Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

Qingsong Ma
Tencent-CSIG, AI Evaluation Lab
qingsong.mqs@gmail.com

Johnny Tian-Zheng Wei
UMass Amherst, CICS
jwei@umass.edu

Ondřej Bojar
Charles University, MFF ÚFAL
bojar@ufal.mff.cuni.cz

Yvette Graham
Dublin City University, ADAPT
graham.yvette@gmail.com

Abstract

This paper presents the results of the WMT19 Metrics Shared Task. Participants were asked to score the outputs of the translations systems competing in the WMT19 News Translation Task with automatic metrics. 13 research groups submitted 24 metrics, 10 of which are reference-less "metrics" and constitute submissions to the joint task with WMT19 Quality Estimation Task, "QE as a Metric". In addition, we computed 11 baseline metrics, with 8 commonly applied baselines (BLEU, SentBLEU, NIST, WER, PER, TER, CDER, and chrF) and 3 reimplementations (chrF+, sacreBLEU-BLEU, and sacreBLEU-chrF). Metrics were evaluated on the system level, how well a given metric correlates with the WMT19 official manual ranking, and segment level, how well the metric correlates with human judgements of segment quality. This year, we use direct assessment (DA) as our only form of manual evaluation.

1 Introduction

To determine system performance in machine translation (MT), it is often more practical to use an automatic evaluation, rather than a manual one. Manual/human evaluation can be costly and time consuming, and so an automatic evaluation metric, given that it sufficiently correlates with manual evaluation, can be useful in developmental cycles. In studies involving hyperparameter tuning or architecture search, automatic metrics are necessary as the amount of human effort implicated in manual evaluation is generally prohibitively large. As objective, reproducible quantities, metrics can also facilitate cross-paper comparisons. The WMT Metrics Shared Task[1] annually serves as a venue to validate the use of existing metrics (including baselines such as BLEU), and to develop new ones; see Koehn and Monz (2006) through Ma et al. (2018).

In the setup of our Metrics Shared Task, an automatic metric compares an MT system's output translations with manual reference translations to produce: either (a) *system-level* score, i.e. a single overall score for the given MT system, or (b) *segment-level* scores for each of the output translations, or both.

This year we teamed up with the organizers of the QE Task and hosted "QE as a Metric" as a joint task. In the setup of the Quality Estimation Task (Fonseca et al., 2019), no human-produced translations are provided to estimate the quality of output translations. Quality estimation (QE) methods are built to assess MT output based on the source or based on the translation itself. In this task, QE developers were invited to perform the same scoring as standard metrics participants, with the exception that they refrain from using a reference translation in production of their scores. We then evaluate the QE submissions in exactly the same way as regular metrics are evaluated, see below. From the point of view of correlation with manual judgements, there is no difference in metrics using or not using references.

The source, reference texts, and MT system outputs for the Metrics task come from the News Translation Task (Barrault et al., 2019, which we denote as Findings 2019). The texts were drawn from the news domain and involve translations of English (en) to/from

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

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 62–90
Florence, Italy, August 1-2, 2019.

Czech (cs), German (de), Finnish (fi), Gujarati (gu), Kazakh (kk), Lithuanian (lt), Russian (ru), and Chinese (zh), but excluding cs-en (15 language pairs). Three other language pairs not including English were also manually evaluated as part of the News Translation Task: German→Czech and German↔French. In total, metrics could participate in 18 language pairs, with 10 target languages.

In the following, we first give an overview of the task (Section 2) and summarize the baseline (Section 3) and submitted (Section 4) metrics. The results for system- and segment-level evaluation are provided in Sections 5.1 and 5.2, respectively, followed by a joint discussion Section 6.

2 Task Setup

This year, we provided task participants with one test set for each examined language pair, i.e. a set of source texts (which are commonly ignored by MT metrics), corresponding MT outputs (these are the key inputs to be scored) and a reference translation (held out for the participants of "QE as a Metric" track).

In the system-level, metrics aim to correlate with a system's score which is an average over many human judgments of segment translation quality produced by the given system. In the segment-level, metrics aim to produce scores that correlate best with a human ranking judgment of two output translations for a given source segment (more on the manual quality assessment in Section 2.3). Participants were free to choose which language pairs and tracks (system/segment and reference-based/reference-free) they wanted to take part in.

2.1 Source and Reference Texts

The source and reference texts we use are *newstest2019* from this year's WMT News Translation Task (see Findings 2019). This set contains approximately 2,000 sentences for each translation direction (except Gujarati, Kazakh and Lithuanian which have approximately 1,000 sentences each, and German to/from French which has 1701 sentences).

The reference translations provided in *newstest2019* were created in the same direction as the MT systems were translating. The exceptions are German→Czech where both sides are translations from English and German↔French which followed last years' practice. Last year and the years before, the dataset consisted of two halves, one originating in the source language and one in the target language. This however lead to adverse artifacts in MT evaluation.

2.2 System Outputs

The results of the Metrics Task are affected by the actual set of MT systems participating in a given translation direction. On one hand, if all systems are very close in their translation quality, then even humans will struggle to rank them. This in turn will make the task for MT metrics very hard. On the other hand, if the task includes a wide range of systems of varying quality, correlating with humans should be generally easier, see Section 6.1 for a discussion on this. One can also expect that if the evaluated systems are of different types, they will exhibit different error patterns and various MT metrics can be differently sensitive to these patterns.

This year, all MT systems included in the Metrics Task come from the News Translation Task (see Findings 2019). There are however still noticeable differences among the various language pairs.

- **Unsupervised MT Systems.** The German→Czech research systems were trained in an unsupervised fashion, i.e. without the access to parallel Czech-German texts (except for a couple of thousand sentences used primarily for validation). We thus expect the research German-Czech systems to be "more creative" and depart further away from the references. The online systems in this language directions are however standard MT systems so the German-Czech evaluation could be to some extent bimodal.
- **EU Election.** The French↔German translation was focused on a sub-domain of news, namely texts related EU Election. Various MT system developers may have invested more or less time to the domain adaptation.
- **Regular News Tasks Systems.** These

are all the other MT systems in the evaluation; differing in whether they are trained only on WMT provided data ("Constrained", or "Unconstrained") as in the previous years. All the freely available web services (online MT systems) are deemed unconstrained.

Overall, the results are based on 233 systems across 18 language pairs.[2]

2.3 Manual Quality Assessment

Direct Assessment (DA, Graham et al., 2013, 2014a, 2016) was employed as the source of the "golden truth" to evaluate metrics again this year. The details of this method of human evaluation are provided in Findings 2019.

The basis of DA is to collect a large number of quality assessments (a number on a scale of 1–100, i.e. effectively a continuous scale) for the outputs of all MT systems. These scores are then standardized per annotator.

In the past years, the underlying manual scores were reference-based (human judges had access to the same reference translation as the MT quality metric). This year, the official WMT19 scores are reference-based (or "monolingual") for some language pairs and reference-free (or "bilingual") for others.[3]

Due to these different types of golden truth collection, reference-based language pairs are in a closer match with the standard reference-based metrics, while the reference-free language pairs are better fit for the "QE as a metric" subtask.

Note that system-level manual scores are different than those of the segment-level. Since for segment-level evaluation, collecting enough DA judgements for each segment is infeasible, so we resort to converting DA judgements to golden truth expressed as relative rankings, see Section 2.3.2.

The exact methods used to calculate correlations of participating metrics with the golden truth are described below, in the two sections for system-level evaluation (Section 5.1) and segment-level evaluation (Section 5.2).

2.3.1 System-level Golden Truth: DA

For the system-level evaluation, the collected continuous DA scores, standardized for each annotator, are averaged across all assessed segments for each MT system to produce a scalar rating for the system's performance.

The underlying set of assessed segments is different for each system. Thanks to the fact that the system-level DA score is an average over many judgments, mean scores are consistent and have been found to be reproducible (Graham et al., 2013). For more details see Findings 2019.

2.3.2 Segment-level Golden Truth: daRR

Starting from Bojar et al. (2017), when WMT fully switched to DA, we had to come up with a solid golden standard for segment-level judgements. Standard DA scores are reliable only when averaged over sufficient number of judgments.[4]

Fortunately, when we have at least two DA scores for translations of the same source input, it is possible to convert those DA scores into a relative ranking judgement, if the difference in DA scores allows conclusion that one translation is better than the other. In the following, we denote these re-interpreted DA judgements as "daRR", to distinguish it clearly from the relative ranking ("RR") golden truth used in the past years.[5]

[2] This year, we do not use the artificially constructed "hybrid systems" (Graham and Liu, 2016) because the confidence on the ranking of system-level metrics is sufficient even without hybrids.

[3] Specifically, the reference-based language pairs were those where the anticipated translation quality was lower or where the manual judgements were obtained with the help of anonymous crowdsourcing. Most of these cases were translations into English (fi-en, gu-en, kk-en, lt-en, ru-en and zh-en) and then the language pairs not involving English (de-cs, de-fr and fr-de). The reference-less (bilingual) evaluations were those where mainly MT researchers themselves were involved in the annotations: en-cs, en-de, en-fi, en-gu, en-kk, en-lt, en-ru, en-zh.

[4] For segment-level evaluation, one would need to collect many manual evaluations of the exact same segment as produced by each MT system. Such a sampling would be however wasteful for the evaluation needed by WMT, so only some MT systems happen to be evaluated for a given input sentence. In principle, we would like to return to DA's standard segment-level evaluation in future, where a minimum of 15 human judgements of translation quality are collected per translation and combined to get highly accurate scores for translations, but this would increase annotation costs.

[5] Since the analogue rating scale employed by DA is marked at the 0-25-50-75-100 points, we use 25 points as the minimum required difference between two system scores to produce daRR judgements. Note that we

	DA>1	Ave	DA pairs	DARR
de-en	2,000	16.0	239,220	85,365
fi-en	1,996	9.5	83,168	38,307
gu-en	1,016	11.0	55,880	31,139
kk-en	1,000	11.0	55,000	27,094
lt-en	1,000	11.0	55,000	21,862
ru-en	1,999	11.9	131,766	46,172
zh-en	2,000	10.1	95,174	31,070
en-cs	1,997	9.1	75,560	27,178
en-de	1,997	19.1	347,109	99,840
en-fi	1,997	8.1	59,129	31,820
en-gu	998	6.9	21,854	11,355
en-kk	998	9.0	37,032	18,172
en-lt	998	9.0	36,435	17,401
en-ru	1,997	8.7	69,503	24,334
en-zh	1,997	9.8	87,501	18,658
de-cs	1,997	8.5	65,039	35,793
de-fr	1,605	4.1	12,055	4,862
fr-de	1,224	3.0	4,258	1,369
newstest2019				

Table 1: Number of judgements for DA converted to DARR data; "DA>1" is the number of source input sentences in the manual evaluation where at least two translations of that same source input segment received a DA judgement; "Ave" is the average number of translations with at least one DA judgement available for the same source input sentence; "DA pairs" is the number of all possible pairs of translations of the same source input resulting from "DA>1"; and "DARR" is the number of DA pairs with an absolute difference in DA scores greater than the 25 percentage point margin.

From the complete set of human assessments collected for the News Translation Task, all possible pairs of DA judgements attributed to distinct translations of the same source were converted into DARR better/worse judgements. Distinct translations of the same source input whose DA scores fell within 25 percentage points (which could have been deemed equal quality) were omitted from the evaluation of segment-level metrics. Conversion of scores in this way produced a large set of DARR judgements for all language pairs, shown in Table 1 due to combinatorial advantage of extracting DARR judgements from all possible pairs of translations of the same source input. We see that only German-French and esp. French-German can suffer from insufficient number of these simulated pairwise comparisons.

The DARR judgements serve as the golden standard for segment-level evaluation in WMT19.

rely on judgements collected from known-reliable volunteers and crowd-sourced workers who passed DA's quality control mechanism. Any inconsistency that could arise from reliance on DA judgements collected from low quality crowd-sourcing is thus prevented.

3 Baseline Metrics

In addition to validating popular metrics, including baselines metrics serves as comparison and prevents "loss of knowledge" as mentioned by Bojar et al. (2016).

