τ between metrics and human judgements.

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 vocabulary size of 60,000. We train SNMT models for English↔German and select models with the most similar (or a slightly lower) BLEU as their UNMT counterparts on newstest2019. The German→English model selected is trained with 1 million sentences of parallel data and a vocabulary size of 64,000, and the English→German model selected is trained with 250,000 sentences of parallel data and a vocabulary size of 48,000. For English↔French and English↔Romanian, 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 English↔French and newstest2016 for English↔Romanian.

The 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 MacroF$_1$ and Bleuin 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, are shown in Table 15 and Table 16.

	$\star$Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
DE-EN	0.828	0.845	0.917	0.883	**0.919**
EN-DE	0.778	0.750	**0.850**	0.783	0.848
EN-ET	0.868	0.868	0.934	0.906	**0.949**
EN-FI	0.901	0.848	0.901	0.879	**0.945**
EN-RU	0.889	0.889	**0.944**	0.889	0.930
EN-ZH	0.736	0.729	0.685	**0.833**	0.827
ET-EN	0.884	0.900	0.884	0.878	**0.904**
FI-EN	0.944	0.944	0.889	0.915	**0.957**
RU-EN	0.786	0.786	**0.929**	0.857	0.869
ZH-EN	0.824	**0.872**	0.738	0.780	0.820
EN-CS	**1.000**	**1.000**	0.949	**1.000**	0.949
Median	0.868	0.868	0.901	0.879	0.919
Mean	0.858	0.857	0.875	0.873	0.902
SD	0.077	0.080	0.087	0.062	0.052
TR-EN	$^\times$0.200	$^\times$0.738	$^\times$0.400	$^\times$0.316	$^\times$0.632
EN-TR	$^\times$0.571	$^\times$0.400	0.837	$^\times$0.571	**0.849**
CS-EN	$^\times$0.800	$^\times$0.800	$^\times$0.600	$^\times$0.800	$^\times$0.738
Wins	1	2	3	2	6

Table 10: WMT18 Metrics task: Kendall's τ between metrics and human judgements.

	$\star$Bleu	Bleu	MacroF$_1$	MicroF$_1$	ChrF$_1$
DE-EN	0.564	0.564	0.734	0.661	**0.744**
EN-CS	0.758	0.751	0.767	0.758	**0.878**
EN-DE	0.714	**0.767**	0.562	0.593	0.720
EN-FI	0.667	0.697	0.769	0.718	**0.782**
EN-RU	0.556	0.556	**0.778**	0.648	0.669
EN-ZH	**0.911**	**0.911**	0.600	0.854	0.899
LV-EN	**0.905**	0.714	**0.905**	**0.905**	**0.905**
RU-EN	0.778	0.611	0.611	0.722	**0.800**
TR-EN	**0.911**	0.778	0.674	0.733	0.907
ZH-EN	0.758	**0.780**	0.736	0.824	0.732
Median	0.758	0.733	0.735	0.728	0.791
Mean	0.752	0.713	0.714	0.742	0.804
SD	0.132	0.110	0.103	0.097	0.088
FI-EN	**0.867**	**0.867**	$^\times$0.733	**0.867**	**0.867**
EN-TR	**0.857**	0.714	$^\times$0.571	0.643	0.849
CS-EN	$^\times$1.000	$^\times$1.000	$^\times$0.667	$^\times$0.667	$^\times$0.913
Wins	5	4	2	2	6

Table 11: WMT17 Metrics task: Kendall's τ between metrics and human judgements.

Translation	SNMT	UNMT	SNMT Name
DE-EN newstest2019	32.7	33.9	*Our Transformer*
EN-DE newstest2019	24.0	24.0	*Our Transformer*
FR-EN newstest2014	31.1	31.2	OnlineA.0
EN-FR newstest2014	25.6	27.1	PROMT-Rule-based.3083
RO-EN newstest2016	30.8	29.6	Online-A.0
EN-RO newstest2016	31.2	31.0	uedin-pbmt.4362

Table 12: SNMT systems are selected such that their Bleu scores are approximately the same as the available pretrained UNMT models.

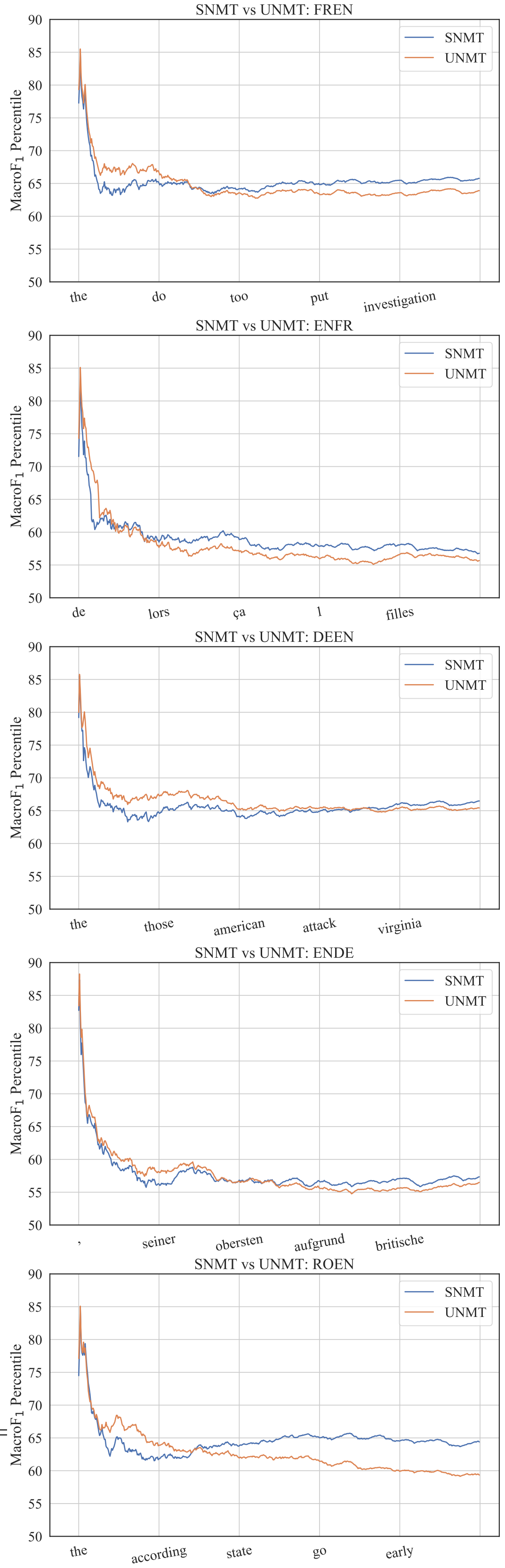

Figure 1: SNMT vs UNMT MacroF$_1$ on the frequent 500 types. UNMT outperforms SNMT on frequent types that are weighed heavily by Bleu however, SNMT is generally better than UNMT on rare types; hence, SNMT has a higher MacroF$_1$.

δ_{MacroF_1}	Fav	Analysis	δ_{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 13: Analysis of the ten FR-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MacroF_1 (left) and Bleu (right). Fav is the favored system by metrics. Actual examples are shown in Appendix Tables 17 and 18.

δ_{MacroF_1}	Fav	Analysis	δ_{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: *NET*, 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*, *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*

Table 14: Analysis of the ten RO-EN test set segments with the most favoritism in SNMT (S) or UNMT (U), according to MacroF_1 (left) and Bleu (right). Fav is the favored system by metrics. Actual examples are shown in Appendix Tables **??** and **??**.