Moses scorer[6] is one of the MT evaluation tools that aggregated several useful metrics over the time. Since Macháček and Bojar (2013), we have been using Moses scorer to provide most of the baseline metrics and kept encouraging authors of well-performing MT metrics to include them in Moses scorer.[7]

The baselines we report are:

BLEU and NIST The metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using `mteval-v13a.pl`[8] from the OpenMT Evaluation Campaign. The tool includes its own tokenization. We run `mteval` with the flag `--international-tokenization`.[9]

TER, WER, PER and CDER. The metrics TER (Snover et al., 2006), WER, PER and CDER (Leusch et al., 2006) were produced by the Moses scorer, which is used in Moses model optimization. We used the standard tokenizer script as available in Moses toolkit for tokenization.

sentBLEU. The metric SENTBLEU is computed using the script `sentence-bleu`, a part of the Moses toolkit. It is a

[6] https://github.com/moses-smt/mosesdecoder/blob/master/mert/evaluator.cpp

[7] If you prefer standard BLEU, we recommend sacreBLEU (Post, 2018a), found at https://github.com/mjpost/sacreBLEU.

[8] http://www.itl.nist.gov/iad/mig/tools/

[9] International tokenization is found to perform slightly better (Macháček and Bojar, 2013).

	Metric	Features	Learned?	Scoring Level Seg	Scoring Level Sys	Citation/Participant	Availability
Baselines	sentBLEU	n-grams		•	–		(mosesdecoder) mert/sentence-bleu
	BLEU	n-grams		–	•	Papineni et al. (2002)	(mosesdecoder) scripts/generic/mteval-v13a.pl
	NIST	n-grams		–	•	Doddington (2002)	(mosesdecoder) scripts/generic/mteval-v13a.pl
	WER	Levenshtein distance		–	•	Leusch et al. (2006)	(mosesdecoder) mert/evaluator
	TER	edit distance, edit types		–	•	Snover et al. (2006)	(mosesdecoder) mert/evaluator
	PER	edit distance, edit types		–	•	Leusch et al. (2003)	(mosesdecoder) mert/evaluator
	CDER	edit distance, edit types		–	•	Leusch et al. (2006)	(mosesdecoder) mert/evaluator
	chrF	character n-grams		•	⊘	Popović (2015)	http://github.com/m-popovic/chrF
	chrF+	character n-grams		•	⊘	Popović (2017)	http://github.com/m-popovic/chrF
	sacreBLEU-BLEU	n-grams		–	•	Post (2018a)	http://github.com/mjpost/sacreBLEU
	sacreBLEU-chrF	n-grams		–	•	Post (2018a)	http://github.com/mjpost/sacreBLEU
Metrics	BEER	char. n-grams, permutation trees	yes	•	⊘	Univ. of Amsterdam, ILCC (Stanojević and Sima'an, 2015)	http://github.com/stanojevic/beer
	BERTr	contextual word embeddings		•	⊘	Univ. of Melbourne (Mathur et al., 2019)	http://github.com/nitikam/mteval-in-context
	characTER	char. edit distance, edit types		•	⊘	RWTH Aachen Univ. (Wang et al., 2016a)	http://github.com/rwth-i6/CharacTER
	EED	char. edit distance, edit types		•	⊘	RWTH Aachen Univ. (Stanchev et al., 2019)	http://github.com/rwth-i6/ExtendedEditDistance
	ESIM	learned neural representations	yes	•	⊘	Univ. of Melbourne (Mathur et al., 2019)	http://github.com/nitikam/mteval-in-context
	LEPORa	surface linguistic features		•	⊘	Dublin City University, ADAPT (Han et al., 2012, 2013)	http://github.com/poethan/LEPOR
	LEPORb	surface linguistic features		•	⊘	Dublin City University, ADAPT (Han et al., 2012, 2013)	http://github.com/poethan/LEPOR
	Meteor++_2.0 (syntax)	word alignments		•	⊘	Peking University (Guo and Hu, 2019)	–
	Meteor++_2.0 (syntax+copy)	word alignments		•	⊘	Peking University (Guo and Hu, 2019)	–
	PReP	psuedo-references, paraphrases		•	⊘	Tokyo Metropolitan Univ. (Yoshimura et al., 2019)	http://github.com/kokeman/PReP
	WMDO	word mover distance		•	⊘	Imperial College London (Chow et al., 2019a)	–
	YiSi-0	semantic similarity		•	⊘	NRC (Lo, 2019)	http://github.com/chikiulo/YiSi
	YiSi-1	semantic similarity		•	⊘	NRC (Lo, 2019)	http://github.com/chikiulo/YiSi
	YiSi-1_srl	semantic similarity		•	⊘	NRC (Lo, 2019)	http://github.com/chikiulo/YiSi
QE Systems	IBM1-morpheme	LM log probs., ibm1 lexicon		•	⊘	Dublin City University, ADAPT (Popovic, 2012)	–
	IBM1-pos4gram	LM log probs., ibm1 lexicon		•	⊘	Dublin City University, ADAPT (Popovic, 2012)	–
	LP	contextual word emb., MT log prob.	yes	•	⊘	Univ. of Tartu (Yankovskaya et al., 2019)	–
	LASIM	contextual word embeddings	yes	•	⊘	Univ. of Tartu (Yankovskaya et al., 2019)	–
	UNI	?	?	•	⊘	?	?
	UNI+	?	?	•	⊘	?	?
	USFD	?	?	•	⊘	Univ. of Sheffield	?
	USFD-TL	?	?	•	⊘	Univ. of Sheffield	?
	YiSi-2	semantic similarity		•	⊘	NRC (Lo, 2019)	http://github.com/chikiulo/YiSi
	YiSi-2_srl	semantic similarity		•	⊘	NRC (Lo, 2019)	http://github.com/chikiulo/YiSi

Table 2: Participants of WMT19 Metrics Shared Task. "•" denotes that the metric took part in (some of the language pairs) of the segment- and/or system-level evaluation. "⊘" indicates that the system-level scores are implied, simply taking arithmetic (macro-)average of segment-level scores. "–" indicates that the metric didn't participate the track (Seg/Sys-level). A metric is learned if it is trained on a QE or metric evaluation dataset (i.e. pretraining or parsers don't count, but training on WMT 2017 metrics task data does). For the baseline metrics available in the Moses toolkit, paths are relative to http://github.com/moses-smt/mosesdecoder/.

smoothed version of BLEU for scoring at the segment-level. We used the standard tokenizer script as available in Moses toolkit for tokenization.

chrF and chrF+. The metrics CHRF and CHRF+ (Popović, 2015, 2017) are computed using their original Python implementation, see Table 2. We ran `chrF++.py` with the parameters `-nw 0 -b 3` to obtain the CHRF score and with `-nw 1 -b 3` to obtain the CHRF+ score. Note that CHRF intentionally removes all spaces before matching the n-grams, detokenizing the segments but also concatenating words.[10]

sacreBLEU-BLEU and sacreBLEU-chrF. The metrics SACREBLEU-BLEU and SACREBLEU-CHRF (Post, 2018a) are re-implementation of BLEU and chrF respectively. We ran SACREBLEU-CHRF with the same parameters as CHRF, but their scores are slightly different. The signature strings produced by sacreBLEU for BLEU and chrF respectively are `BLEU+case.lc+lang.de-en+numrefs.1+smooth.exp+tok.intl+version.1.3.6` and `chrF3+case.mixed+lang.de-en+numchars.6+numrefs.1+space.False+tok.13a+version.1.3.6`.

The baselines serve in system and segment-level evaluations as customary: BLEU, TER, WER, PER, CDER, SACREBLEU-BLEU and SACREBLEU-CHRF for system-level only; SENTBLEU for segment-level only and CHRF for both.

Chinese word segmentation is unfortunately not supported by the tokenization scripts mentioned above. For scoring Chinese with baseline metrics, we thus pre-processed MT outputs and reference translations with the script `tokenizeChinese.py`[11] by Shujian Huang, which separates Chinese characters from each other and also from non-Chinese parts.

[10] We originally planned to use the CHRF implementation which was recently made available in Moses Scorer but it mishandles Unicode characters for now.

[11] http://hdl.handle.net/11346/WMT17-TVXH

4 Submitted Metrics

Table 2 lists the participants of the WMT19 Shared Metrics Task, along with their metrics and links to the source code where available. We have collected 24 metrics from a total of 13 research groups, with 10 reference-less "metrics" submitted to the joint task "QE as a Metrich" with WMT19 Quality Estimation Task.

The rest of this section provides a brief summary of all the metrics that participated.

4.1 BEER

BEER (Stanojević and Sima'an, 2015) is a trained evaluation metric with a linear model that combines sub-word feature indicators (character n-grams) and global word order features (skip bigrams) to achieve a language agnostic and fast to compute evaluation metric. BEER has participated in previous years of the evaluation task.

4.2 BERTr

BERTr (Mathur et al., 2019) uses contextual word embeddings to compare the MT output with the reference translation.

The BERTr score of a translation is the average recall score over all tokens, using a relaxed version of token matching based on BERT embeddings: namely, computing the maximum cosine similarity between the embedding of a reference token against any token in the MT output. BERTr uses `bert_base_uncased` embeddings for the to-English language pairs, and `bert_base_multilingual_cased` embeddings for all other language pairs.

4.3 CharacTER

CHARACTER (Wang et al., 2016b,a), identical to the 2016 setup, is a character-level metric inspired by the commonly applied translation edit rate (TER). It is defined as the minimum number of character edits required to adjust a hypothesis, until it completely matches the reference, normalized by the length of the hypothesis sentence. CHARACTER calculates the character-level edit distance while performing the shift edit on word level. Unlike the strict matching criterion in TER, a hypothesis word is considered to match a reference word and could be shifted, if the edit dis-

tance between them is below a threshold value. The Levenshtein distance between the reference and the shifted hypothesis sequence is computed on the character level. In addition, the lengths of hypothesis sequences instead of reference sequences are used for normalizing the edit distance, which effectively counters the issue that shorter translations normally achieve lower TER.

Similarly to other character-level metrics, CHARACTER is generally applied to non-tokenized outputs and references, which also holds for this year's submission with one exception. This year tokenization was carried out for en-ru hypotheses and references before calculating the scores, since this results in large improvements in terms of correlations. For other language pairs, no tokenizer was used for pre-processing.

4.4 EED

EED (Stanchev et al., 2019) is a character-based metric, which builds upon CDER. It is defined as the minimum number of operations of an extension to the conventional edit distance containing a "jump" operation. The edit distance operations (insertions, deletions and substitutions) are performed at the character level and jumps are performed when a blank space is reached. Furthermore, the coverage of multiple characters in the hypothesis is penalised by the introduction of a coverage penalty. The sum of the length of the reference and the coverage penalty is used as the normalisation term.

4.5 ESIM

Enhanced Sequential Inference Model (ESIM; Chen et al., 2017; Mathur et al., 2019) is a neural model proposed for Natural Language Inference that has been adapted for MT evaluation. It uses cross-sentence attention and sentence matching heuristics to generate a representation of the translation and the reference, which is fed to a feedforward regressor. The metric is trained on singly-annotated Direct Assessment data that has been collected for evaluating WMT systems: all WMT 2018 to-English data for the to-English language pairs, and all WMT 2018 data for all other language pairs.

4.6 hLEPORb_baseline, hLEPORa_baseline

The submitted metric HLEPOR_BASELINE is a metric based on the factor combination of length penalty, precision, recall, and position difference penalty. The weighted harmonic mean is applied to group the factors together with tunable weight parameters. The system-level score is calculated with the same formula but with each factor weighted using weight estimated at system-level and not at segment-level.

In this submitted baseline version, HLEPOR_BASELINE was not tuned for each language pair separately but the default weights were applied across all submitted language pairs. Further improvements can be achieved by tuning the weights according to the development data, adding morphological information and applying n-gram factor scores into it (e.g. part-of-speech, n-gram precision and n-gram recall that were added into LEPOR in WMT13.). The basic model factors and further development with parameters setting were described in the paper (Han et al., 2012) and (Han et al., 2013).

For sentence-level score, only HLEPORA_BASELINE was submitted with scores calculated as the weighted harmonic mean of all the designed factors using default parameters.

For system-level score, both HLEPORA_BASELINE and HLEPORB_BASELINE were submitted, where HLEPORA_BASELINE is the the average score of all sentence-level scores, and HLEPORB_BASELINE is calculated via the same sentence-level hLEPOR equation but replacing each factor value with its system-level counterpart.

4.7 Meteor++_2.0 (syntax), Meteor++_2.0 (syntax+copy)

METEOR++ 2.0 (Guo and Hu, 2019) is a metric based on Meteor (Denkowski and Lavie, 2014) that takes syntactic-level paraphrase knowledge into consideration, where paraphrases may sometimes be skip-grams. i.e. (protect...from, protect...against). As the original Meteor-based metrics only pay attention to consecutive string matching,

they perform badly when reference-hypothesis pairs contain skip n-gram paraphrases. METEOR++ 2.0 extracts the knowledge from the Paraphrase Database (PPDB; Bannard and Callison-Burch, 2005) and integrates it into Meteor-based metrics.