δ_{MACROF_1}	Source	Reference	SNMT	UNMT
0.071	Es wird davon ausgegangen, dass sie über eine leistungsstarke Kanone, eine Reihe von Flugabwehr- und Schiffsabwehrraketen sowie einige Stealth-Technologien verfügen, wie z. B. reduzierte Radar-, Infrarot- und akustische Signaturen.	It is understood they will feature a powerful cannon, an array of anti-aircraft and anti-ship missiles as well as some stealth technologies, such as reduced radar, infrared and acoustic signatures.	It is assumed that they have a powerful cannon, a series of anti-aircraft and anti-ship missiles, as well as some steam technologies, such as reduced radar, infrared and acoustic signatures.	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, infrarot and akustic signposts.
0.064	Eine Gruppe maskierter Pro-Separatisten, die von der Bereitschaftspolizei zurückgehalten wurden, bewarfen sie mit Eiern und schleuderte Pulverfarbe und erzeugte in den StraSSen, die normalerweise von Touristen überfüllt waren, dunkle Staubwolken.	A group of masked pro-separatists held back by riot police pelted them with eggs and hurled powder paint, creating dark clouds of dust in streets that would usually be thronged with tourists.	A group of masked pro-separatists held hostage by the riot police brought them to eggs and ignited powder paint and produced dark clouds in the streets that were usually crowded by tourists.	A group of masked pro-independence separforces, who were kept away by the Bereitschaftpolice, beheaded them with evocative paint and poured pulverforce paint and created dark Staubations in the streets normally clogged by tourists.
-0.055	Il faut bien le faire.	Il faut bien le faire.	La gentillesse du personnel et la disponibil.	Il faut bien faire.
0.052	In einem Abschnitt gibt es ein Bild eines Schlafsaals, in dem die Studenten auf Kaffeetassen, Vorhänge, Trainer und Bücher klicken, um über die Auswirkungen von Koffein und Licht informiert zu werden und darüber, wie sich die sportliche Leistung durch Schlafmangel und die Bedeutung einer Schlafenszeitroutine beeinflusst.	In one section there is an image of a dorm room, where students click on coffee cups, curtains, trainers and books to be told about the effects of caffeine and light and how athletic performance is impacted by sleep deficiency, and the importance of a bedtime routine.	In a section there is an image of a bedroom where students click on coffee cups, curtains, trainers, and books to be informed about the effects of caffeine and light, and about how sporting performance is affected by lack of sleep and the importance of sleeping time routine.	In one section, there is a picture of a sleeping sauna where students click on coffee cups, forecourts, coaches and books to be educated about the impact of Koffein and light and about how athletic performance is influenced by sleep loss and the importance of a sleep day routine.
-0.045	Nickelbergbau ist auch in der Provinz wichtig, wird aber hauptsächlich in Morowali betrieben, an der gegenüberliegenden Küste von Sulawesi.	Nickel mining is also important in the province, but is mostly concentrated in Morowali, on the opposite coast of Sulawesi.	Nickelergbau is also important in the province, but is mainly operated in Morewali, on the opposite coast of Sulawesi.	Nickel mining is also important in the province, but is mostly operated in Morowali, on the opposite coast of Sulawesi.
0.044	Vor 32 Jahren schloss ich mich als Schüler, wegen der Vernachlässigung der Thatcher-Regierung, Labour an. Diese Vernachlässigung hatte dazu geführt, dass mein Klassenzimmer buchstäblich zusammengebrochen war. Infolgedessen habe ich versucht, mich für bessere öffentliche Dienstleistungen für diejenigen einzusetzen, die sie am meisten brauchen. Egal ob als Gemeinderat oder Minister.	Ever since I joined Labour 32 years ago as a school pupil, provoked by the Thatcher government's neglect that had left my comprehensive school classroom literally falling down, I've sought to champion better public services for those who need them most - whether as a local councillor or government minister.	32 years ago, I joined Labour as a student because of the neglect of the Thatcher government, which had led to my classroom literally collapsed, and as a result I tried to promote better public services for those who need it most, whether as a local council or ministers.	Last 32 years ago, as a student, because of the disdain for the Thatcher-era government, Labour joined Labour.
0.044	UN-Gesandter Staffan de Mistura hofft, bald die ersten Treffen eines neuen Ausschusses aus Regierungs- und Oppositionsmitgliedern einzuberufen, um eine Nachkriegsverfassung für Syrien zu entwerfen und den Weg zu Wahlen zu ebnen.	UN envoy Staffan de Mistura is hoping to soon convene the first meetings of a new committee comprised of government and opposition members to draft a post-war constitution for Syria and pave the way to elections.	UN envoy Staffan de Mistura hopes to convene soon the first meetings of a new committee of government and opposition members to draw up a post-war constitution for Syria and pave the way for elections.	U.N. Secretary General Staffan de Mistura hopes to soon join the first meetings of a new committee of government and opposition leaders to design a Nachkriegsrewrite for Syria and clear the path to elections.
0.043	CBS hatte 3,1 Millionen, NBC 2,94 Millionen, MSNBC 2,89 Millionen und CNN 2,52 Millionen, so Nielsen.	CBS had 3.1 million, NBC had 2.94 million, MSNBC had 2.89 million and CNN had 2.52 million, Nielsen said.	CBS had 3.1 million, NBC 2.94 million, MSNBC 2.89 million and CNN 2.52 million, says Nielsen.	CBS had 3.8 million, NBC 3.94 million, MSNBC 3.89 million and CNN 3.52 million, Nielsen said.
-0.041	Den Rangers gelangen nur zwei Schüsse in der ersten Hälfte, aber der ehemalige Ibrox-Torhüter Liam Kelly war kaum von Lassana Coulibalys Kopfsprung und dem Treffer eines bisslosen Ovie Ejaria aus der Ruhe zu bringen.	Rangers managed just two first-half shots on target but former Ibrox goalkeeper Liam Kelly was barely troubled by Lassana Coulibaly's header and a tame Ovie Ejaria strike.	The Ranners only reach two shots in the first half, but the former Ibrox-Torkeeper Liam Kelly was hardly the head of Lassanna Coulibys and the hit of a bissloze Ovi Ejaria.	The Rangers managed only two shots in the first half but former Ibrox goalkeeper Liam Kelly was unlikely to be helped by Lassana Coulibaly's headfirst tackle and the goal of a bisected Ovie Ejaria.
0.041	Liverpool tritt am MIttwoch um 15.00 Uhr im Stadio San Paolo in Neapel, Italien, gegen Napoli an.	Liverpool battles Napoli in the group stage of the Champions League at 3 p.m. on Wednesday at Stadio San Paolo in Naples, Italy.	Liverpool will take place at 3 p.m. at the Stadio San Paolo in Naples, Italy, against Napoli.	Liverpool v Napoli at the MItch Stadium at 15.00 pm in Neapel, Italy, on MItch.