4.8 PReP

PREP (Yoshimura et al., 2019) is a method for filtering pseudo-references to achieve a good match with a gold reference.

At the beginning, the source sentence is translated with some off-the-shelf MT systems to create a set of pseudo-references. (Here the MT systems were Google Translate and Microsoft Bing Translator.) The pseudo-references are then filtered using BERT (Devlin et al., 2019) fine-tuned on the MPRC corpus (Dolan and Brockett, 2005), estimating the probability of the paraphrase between gold reference and pseudo-references. Thanks to the high quality of the underlying MT systems, a large portion of their outputs is indeed considered as a valid paraphrase.

The final metric score is calculated simply with SentBLEU with these multiple references.

4.9 WMDO

WMDO (Chow et al., 2019b) is a metric based on distance between distributions in the semantic vector space. Matching in the semantic space has been investigated for translation evaluation, but the constraints of a translation's word order have not been fully explored. Building on the Word Mover's Distance metric and various word embeddings, WMDO introduces a fragmentation penalty to account for fluency of a translation. This word order extension is shown to perform better than standard WMD, with promising results against other types of metrics.

4.10 YiSi-0, YiSi-1, YiSi-1_srl, YiSi-2, YiSi-2_srl

YiSi (Lo, 2019) is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources.

YiSi-1 is a MT evaluation metric that measures the semantic similarity between a machine translation and human references by aggregating the idf-weighted lexical semantic similarities based on the contextual embeddings extracted from BERT and optionally incorporating shallow semantic structures (denoted as YiSi-1_srl).

YiSi-0 is the degenerate version of YiSi-1 that is ready-to-deploy to any language. It uses longest common character substring to measure the lexical similarity.

YiSi-2 is the bilingual, reference-less version for MT quality estimation, which uses the contextual embeddings extracted from BERT to evaluate the crosslingual lexical semantic similarity between the input and MT output. Like YiSi-1, YiSi-2 can exploit shallow semantic structures as well (denoted as YiSi-2_srl).

4.11 QE Systems

In addition to the submitted standard metrics, 10 quality estimation systems were submitted to the "QE as a Metric" track. The submitted QE systems are evaluated in the same settings as metrics to facilitate comparison. Their descriptions can be found in the Findings of the WMT 2019 Shared Task on Quality Estimation (Fonseca et al., 2019).

5 Results

We discuss system-level results for news task systems in Section 5.1. The segment-level results are in Section 5.2.

5.1 System-Level Evaluation

As in previous years, we employ the Pearson correlation (r) as the main evaluation measure for system-level metrics. The Pearson correlation is as follows:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \quad (1)$$

where H_i are human assessment scores of all systems in a given translation direction, M_i are the corresponding scores as predicted by a given metric. $\overline{H}$ and $\overline{M}$ are their means, respectively.

Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human assessment, while error metrics, such as TER, aim for a strong negative correlation we compare metrics via the absolute value $|r|$ of a

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
n	16	12	11	11	11	14	15
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	0.906	**0.993**	0.952	0.986	0.947	0.915	0.942
BERTr	**0.926**	0.984	0.938	0.990	0.948	**0.971**	0.974
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER	0.890	**0.988**	0.876	0.967	**0.975**	0.892	0.917
CharacTER	0.898	**0.990**	0.922	0.953	0.955	0.923	0.943
chrF	**0.917**	**0.992**	0.955	0.978	0.940	0.945	0.956
chrF+	**0.916**	**0.992**	0.947	0.976	0.940	0.945	0.956
EED	0.903	**0.994**	0.976	0.980	0.929	0.950	0.949
ESIM	**0.941**	0.971	0.885	0.986	**0.989**	**0.968**	**0.988**
hLEPORA_baseline	–	–	–	0.975	–	–	0.947
hLEPORB_baseline	–	–	–	0.975	0.906	–	0.947
Meteor++_2.0(syntax)	0.887	**0.995**	0.909	0.974	0.928	**0.950**	0.948
Meteor++_2.0(syntax+copy)	0.896	**0.995**	0.900	0.971	0.927	**0.952**	0.952
NIST	0.813	0.986	0.930	0.942	0.944	0.925	0.921
PER	0.883	**0.991**	0.910	0.737	0.947	0.922	0.952
PReP	0.575	0.614	0.773	0.776	0.494	0.782	0.592
sacreBLEU.BLEU	0.813	0.985	0.834	0.946	0.955	0.873	0.903
sacreBLEU.chrF	0.910	**0.990**	0.952	0.969	0.935	0.919	0.955
TER	0.874	**0.984**	0.890	0.799	0.960	0.917	0.840
WER	0.863	0.983	0.861	0.793	0.961	0.911	0.820
WMDO	0.872	**0.987**	0.983	**0.998**	0.900	0.942	0.943
YiSi-0	0.902	**0.993**	**0.993**	0.991	0.927	**0.958**	0.937
YiSi-1	**0.949**	**0.989**	0.924	0.994	0.981	**0.979**	**0.979**
YiSi-1_srl	**0.950**	**0.989**	0.918	0.994	**0.983**	**0.978**	0.977
QE as a Metric:							
ibm1-morpheme	0.345	0.740	–	–	0.487	–	–
ibm1-pos4gram	0.339	–	–	–	–	–	–
LASIM	0.247	–	–	–	–	0.310	–
LP	0.474	–	–	–	–	0.488	–
UNI	0.846	0.930	–	–	–	0.805	–
UNI+	0.850	0.924	–	–	–	0.808	–
YiSi-2	0.796	0.642	0.566	0.324	0.442	0.339	0.940
YiSi-2_srl	0.804	–	–	–	–	–	0.947
			newstest2019				

Table 3: Absolute Pearson correlation of to-English system-level metrics with DA human assessment in newstest2019; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	en-cs	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**
n	11	22	12	11	11	12	12	12
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.990**	0.983	**0.989**	0.829	0.971	**0.982**	0.977	0.803
BLEU	0.897	0.921	**0.969**	0.737	0.852	**0.989**	0.986	0.901
CDER	0.985	0.973	**0.978**	0.840	0.927	**0.985**	**0.993**	0.905
CHARACTER	**0.994**	**0.986**	0.968	**0.910**	0.936	0.954	**0.985**	0.862
CHRF	0.990	0.979	**0.986**	**0.841**	**0.972**	**0.981**	0.943	0.880
CHRF+	**0.991**	0.981	**0.986**	**0.848**	**0.974**	**0.982**	0.950	0.879
EED	**0.993**	**0.985**	**0.987**	**0.897**	**0.979**	0.975	0.967	0.856
ESIM	–	**0.991**	0.957	–	**0.980**	**0.989**	**0.989**	0.931
HLEPORA_BASELINE	–	–	–	0.841	0.968	–	–	–
HLEPORB_BASELINE	–	–	–	0.841	0.968	0.980	–	–
NIST	0.896	0.321	0.971	0.786	0.930	**0.993**	**0.988**	0.884
PER	0.976	0.970	**0.982**	0.839	0.921	0.985	0.981	0.895
SACREBLEU.BLEU	**0.994**	0.969	**0.966**	0.736	0.852	**0.986**	0.977	0.801
SACREBLEU.CHRF	0.983	0.976	0.980	0.841	**0.967**	0.966	**0.985**	0.796
TER	0.980	0.969	**0.981**	**0.865**	0.940	**0.994**	**0.995**	0.856
WER	0.982	0.966	**0.980**	**0.861**	0.939	**0.991**	**0.994**	0.875
YISI-0	**0.992**	0.985	**0.987**	0.863	0.974	0.974	0.953	0.861
YISI-1	0.962	**0.991**	0.971	**0.909**	**0.985**	0.963	**0.992**	**0.951**
YISI-1_SRL	–	**0.991**	–	–	–	–	–	**0.948**
QE as a Metric:								
IBM1-MORPHEME	0.871	0.870	0.084	–	–	0.810	–	–
IBM1-POS4GRAM	–	0.393	–	–	–	–	–	–
LASIM	–	0.871	–	–	–	–	0.823	–
LP	–	0.569	–	–	–	–	0.661	–
UNI	0.028	0.841	0.907	–	–	–	0.919	–
UNI+	–	–	–	–	–	–	0.918	–
USFD	–	0.224	–	–	–	–	0.857	–
USFD-TL	–	0.091	–	–	–	–	0.771	–
YISI-2	0.324	0.924	0.696	0.314	0.339	0.055	0.766	0.097
YISI-2_SRL	–	0.936	–	–	–	–	–	0.118
			newstest2019					

Table 4: Absolute Pearson correlation of out-of-English system-level metrics with DA human assessment in newstest2019; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

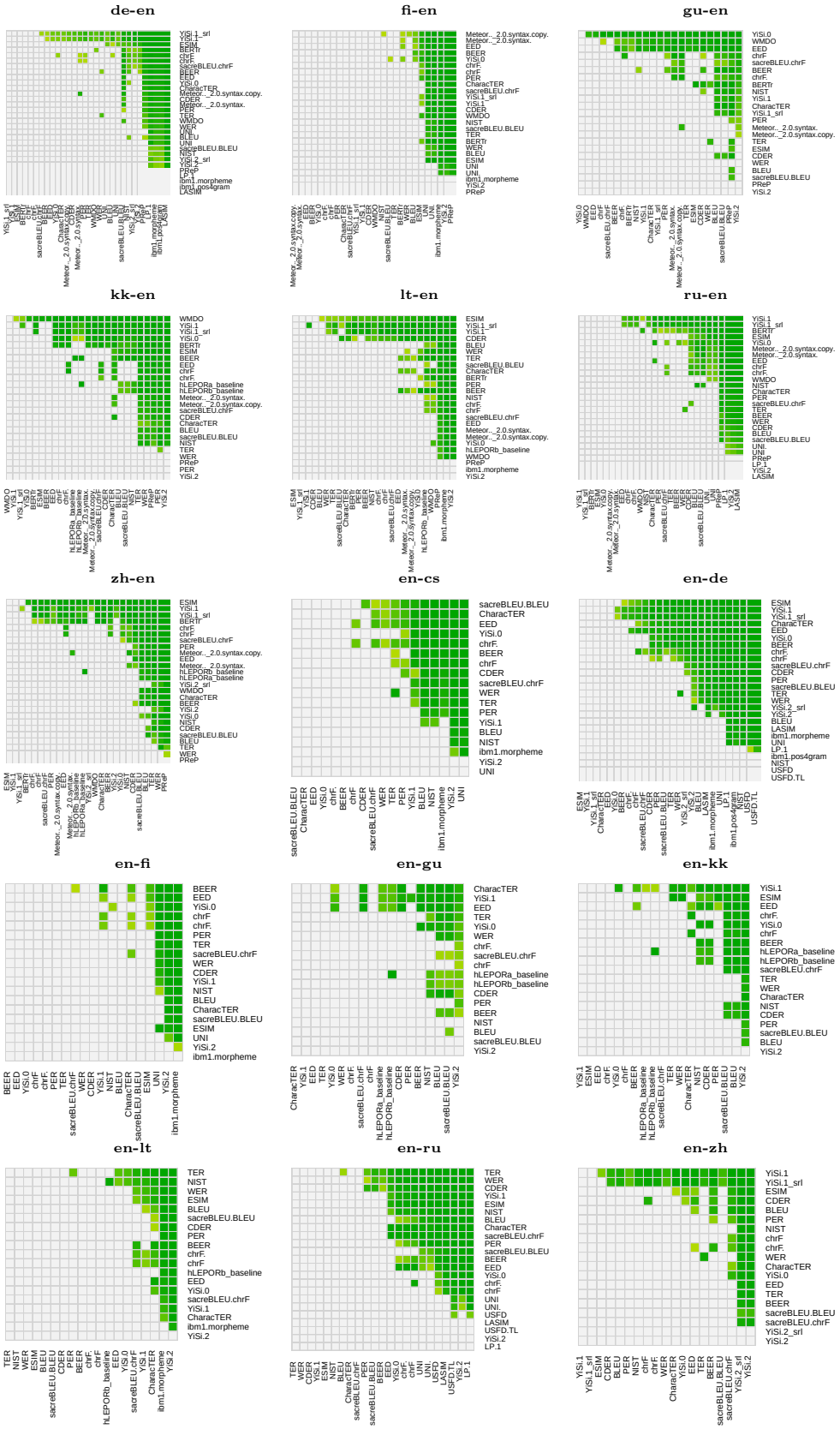


Figure 1: System-level metric significance test results for DA human assessment for into English and out-of English language pairs (newstest2019): Green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

given metric's correlation with human assessment.