Table 15: Top 10 segments by $|\delta_{\text{MACROF}_1}(i, h_S, h_U)|$ on DE-EN.

δ_{BLEU}	Source	Reference	SNMT	UNMT
0.048	In der letzten Woche wurden mittlere Konzentrationen in Küstennähe und auf offener See in Pinellas County gemeldet, geringe bis hohe Konzentrationen auf offener See in Hillsborough County, Hintergrund- bis hohe Konzentrationen in Manatee County, Hintergrund- bis hohe Konzentrationen in Küstennähe und auf offener See in Sarasota County, Hintergrund- bis mittlere Konzentrationen in Charlotte County, Hintergrund- bis hohe Konzentrationen in Küstennähe und auf hoher See in Lee County sowie geringe 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 or offshore of Sarasota County, background to medium concentrations in Charlotte County, background to high 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 Mantee County, high concentrations in coastal and open seas in Sarsota 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 Konzentrof lead in Küstas County were reported in Pinellas County, low to high Konzentrof lead levels on open water in Hillsborough County, Hintergrundto high levels in Manatee County, Hintergrundto high to high Konzentrin Küstas and on open water in Sarasota County and low Konzentrationen in Charlotte County, Hintergrundto high to high Konzentrin Küstennähe and on open water in Sarasota County.
0.046	Moskau hat wiederholt betont, dass die 11-Milliarden-Dollar-Pipeline Nord Stream 2, die die bestehende Pipeline-Kapazität auf 110 Milliarden Kubikmeter verdoppeln soll, ein rein wirtschaftliches Projekt ist.	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 meters, is a purely economic project.	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.	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.
0.044	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 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.	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.	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.
0.042	George W. Bush telefonierte mit Senatoren, um diese zu überreden, Herrn Kavanaugh zu unterstützen, der im WeiSSen Haus für Herrn Bush gearbeitet hatte und durch ihn seine Frau Ashley traf, die die persönliche Sekretärin von Herrn Bush war.	George W. Bush has been picking up the phone to call Senators, lobbying them to support Mr Kavanaugh, who worked in the White House for Mr Bush and through him met his wife Ashley, who was Mr Bush's personal secretary.	George W. Bush contacted senators to persuade them to support Mr Kavanaugh, who worked in the White House for Mr Bush and met his wife Ashley, who was Mr Bush's personal secretary.	George W. Bush spoke to senators to help him overture to support Mr. Kavanaugh, who had worked in the White House for Mr. Bush and met through him his wife, Ashley, who was the personal secretary to Mr. Bush.
-0.039	Eine Woche nachdem eine offizielle chinesische Zeitung eine 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 amerikanische Presse zur Verbreitung von Propaganda zu verwenden.	A week after an official Chinese newspaper ran a four-page ad in a U.S. daily touting the mutual benefits of U.S.-China trade, the U.S. ambassador to China accused Beijing of using the American press to spread propaganda.	A week after an official Chinese newspaper published a four-page display in a US daily on the mutual benefit of US China trade, the US ambassador to China published in Beijing to use the American press for 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 Beijing of using the American press to spread propaganda.
-0.037	Sie kümmern sich nicht darum, wen sie verletzen, wen sie überfahren müssen, um Macht und Kontrolle zu bekommen, das ist, was sie wollen, Macht und Kontrolle, wir werden es ihnen nicht geben.	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 is power and control, we're not going to give it to them."	They do not care about who they hurt whom they must 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 to get power and control, that's what they want, power and control, we won't give it to them.
-0.034	Mayorga behauptet, Ronaldo sei nach dem angeblichen Vorfall auf die Knie gefallen und habe ihr gesagt, er sei zu 99 Prozent ein guter Kerl, der von den ein Prozent im Stich gelassen wurde.	Mayorga claims Ronaldo fell to his knees after the alleged incident and told her he was "99 percent" a "good guy" let down by the "one percent."	Mayorga claims that Ronaldo fell to the knees after the alleged incident, saying that he was 99% a good guy left in the lurch by the one percent .	Mayorga claims Ronaldo fell on his knee after the alleged incident and told her he was "to 99 percent" a "good guy" who was left in the dark by the "one percent."
-0.032	Palin, 29, aus Wasilla, Alaska, wurde wegen des Verdachts auf häusliche Gewalt verhaftet. Gegen ihn liegt bereits ein Bericht über häusliche 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 on suspicion of domestic violence, interfering with a report of domestic violence and resisting arrest, according to a report released Saturday by Alaska State Troopers.	Palin, 29, from Wasilla, Alaska, was arrested for alleged domestic violence, and a report on domestic violence and opposition to arrest has already been published on Saturday by the Alaska State Trooperator.	Palin, 29, of Wasilla, Alaska, was arrested on charges of domestic violence. – Against him, a report of domestic violence and resistance in arrest was already released Saturday, according to a report released Saturday by Alaska State Troopers.
-0.032	"Ich habe [...] nicht versteckt Fords Behauptungen, ich habe ihre Geschichte nicht geleakt, erzählte Feinstein dem Komitee, berichtete The Hill.	"I did not hide Dr. Ford's allegations, I did not leak her story," Feinstein told the committee, The Hill reported.	"I have [...] not hidden Ford's claims that I have not lived their history," told Finestein the committee, reported The Hill.	"I did not hide [Forman's claims, I didn't geleast her story," Feinstein told the committee, The Hill reported.
-0.031	Briefings werden immer noch stattfinden, sagte Sanders, aber sollte die Presse die Chance haben, dem Präsidenten der Vereinigten Staaten die Fragen direkt zu stellen, so sei das unendlich besser, als mit ihr zu sprechen.	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.	briefing is still going to take place, Sanders said, but the press should have the opportunity to put the questions directly to the President of the United States, if that is infinitely better than to talk to her.	Briefings will still take place, Sanders said, but if the press has the chance to ask the president of the United States directly, so that is unendlich better than talking to her.

Table 16: Top 10 segments by $|\delta_{\text{BLEU}}(i, h_S, h_U)|$ on DE-EN.

δ_{MacroF_1}	Source	Reference	SNMT	UNMT
0.044	Il ne fallait qu'en déployer les accidents, et l'affaire, jacobinisme oblige, était confiée aux préfets et aux sous-préfets, interprètes autorisés.	All it took was to highlight its mistakes and, in keeping with Jacobinism, the issue would be entrusted to prefects and sub-prefects - the authorised interpreters.	It should deploy the accidents, and the case, Jacobinism obliges, was entrusted to the prefects and the sub-prefects, authorized interpreters.	It only took to deploy the accidents, and the matter, jacobinite oblige, was handed to the préfets and the sous-préfets, authorized interprètes.
-0.038	Les spécialistes disent que les personnes sont systématiquement contraintes à faire leurs aveux, 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.	Experts say confessions are still routinely coerced, despite a change in the law earlier this year banning the authorities from forcing anyone to incriminate themselves.	The experts say that people are systematically forced to make their confessions, despite a change in the law which was passed earlier this year prohibiting the authorities to force anyone to incriminating himself.	Experts say people are routinely forced to make their confessions, despite a change in the law that was passed earlier in the year banning officials from forcing anyone to incriminate themselves.
0.035	Ils sont intersexués, l'intersexualité faisant partie du groupe de la soixantaine de maladies diagnostiquées comme désordres du développement sexuel, un terme générique désignant les personnes possédant des chromosomes ou des gonades (ovaires ou testicules) atypiques ou des organes sexuels anormalement développés.	They are intersex, part of a group of about 60 conditions that fall under the diagnosis of disorders of sexual development, an umbrella term for those with atypical chromosomes, gonads (ovaries or testes), or unusually developed genitalia.	They are intersex, intersex forming part of the Group of 60 diseases diagnosed as disorders of sexual development, a generic term for people with chromosomes or atypical gonads (ovaries or testes) or abnormally developed sexual organs.	They are intersexuzed, with intersexuality making up the group of the soixantaine of diseases diagnosed as disordered sexual development, a generic term dissignant people possessing chromosomes or gonades (ovaries or testicules) atypiques or anormally developed sexual organs.
-0.034	Ces violences sont de plus en plus meurtrières en dépit de mesures de sécurité renforcées et d'opérations militaires d'envergure lancées depuis des mois par le gouvernement de Nouri Al Maliki, dominé par les chiites.	The violence is becoming more and more deadly in spite of reinforced security measures and large-scale military operations undertaken in recent months by Nouri Al Maliki's government, which is dominated by Shiites.	Such violence are more lethal despite measures enhanced security and large-scale military operations launched by the Government of Nouri Al Maliki, the Shia-dominated for months.	Those violence is increasingly deadly in the face of increased security measures and major military operations launched for months by Nouri Al Maliki's government, dominated by Shiites.
-0.034	Du côté du gouvernement, on estime que 29 954 membres des forces armées du président Bachar el-Assad ont trouvé la mort, dont 18 678 étaient des combattants des forces pro-gouvernementales et 187 des militants du Hezbollah libanais.	On the government side, it said 29,954 are members of President Bashar Assad's armed forces, 18,678 are pro-government fighters and 187 are Lebanese Hezbollah militants.	On the side of the Government, it is estimated that 29 954 members of the armed forces of president Bachar Al-Assad died, whose 18 678 were 187 Lebanese Hezbollah militants and fighters of the pro-Government forces.	On the government side, one estimate says 29,954 members of President Bachar al-Assad's armed forces have found their way, including 18,678 were fighters from pro-government forces and 187 from Lebanese Hezbollah militants.
-0.033	Mercredi, le Centre américain de contrôle et de prévention des maladies a publié une série de directives indiquant comment gérer les allergies alimentaires des enfants à l'école.	On Wednesday, the Centers for Disease Control and Prevention released a set of guidelines to manage children's food allergies at school.	Wednesday, the US Centre of disease prevention and control issued a set of guidelines indicating how to manage food allergies of children at the school.	Wednesday, the U.S. Centers for Disease Control and Prevention issued a series of directives indicating how to handle children's food allergies at school.