5.1.1 System-Level Results

Tables 3, 4 and 5 provide the system-level correlations of metrics evaluating translation of newstest2019. The underlying texts are part of the WMT19 News Translation test set (newstest2019) and the underlying MT systems are all MT systems participating in the WMT19 News Translation Task.

As recommended by Graham and Baldwin (2014), we employ Williams significance test (Williams, 1959) to identify differences in correlation that are statistically significant. Williams test is a test of significance of a difference in dependent correlations and therefore suitable for evaluation of metrics. Correlations not significantly outperformed by any other metric for the given language pair are highlighted in bold in Tables 3, 4 and 5.

Since pairwise comparisons of metrics may be also of interest, e.g. to learn which metrics significantly outperform the most widely employed metric BLEU, we include significance test results for every competing pair of metrics including our baseline metrics in Figure 1 and Figure 2.

This year, the increased number of systems participating in the news tasks has provided a larger sample of system scores for testing metrics. Since we already have sufficiently conclusive results on genuine MT systems, we do not need to generate hybrid system results as in Graham and Liu (2016) and past metrics tasks.

5.2 Segment-Level Evaluation

Segment-level evaluation relies on the manual judgements collected in the News Translation Task evaluation. This year, again we were unable to follow the methodology outlined in Graham et al. (2015) for evaluation of segment-level metrics because the sampling of sentences did not provide sufficient number of assessments of the same segment. We therefore convert pairs of DA scores for competing translations to daRR better/worse preferences as described in Section 2.3.2.

We measure the quality of metrics' segment-level scores against the daRR golden truth using a Kendall's Tau-like formulation, which is an adaptation of the conventional Kendall's Tau coefficient. Since we do not have a total order ranking of all translations, it is not possible to apply conventional Kendall's Tau (Graham et al., 2015).

Our Kendall's Tau-like formulation, τ, is as follows:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (2)$$

where *Concordant* is the set of all human comparisons for which a given metric suggests the same order and *Discordant* is the set of all human comparisons for which a given metric disagrees. The formula is not specific with respect to ties, i.e. cases where the annotation says that the two outputs are equally good.

The way in which ties (both in human and metric judgement) were incorporated in computing Kendall τ has changed across the years of WMT Metrics Tasks. Here we adopt the version used in WMT17 daRR evaluation. For a detailed discussion on other options, see also Macháček and Bojar (2014).

Whether or not a given comparison of a pair of distinct translations of the same source input, s_1 and s_2, is counted as a concordant (Conc) or disconcordant (Disc) pair is defined by the following matrix:

		Metric		
		$s_1 < s_2$	$s_1 = s_2$	$s_1 > s_2$
Human	$s_1 < s_2$	Conc	Disc	Disc
	$s_1 = s_2$	–	–	–
	$s_1 > s_2$	Disc	Disc	Conc

In the notation of Macháček and Bojar (2014), this corresponds to the setup used in WMT12 (with a different underlying method of manual judgements, RR):

	WMT12	Metric <	=	>
Human	<	1	-1	-1
	=	X	X	X
	>	-1	-1	1

The key differences between the evaluation used in WMT14–WMT16 and evaluation used in WMT17–WMT19 were (1) the move from RR to daRR and (2) the treatment of ties. In the years 2014-2016, ties in metrics scores were not penalized. With the move to daRR, where the quality of the two candidate translations

	de-cs	de-fr	fr-de
n	11	11	10
Correlation	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.978**	**0.941**	0.848
BLEU	**0.941**	0.891	0.864
CDER	0.864	**0.949**	0.852
CHARACTER	0.965	0.928	0.849
CHRF	**0.974**	0.931	0.864
CHRF+	0.972	0.936	0.848
EED	**0.982**	**0.940**	0.851
ESIM	**0.980**	**0.950**	**0.942**
HLEPORA_BASELINE	0.941	0.814	–
HLEPORB_BASELINE	**0.959**	0.814	–
NIST	**0.954**	**0.916**	0.862
PER	0.875	0.857	**0.899**
SACREBLEU-BLEU	0.869	0.891	0.869
SACREBLEU-CHRF	**0.975**	**0.952**	0.882
TER	0.890	**0.956**	**0.895**
WER	0.872	**0.956**	**0.894**
YISI-0	**0.978**	**0.952**	0.820
YISI-1	0.973	**0.969**	**0.908**
YISI-1_SRL	–	–	**0.912**
QE as a Metric:			
IBM1-MORPHEME	0.355	0.509	0.625
IBM1-POS4GRAM	–	0.085	0.478
YISI-2	0.606	0.721	0.530
	newstest2019		

Table 5: Absolute Pearson correlation of system-level metrics for language pairs not involving English with DA human assessment in newstest2019; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

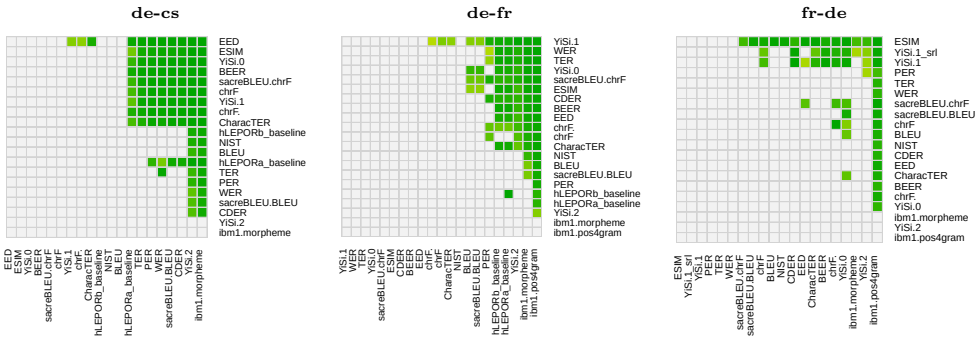


Figure 2: System-level metric significance test results for DA human assessment in newstest2019 for German to Czech, German to French and French to German; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	85,365	38,307	31,139	27,094	21,862	46,172	31,070
BEER	0.128	0.283	0.260	0.421	0.315	0.189	0.371
BERTr	0.142	0.331	0.291	0.421	0.353	0.195	0.399
CharacTER	0.101	0.253	0.190	0.340	0.254	0.155	0.337
chrF	0.122	0.286	0.256	0.389	0.301	0.180	0.371
chrF+	0.125	0.289	0.257	0.394	0.303	0.182	0.374
EED	0.120	0.281	0.264	0.392	0.298	0.176	0.376
ESIM	0.167	**0.337**	**0.303**	**0.435**	0.359	0.201	0.396
hLEPORA_baseline	–	–	–	0.372	–	–	0.339
Meteor++_2.0(syntax)	0.084	0.274	0.237	0.395	0.291	0.156	0.370
Meteor++_2.0(syntax+copy)	0.094	0.273	0.244	0.402	0.287	0.163	0.367
PReP	0.030	0.197	0.192	0.386	0.193	0.124	0.267
sentBLEU	0.056	0.233	0.188	0.377	0.262	0.125	0.323
WMDO	0.096	0.281	0.260	0.420	0.300	0.162	0.362
YiSi-0	0.117	0.271	0.263	0.402	0.289	0.178	0.355
YiSi-1	0.164	**0.347**	**0.312**	**0.440**	**0.376**	**0.217**	**0.426**
YiSi-1_srl	**0.199**	**0.346**	**0.306**	**0.442**	**0.380**	**0.222**	**0.431**
QE as a Metric:							
ibm1-morpheme	−0.074	0.009	–	–	0.069	–	–
ibm1-pos4gram	−0.153	–	–	–	–	–	–
LASIM	−0.024	–	–	–	–	0.022	–
LP	−0.096	–	–	–	–	−0.035	–
UNI	0.022	0.202	–	–	–	0.084	–
UNI+	0.015	0.211	–	–	–	0.089	–
YiSi-2	0.068	0.126	−0.001	0.096	0.075	0.053	0.253
YiSi-2_srl	0.068	–	–	–	–	–	0.246
			newstest2019				

Table 6: Segment-level metric results for to-English language pairs in newstest2019: absolute Kendall's Tau formulation of segment-level metric scores with DA scores; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	27,178	99,840	31,820	11,355	18,172	17,401	24,334	18,658
BEER	0.443	0.316	0.514	0.537	0.516	0.441	0.542	0.232
CHARACTER	0.349	0.264	0.404	0.500	0.351	0.311	0.432	0.094
CHRF	0.455	0.326	0.514	0.534	0.479	0.446	0.539	0.301
CHRF+	**0.458**	0.327	0.514	0.538	0.491	**0.448**	0.543	0.296
EED	0.431	0.315	0.508	**0.568**	0.518	0.425	0.546	0.257
ESIM	–	0.329	0.511	–	0.510	0.428	**0.572**	**0.339**
HLEPORA_BASELINE	–	–	–	0.463	0.390	–	–	–
SENTBLEU	0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270
YISI-0	0.406	0.304	0.483	0.539	0.494	0.402	0.535	0.266
YISI-1	**0.475**	0.351	**0.537**	**0.551**	**0.546**	**0.470**	**0.585**	**0.355**
YISI-1_SRL	–	**0.368**	–	–	–	–	–	**0.361**
QE as a Metric:								
IBM1-MORPHEME	−0.135	−0.003	−0.005	–	–	−0.165	–	–
IBM1-POS4GRAM	–	−0.123	–	–	–	–	–	–
LASIM	–	0.147	–	–	–	–	−0.24	–
LP	–	−0.119	–	–	–	–	−0.158	–
UNI	0.060	0.129	0.351	–	–	–	0.226	–
UNI+	–	–	–	–	–	–	0.222	–
USFD	–	−0.029	–	–	–	–	0.136	–
USFD-TL	–	−0.037	–	–	–	–	0.191	–
YISI-2	0.069	0.212	0.239	0.147	0.187	0.003	−0.155	0.044
YISI-2_SRL	–	0.236	–	–	–	–	–	0.034
			newstest2019					

Table 7: Segment-level metric results for out-of-English language pairs in newstest2019: absolute Kendall's Tau formulation of segment-level metric scores with DA scores; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	de-cs	de-fr	fr-de
Human Evaluation	DARR	DARR	DARR
n	35,793	4,862	1,369
BEER	0.337	0.293	**0.265**
CHARACTER	0.232	0.251	**0.224**
CHRF	0.326	0.284	**0.275**
CHRF+	0.326	0.284	**0.278**
EED	0.345	0.301	**0.267**
ESIM	0.331	0.290	**0.289**
HLEPORA_BASELINE	0.207	0.239	–
SENTBLEU	0.203	0.235	0.179
YISI-0	0.331	0.296	**0.277**
YISI-1	**0.376**	**0.349**	**0.310**
YISI-1_SRL	–	–	**0.299**
QE as a Metric:			
IBM1-MORPHEME	0.048	−0.013	−0.053
IBM1-POS4GRAM	–	−0.074	−0.097
YISI-2	0.199	0.186	0.066
	newstest2019		

Table 8: Segment-level metric results for language pairs not involving English in newstest2019: absolute Kendall's Tau formulation of segment-level metric scores with DA scores; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

is deemed substantially different and no ties in human judgements arise, it makes sense to penalize ties in metrics' predictions in order to promote discerning metrics.

Note that the penalization of ties makes our evaluation asymmetric, dependent on whether the metric predicted the tie for a pair where humans predicted $<$, or $>$. It is now important to interpret the meaning of the comparison identically for humans and metrics. For error metrics, we thus reverse the sign of the metric score prior to the comparison with human scores: higher scores have to indicate better translation quality. In WMT19, the original authors did this for CharacTER.

To summarize, the WMT19 Metrics Task for segment-level evaluation:

- ensures that error metrics are first converted to the same orientation as the human judgements, i.e. higher score indicating higher translation quality,
- excludes all human ties (this is already implied by the construction of DARR from DA judgements),

Figure 3: DARR segment-level metric significance test results for into English and out-of English language pairs (newstest2019): Green cells denote a significant win for the metric in a given row over the metric in a given column according bootstrap resampling.