0.033	N'est-il pas surprenant de lire dans les colonnes du Monde à quelques semaines d'intervalle d'une part la reproduction de la correspondance diplomatique américaine et d'autre part une condamnation des écoutes du Quai d'Orsay par la NSA ?	And is it not surprising to read in the pages of Le Monde, on the one hand, a reproduction of diplomatic correspondence with the US and, on the other, condemnation of the NSA's spying on the Ministry of Foreign Affairs on the Quai d'Orsay, within a matter of weeks?	Is it not surprising to read in the columns of the world a few weeks apart on the one hand the reproduction of American diplomatic correspondence and on the other hand a condemnation of the Quai d'Orsay by the NSA listens?	Isn't it surprising to read in the Times' pages just weeks apart of one side's reproduction of the American diplomatic correspondance and of another a condamnation of the Quai d' Orsay's écoutes by the NSA?
0.032	Les ministères appellent à présent les personnes qui auraient été mordues, griffées, égratignées, ou léchées sur une muqueuse ou sur une peau lésée par ce 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.	The ministries are currently asking anyone who might have been bitten, clawed, scratched or licked on a mucous membrane or on damaged skin by the kitten, or who own an animal that may have been in contact with the kitten between 08 to 28 October, to contact them on 08 11 00 06 95 between 10am and 6pm from 01 November.	Departments now call people who have been bitten, scratched, scratched or licked on mucous membranes or skin injured by this kitten or where the animal would have been in contact with this kitten between 8 and 28 October to contact the 08.11.00.06.95 between 10 a.m. and 6 p.m. from November 1.	The at present call on people who may have been morbid, griffon, egregious or layed on a mug or on a skin léché by this chateau or whose animal may have been in contact with that chateau between 8 and 28 October to contact 08.11.00.095 between 10 and 18 November.
0.03	A cette IIIe République, moment central et créateur, Pierre Nora a montré beaucoup d'intérêt et même de tendresse: saluant ceux qui se sont alors employés à réparer la fracture révolutionnaire, en enseignant aux écoliers tout ce qui dans l'ancienne France préparait obscurément la France moderne et en leur proposant une version unifiée de leur histoire.	Pierre Nora has shown has shown great interest and even tenderness for this Third Republic: he salutes those who tried at the time to repair the divide created by the Revolution by teaching students about everything in the former France that obscurely paved the way for the modern France, and by offering them a unified version of their history.	This third Republic, while central and creator, Pierre Nora has shown great interest and even tenderness: saluting those who then worked to repair the revolutionary divide, by teaching students what in the former France preparing darkly modern France and offering them a version unified in their history.	At this IIIe République, central and creator moment, Pierre Nora showed much interest and even tendresse: praising those who then helped to repair the revolutionary fracture, teaching schoolchildren everything in the former French Republic that obscurantly prepared modern France and offering them a unifying version of their history.
0.03	La théorie dominante sur la façon de traiter les enfants pourvus d'organes sexuels ambigus a été lancée par le Dr John Money, de l'université Johns-Hopkins, qui considérait que le genre est malléable.	The prevailing theory on how to treat children with ambiguous genitalia was put forward by Dr. John Money at Johns Hopkins University, who held that gender was malleable.	The prevailing theory about how to treat children with ambiguous sexual organs was launched by Dr. John Money of the Johns Hopkins University, who considered that the genre is malleable.	The dominant theory about how to treat children armed with ambigüous sex organs was launched by Dr. John Money, of Johns-Hopkins University, who considered the genre maudlin.