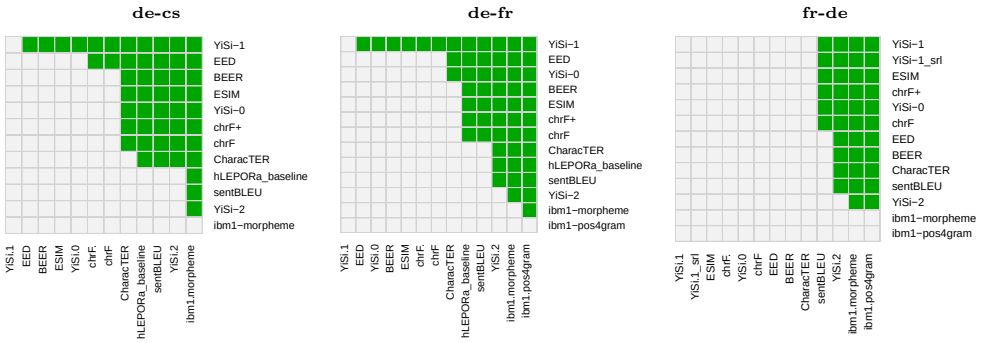


Figure 4: DARR segment-level metric significance test results for German to Czech, German to French and French to German (newstest2019): Green cells denote a significant win for the metric in a given row over the metric in a given column according bootstrap resampling.

- counts metric's ties as a *Discordant* pairs.

We employ bootstrap resampling (Koehn, 2004; Graham et al., 2014b) to estimate confidence intervals for our Kendall's Tau formulation, and metrics with non-overlapping 95% confidence intervals are identified as having statistically significant difference in performance.

5.2.1 Segment-Level Results

Results of the segment-level human evaluation for translations sampled from the News Translation Task are shown in Tables 6, 7 and 8, where metric correlations not significantly outperformed by any other metric are highlighted in bold. Head-to-head significance test results for differences in metric performance are included in Figures 3 and 4.

6 Discussion

This year, human data was collected from reference-based evaluations (or "monolingual") and reference-free evaluations (or "bilingual"). The reference-based (monolingual) evaluations were obtained with the help of anonymous crowdsourcing, while the reference-less (bilingual) evaluations were mainly from MT researchers who committed their time contribution to the manual evaluation for each submitted system.

6.1 Stability across MT Systems

The observed performance of metrics depends on the underlying texts and systems that participate in the News Translation Task (see Section 2). For the strongest MT systems, distinguishing which system outputs are better is

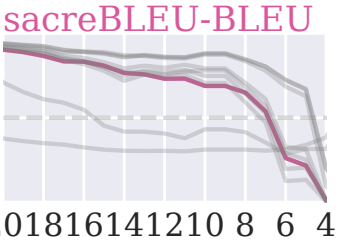


Figure 5: Pearson correlations of SACREBLEU-BLEU for English-German system-level evaluation for all systems (left) down to only top 4 systems (right). The y-axis spans from -1 to +1, baseline metrics for the language pair in grey.

hard, even for human assessors. On the other hand, if the systems are spread across a wide performance range, it will be easier for metrics to correlate with human judgements.

To provide a more reliable view, we created plots of Pearson correlation when the underlying set of MT systems is reduced to top n ones. One sample such plot is in Figure 5, all language pairs and most of the metrics are in Appendix A.

As the plot documents, the official correlations reported in Tables 3 to 5 can lead to wrong conclusions. SACREBLEU-BLEU correlates at .969 when all systems are considered, but as we start considering only the top n systems, the correlation falls relatively quickly. With 10 systems, we are below .5 and when only the top 6 or 4 systems are considered, the correlation falls even to the negave values. Note that correlations point estimates (the value in the y-axis) become noiser with the decreasing number of the underlying MT systems.

Figure 6 explains the situation and illus-

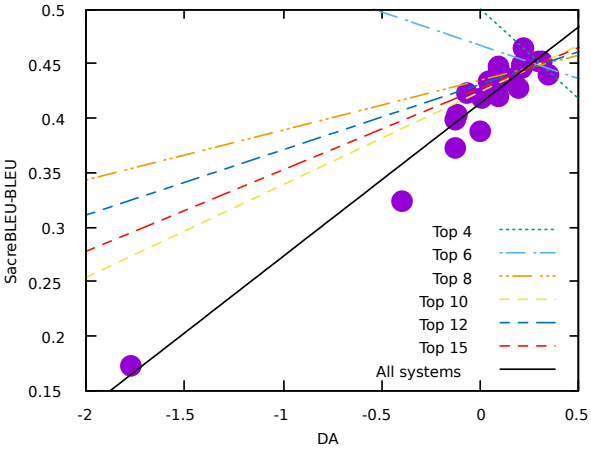


Figure 6

trates the sensitivity of the observed correlations to the exact set of systems. On the full set of systems, the single outlier (the worst-performing system called EN__DE__TASK) helps to achieve a great positive correlation. The majority of MT systems however form a cloud with Pearson correlation around .5 and the top 4 systems actually exhibit a negative correlation of the human score and SACREBLEU-BLEU.

In Appendix A, baseline metrics are plotted in grey in all the plots, so that their trends can be observed jointly. In general, most baselines have similar correlations, as most baselines use similar features (n-gram or word-level features, with the exception of CHRF). In a number of language pairs (de-en, de-fr, en-de, en-kk, lt-en, ru-en, zh-en), baseline correlations tend towards 0 (no correlation) or even negative Pearson correlation. For a widely applied metric such as SACREBLEU-BLEU, our analysis reveals weak correlation in comparing top state-of-the-art systems in these language pairs, especially in en-de, de-en, ru-en, and zh-en.

We will restrict our analysis to those language pairs where the baseline metrics have an obvious downward trend (de-en, de-fr, en-de, en-kk, lt-en, ru-en, zh-en). Examining the top-n correlation in the submitted metrics (not including QE systems), most metrics show the same degredation in correlation as the baselines. We note BERTR as the one exception consistently degrading less and retaining positive correlation compared to other submitted metrics and baselines, in the language pairs where it participated.

For QE systems, we noticed that in some instances, QE systems have upward correlation trends when other metrics and baselines have downward trends. For instance, LP, UNI, and UNI+ in the de-en language pair, YISI-2 in en-kk, and UNI and UNI+ in ru-en. These results suggest that QE systems such as UNI and UNI+ perform worse on judging systems of wide ranging quality, but better for top performing systems, or perhaps for systems closer in quality.

If our method of human assessment is sound, we should believe that BLEU, a widely applied metric, is no longer a reliable metric for judging our best systems. Future investigations are needed to understand when BLEU applies well, and why BLEU is not effective for output from our state of the art models.

Metrics and QE systems such as BERTR, ESIM, YISI that perform well at judging our best systems often use more semantic features compared to our n-gram/char-gram based baselines. Future metrics may want to explore a) whether semantic features such as contextual word embeddings are achieving semantic understanding and b) whether semantic understanding is the true source of a metric's performance gains.

It should be noted that *some* language pairs do not show the strong degrading pattern with top-n systems this year, for instance en-cs, en-gu, en-ru, or kk-en. English-Chinese is particularly interesting because we see a clear trend towards *better* correlations as we reduce the set of underlying systems to the top scoring ones.

6.2 Overall Metric Performance

6.2.1 System-Level Evaluation

In system-level evaluation, the series of YISI metrics achieve the highest correlations in several language pairs and it is not significantly outperformed by any other metrics (denoted as a "win" in the following) for almost all language pairs.

The new metric ESIM performs best on 5 language languages (18 language pairs) and obtains 11 "wins" out of 16 language pairs in which ESIM participated.

The metric EED performs better for language pairs out-of English and excluding En-

glish compared to into-English language pairs, achieving 7 out of 11 "wins" there.

6.2.2 Segment-Level Evaluation

For segment-level evaluation, most language pairs are quite discerning, with only one or two metrics taking the "winner" position (of not being significantly surpassed by others). Only French-German differs, with all metrics performing similarly except the significantly worse sentBLEU.

YiSi-1_srl stands out as the "winner" for all language pairs in which it participated. The excluded language pairs were probably due to the lack of semantic information required by YiSi-1_srl. YiSi-1 participated all language pairs and its correlations are comparable with those of YiSi-1_srl.

ESIM obtain 6 "winners" out of all 18 languages pairs.

Both YiSi and ESIM are based on neural networks (YiSi via word and phrase embeddings, as well as other types of available resources, ESIM via sentence embeddings). This is a confirmation of a trend observed last year.

6.2.3 QE Systems as Metrics

Generally, correlations for the standard reference-based metrics are obviously better than those in "QE as a Metric" track, both when using monolingual and bilingual golden truth.

In system-level evaluation, correlations for "QE as a Metric" range from 0.028 to 0.947 across all language pairs and all metrics but they are very unstable. Even for a single metric, take UNI for example, the correlations range from 0.028 to 0.930 across language pairs.

In segment-level evaluation, correlations for QE metrics range from -0.153 to 0.351 across all language pairs and show the same instability across language pairs for a given metric.

In either case, we do not see any pattern that could explain the behaviour, e.g. whether the manual evaluation was monolingual or bilingual, or the characteristics of the given language pair.

6.3 Dependence on Implementation

As it already happened in the past, we had multiple implementations for some metrics, BLEU and chrF in particular.

The detailed configuration of BLEU and sacreBLEU-BLEU differ and hence their scores and correlation results are different.

chrF and sacreBLEU-chrF use the same parameters and should thus deliver the same scores but we still observe some differences, leading to different correlations. For instance for German-French Pearson correlation, chrF obtains 0.931 (no win) but sacreBLEU-chrF reaches 0.952, tying for a win with other metrics.

We thus fully support the call for clarity by Post (2018b) and invite authors of metrics to include their implementations either in Moses scorer or sacreBLEU to achieve a long-term assessment of their metric.

7 Conclusion

This paper summarizes the results of WMT19 shared task in machine translation evaluation, the Metrics Shared Task. Participating metrics were evaluated in terms of their correlation with human judgement at the level of the whole test set (system-level evaluation), as well as at the level of individual sentences (segment-level evaluation).

We reported scores for standard metrics requiring the reference as well as quality estimation systems which took part in the track "QE as a metric", joint with the Quality Estimation task.

For system-level, best metrics reach over 0.95 Pearson correlation or better across several language pairs. As expected, QE systems are visibly in all language pairs but they can also reach high system-level correlations, up to .947 (Chinese-English) or .936 (English-German) by YiSi-1_srl or over .9 for multiple language pairs by UNI.

An important caveat is that the correlations are heavily affected by the underlying set of MT systems. We explored this by reducing the set of systems to top-n ones for various ns and found out that for many language pairs, system-level correlations are much worse when based on only the better performing systems. With both good and bad MT systems partic-

ipating in the news task, the metrics results can be overly optimistic compared to what we get when evaluating state-of-the-art systems.

In terms of segment-level Kendall's τ results, the standard metrics correlations varied between 0.03 and 0.59, and QE systems obtained even negative correlations.

The results confirm the observation from the last year, namely metrics based on word or sentence-level embeddings (YiSi and ESIM), achieve the highest performance.

Acknowledgments

Results in this shared task would not be possible without tight collaboration with organizers of the WMT News Translation Task. We would like to thank Marcin Junczys-Dowmunt for the suggestion to examine metrics performance across varying subsets of MT systems, as we did in Appendix A.

This study was supported in parts by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund, and Charles University Research Programme "Progres" Q18+Q48.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pages 27–34, Portorose, Slovenia.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Julian Chow, Pranava Madhyastha, and Lucia Specia. 2019a. Wmdo: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of Fourth Conference on Machine Translation*.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019b. WMDO: Fluency-based Word Mover's Distance for Machine Translation Evaluation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared

Task on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Yvette Graham and Timothy Baldwin. 2014. Testing for Significance of Increased Correlation with Human Judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014a. Is Machine Translation Getting Better over Time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Yvette Graham and Qun Liu. 2016. Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014b. Randomized significance tests in machine translation. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, pages 266–274. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate Evaluation of Segment-level Machine Translation Metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, Denver, Colorado.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt Syntactic Level Paraphrase Knowledge into Machine Translation Evaluation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Aaron L.-F. Han, Derek F. Wong, and Lidia S. Chao. 2012. Lepor: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 441–450. Association for Computational Linguistics.

Aaron L.-F. Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation Between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 102–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Mt Summit IX*, pages 240–247.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *In Proceedings of EACL*, pages 241–248.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD, USA. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceed-*

ings of the Eighth Workshop on Statistical Machine Translation, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proc. of ACL (short papers)*. To appear.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318.

Maja Popovic. 2012. Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT@NAACL-HLT 2012, June 7-8, 2012, Montréal, Canada*, pages 133–137.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018a. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matt Post. 2018b. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended Edit Distance Measure for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016a. Character: Translation edit rate on character level. In *ACL 2016 First Conference on Machine Translation*, pages 505–510, Berlin, Germany.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016b. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

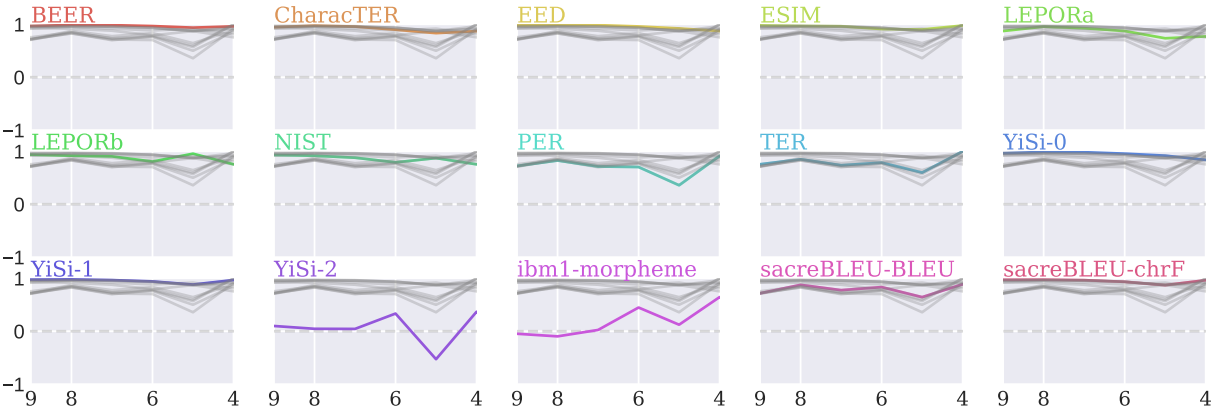
Evan James Williams. 1959. *Regression analysis*, volume 14. Wiley New York.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

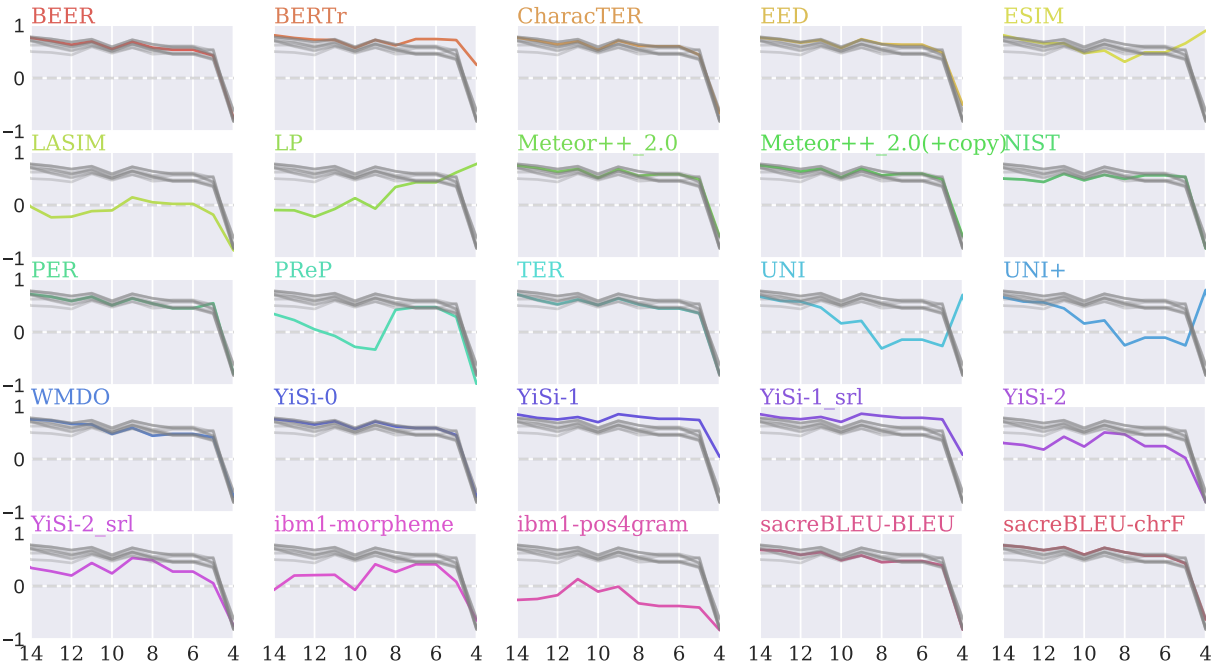
Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering Pseudo-References by Paraphrasing for Automatic Evaluation of Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