Table 17: Top 10 segments by $|\delta_{\text{MacroF}_1}(i, h_S, h_U)|$ on FR-EN.

δ_{BLEU}	Source	Reference	SNMT	UNMT
-0.026	Mais le représentant Bill Shuster (R-Pa.), président du Comité des transports de la Chambre des représentants, a déclaré qu'il le considérait aussi comme l'alternative la plus viable à long terme.	But Rep. Bill Shuster (R-Pa.), chairman of the House Transportation Committee, has said he, too, sees it as the most viable long-term alternative.	But Congressman Bill Shuster (R - PA.), Chairman of the House of representatives Transportation Committee, said that he considered as the most viable alternative in the long term.	But Rep. Bill Shuster (R-Pa.), chairman of the House Transportation Committee, said he also considered it the most viable long-term alternative.
0.025	Les neuf premiers épisodes de Sheriff Callie's Wild West seront disponibles à partir du 24 novembre sur le site watchdisneyjunior.com ou via son application pour téléphones et tablettes.	The first nine episodes of Sheriff Callie's Wild West will be available from November 24 on the site watchdisneyjunior.com or via its application for mobile phones and tablets.	The first nine episodes of Sheriff Callie's Wild West will be available from November 24 on the site watchdisneyjunior.com or via its application for phones and tablets.	Sheriff first nine episodes of Sheriff Callie' s Wild West will be available as of November 24 on the watchdisneyjunior.com website or via its application for phones and computers.
0.024	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.	President Xi Jinping, who took office last March, has made the fight against corruption a national priority, believing that the phenomenon is a threat to the very existence of the Communist Party.	President Xi Jinping, who took office in March, has made the fight against corruption a national priority, believing that the phenomenon posed a threat to the very existence of the Communist Party.	President Xi Jinping, who took office in March, has made fighting corruption a national priority, saying the phenomenon posed a threat to the Communist Party's existence-free existence.
0.021	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 engin pilonnant les positions du M23 sur les collines.	A little earlier, on the road to Bunagana, the frontier post with Uganda, soldiers assisted by civilians loaded up a multiple rocket launcher mounted on a brand new truck belonging to the FARDC, intended to take over from another device pounding the positions of the M23 in the hills.	Earlier, on the road leading to Bunagana border post with Uganda, soldiers helped civilians loaded a multiple rocket launcher mounted on a truck brand new FARDC, to ensure succession of another engine pounding the positions of the M23 in the hills.	A day earlier, on the road leading to Bunagana, a postcode with Uganda, military personnel aided by civilians were loading a multiple rocket lance-roquettes fire on a flambant neuf FARDC truck, expected to provide the lead for another device pilfering M23 positions on the hills.
0.021	Il y a, avec la crémation, "une violence faite au corps aimé", qui va être "réduit à un tas de cendres" en très peu de temps, et non après un processus de décomposition, qui "accompagnerait les phases du deuil".	With cremation, there is a sense of "violence committed against the body of a loved one", which will be "reduced to a pile of ashes" in a very short time instead of after a process of decomposition that "would accompany the stages of grief".	There, with the cremation, "a violence made to the beloved body", which will be "reduced to a pile of ashes" in a very short time, and not after a process of decomposition, which "would accompany the phases of mourning".	There is, with cremation, "violence done to the loved one," which is going to be "reduced to a tas of ashes" in very little time, and not after a process of disablement, which would "accompany the phases of grief."
0.021	Scott Brown, le capitaine du Celtic Glasgow, a vu son appel rejeté et sera bien suspendu pour les deux prochains matches de Ligue des champions de son club, contre l'Ajax et l'AC Milan.	Scott Brown, Glasgow Celtic captain, has had his appeal rejected and will miss his club's next two Champion's League matches, against Ajax and AC Milan.	Scott Brown, the captain of the Glasgow Celtic, saw his appeal dismissed and will be well suspended for the next two matches of the champions League for his club against Ajax and AC Milan.	Scott Brown, the Celtic captain, has had his appeal rejected and will be well suspended for his club's next two Champions League matches, against Ajax and AC Milan.
-0.021	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.	The diehards of the M23, who are several hundreds in number, had entrenched themselves at an altitude of almost 2,000 metres in the farmland hills of Chanzu, Runyonyi and Mbuzi, close to Bunagana and Jomba, two towns located around 80km north of Goma, the capital of North Kivu province.	The irreducible m23, or a few hundred fighters, were cut off to nearly 2000 metres above sea level on the hills agricultural Chanzu, Runyonyi and Mbuzi, near Bunagana and Jomba, located about 80 km north of Goma, the capital of the province of North Kivu.	The irréductibles M23, or some hundred fighters, were retranchés at nearly 2000 feet of altitude on the agricultural hills of Chanzu, Runyonyi and Mbuzi, close to Bunagana and Jomba, two towns located about 80 miles north of Goma, the capital of North Kivu province.
-0.021	Il a indiqué que le nouveau tribunal des médias ń sera toujours partial car il s'agit d'un prolongement du gouvernement ż et que les restrictions relatives au contenu et à la publicité nuiraient à la place du Kenya dans l'économie mondiale.	He said the new media tribunal "will always be biased because it's an extension of the government," and that restrictions on content and advertising would damage Kenya's place in the global economy.	He said as the new media tribunal ' will be always partial because it is an extension of the Government "and content and advertising restrictions hurt instead of the Kenya into the world economy.	He said the new media tribunal "will always be partial because it is a extension of the government" and that restrictions relating to content and advertising would hurt Kenya's place in the global economy.
0.021	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, berceau de la révolte libyenne.	In "Les Fous de Benghazi", he was the first to reveal the existence of a secret CIA command centre in the city, the cradle of the Libyan revolt.	In "Les Fous de Benghazi", he was the first to reveal the existence of a secret CIA command center in this town, cradle of the Libyan revolt.	In "The Facts of Libya," he had been the first to reveal the existence of a secret CIA command center in that city, the birthplace of the Libyan uprising.
-0.02	Le Sénat américain a approuvé un projet pilote de 90 M$ l'année dernière qui aurait porté sur environ 10 000 voitures.	The U.S. Senate approved a $90-million pilot project last year that would have involved about 10,000 cars.	The US Senate has approved a pilot project of 90 M$ last year which would have covered about 10,000 cars.	The U.S. Senate approved a $90 million pilot project last year that would have focused on about 10,000 cars.

Table 18: Top 10 segments by $|\delta_{\text{BLEU}}(i, h_S, h_U)|$ on FR-EN.