A Correlations for Top-N Systems

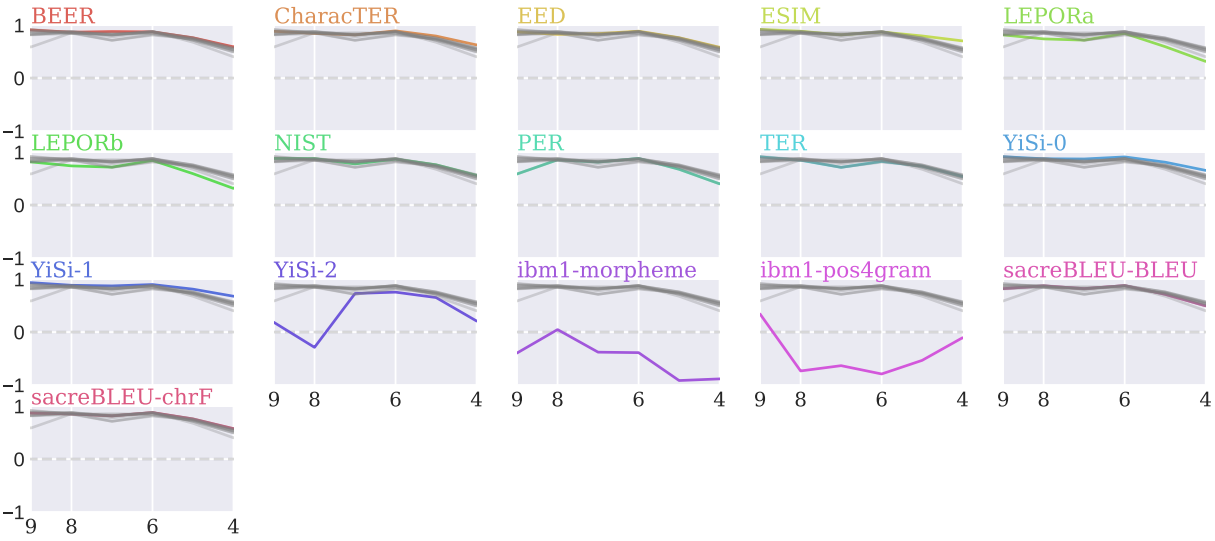
A.1 de-cs



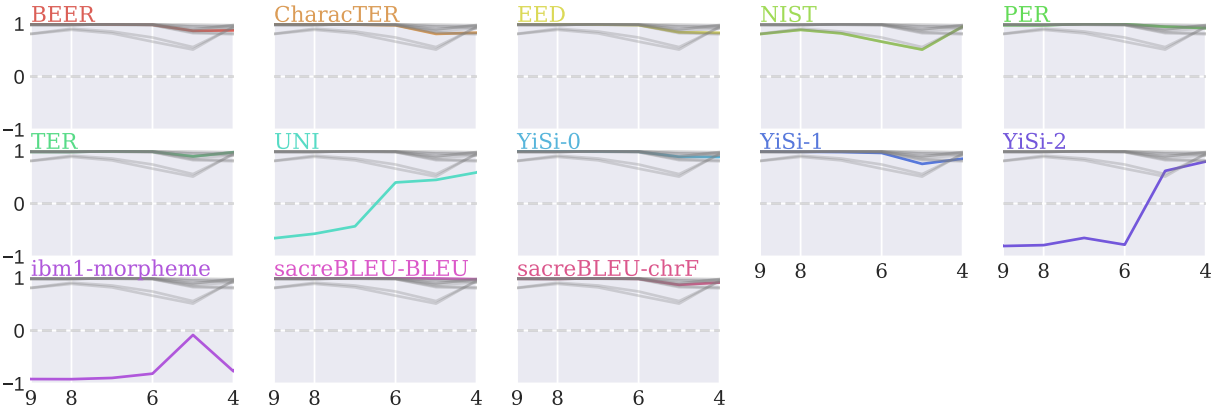
A.2 de-en



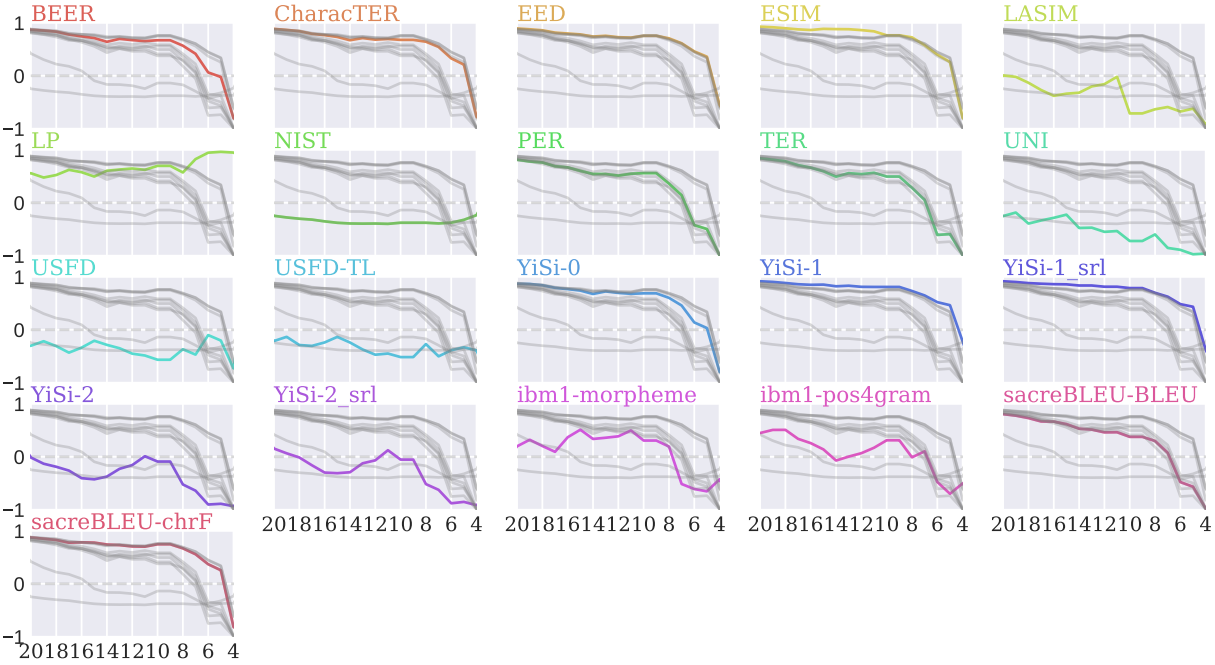
A.3 de-fr



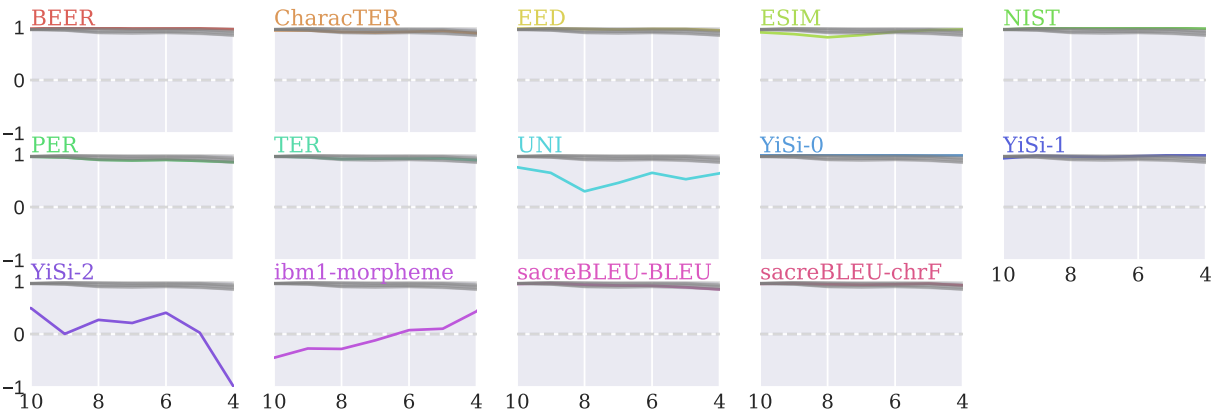
A.4 en-cs



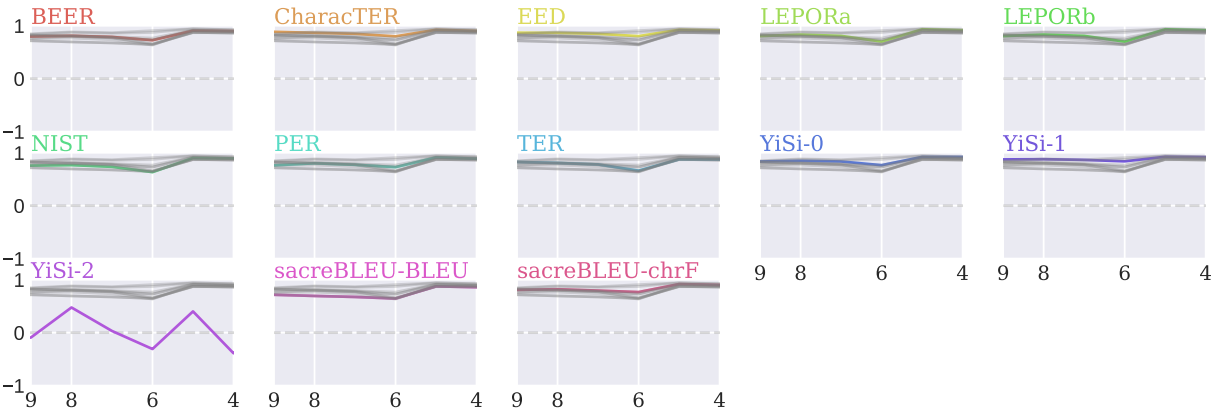
A.5 en-de



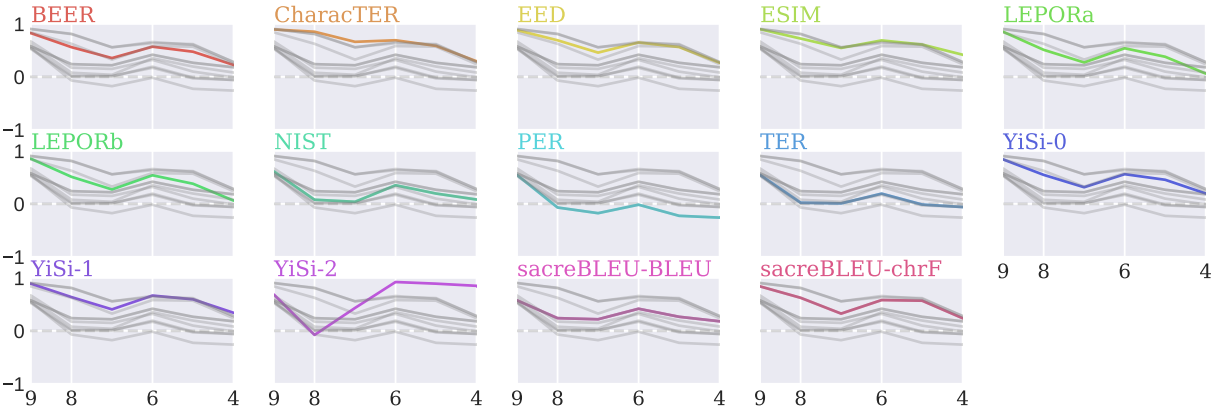
A.6 en-fi



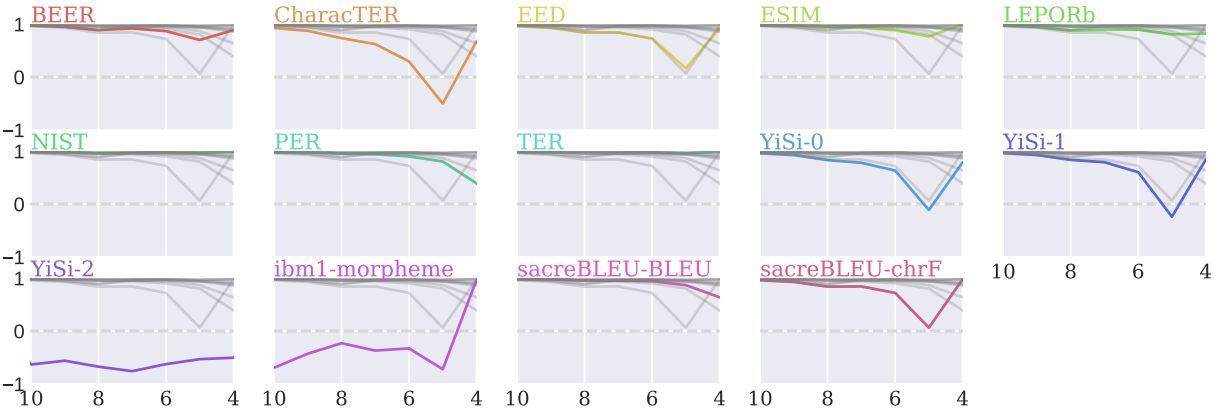
A.7 en-gu



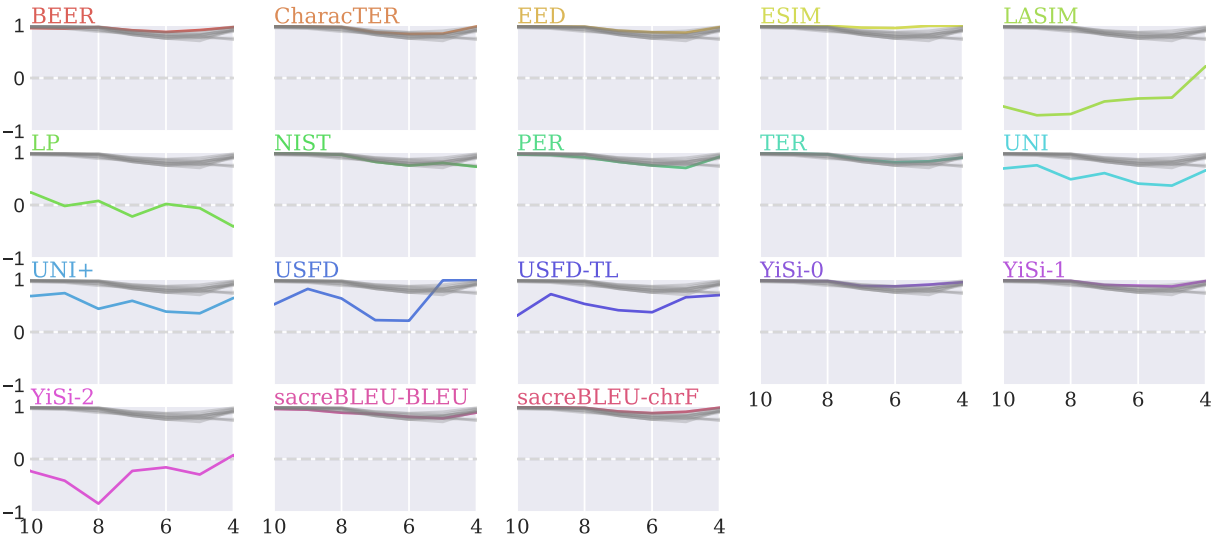
A.8 en-kk



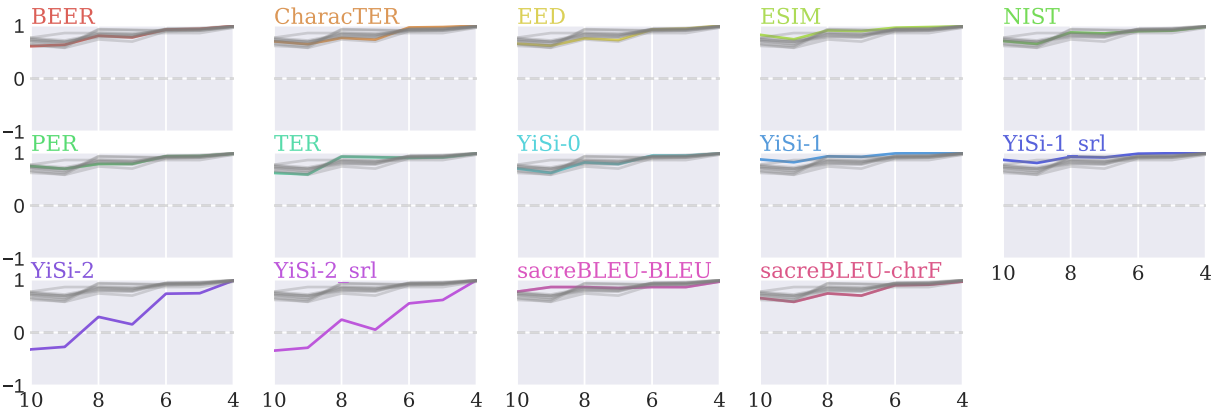
A.9 en-lt



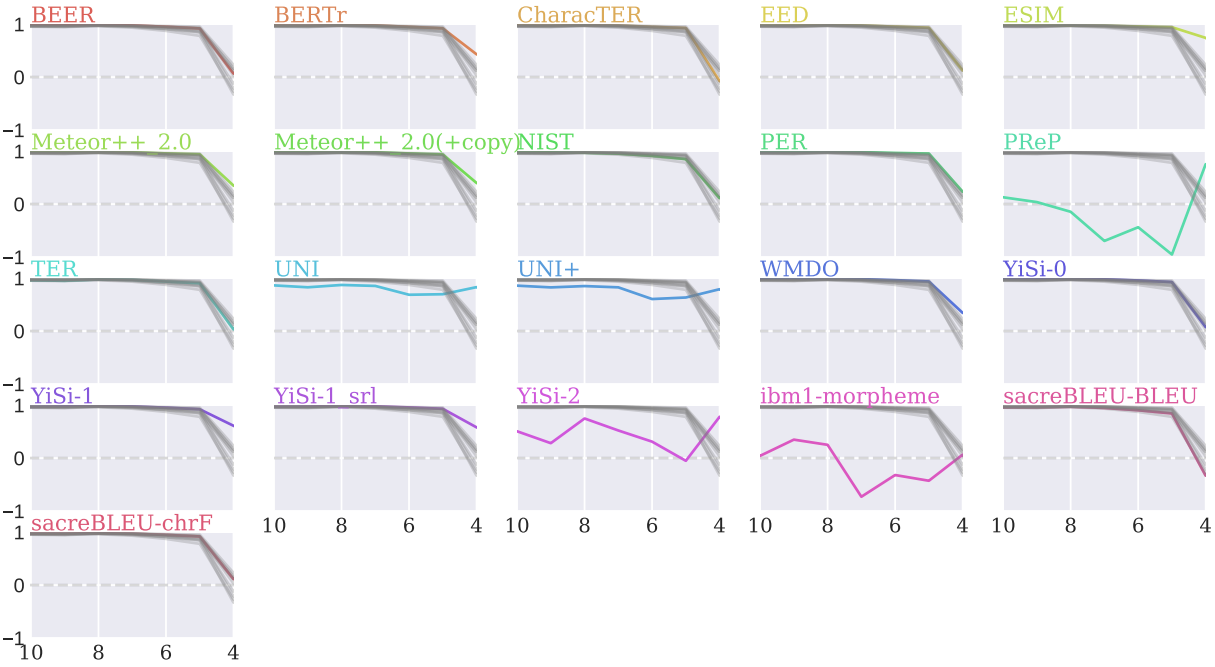
A.10 en-ru



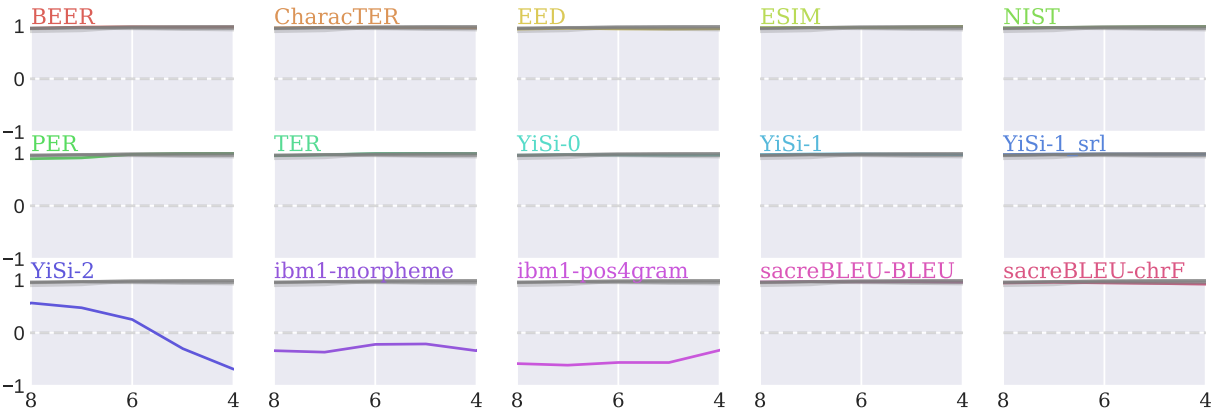
A.11 en-zh



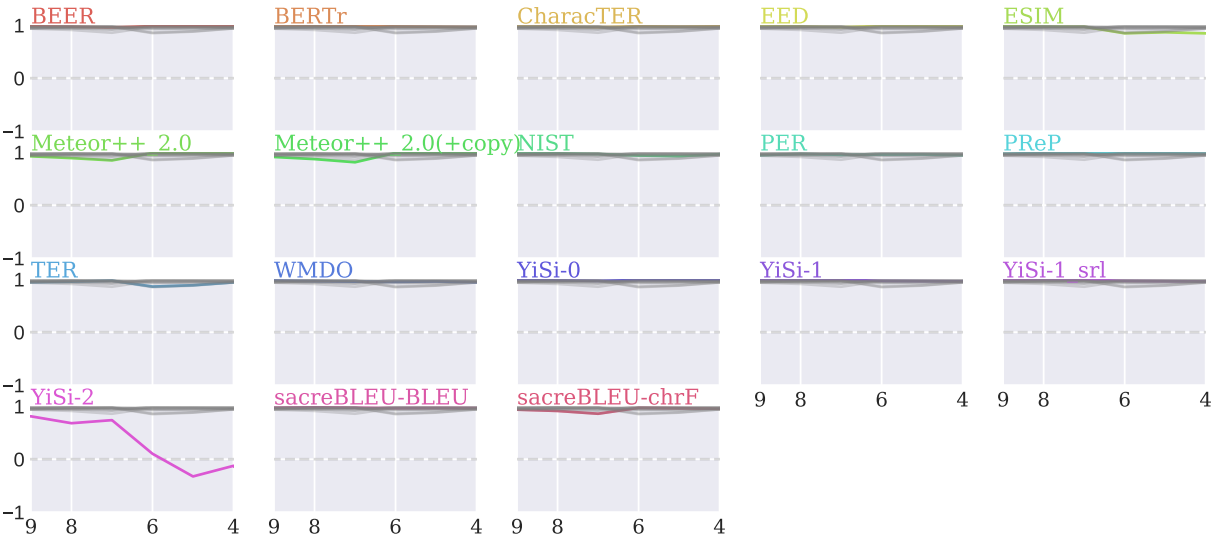
A.12 fi-en



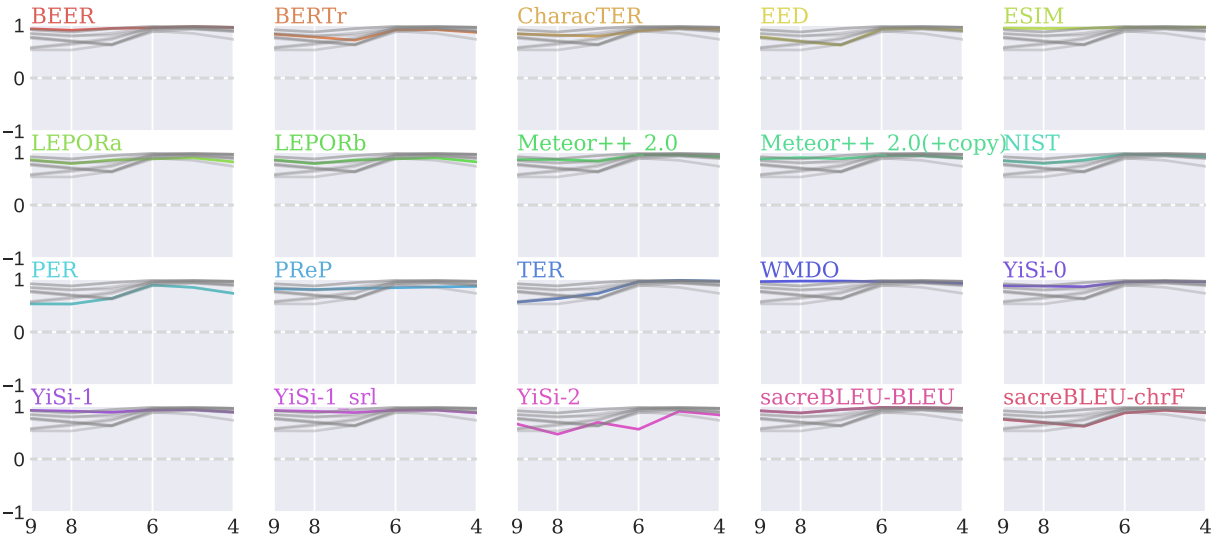
A.13 fr-de



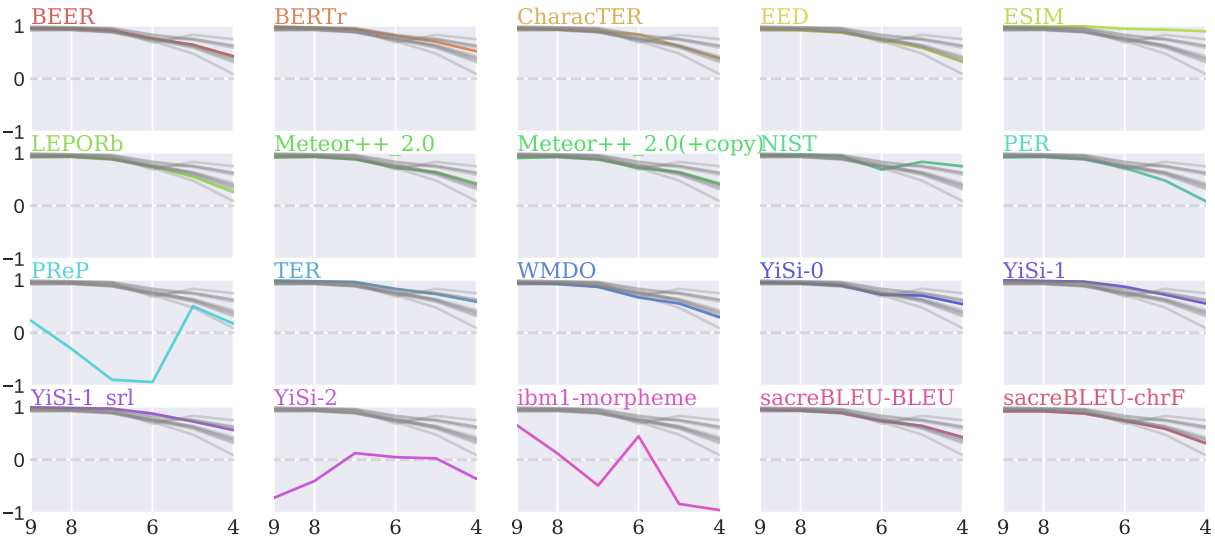
A.14 gu-en



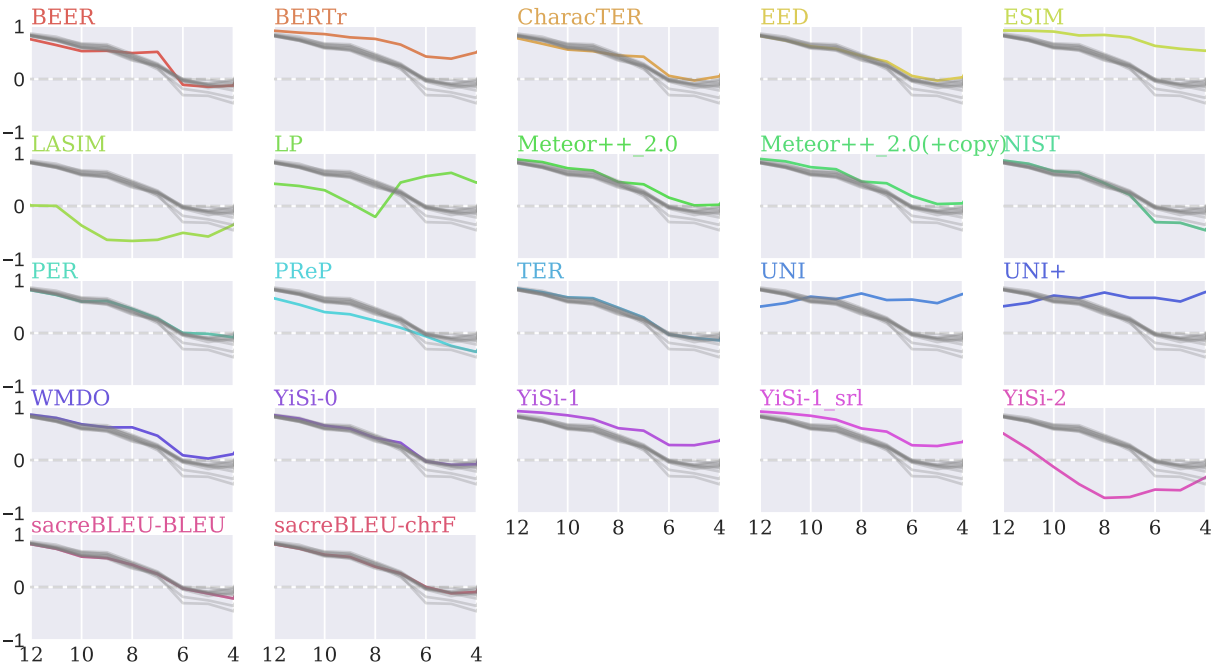
A.15 kk-en



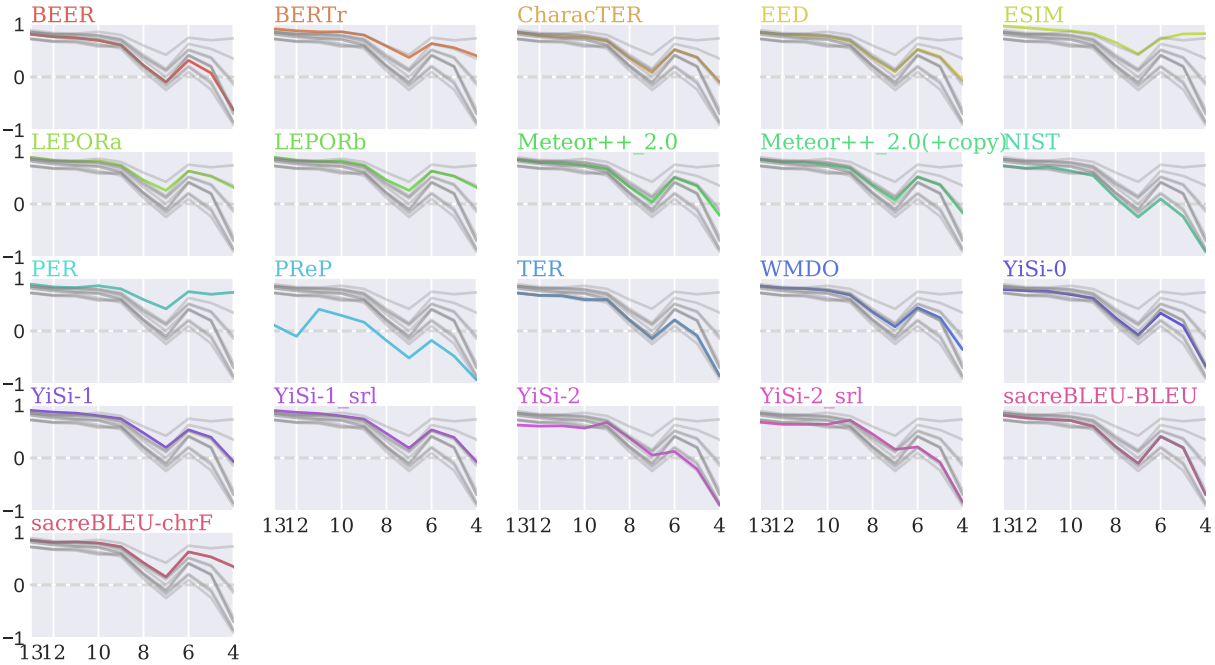
A.16 lt-en



A.17 ru-en



A.18 zh-en


BEER
BERTr
CharacTER
EED
ESIM
LEPORa
LEPORb
Meteor++_2.0
Meteor++_2.0(+copy)
NIST
PER
PReP
TER
WMDO
YiSi-0
YiSi-1
YiSi-1_srl
YiSi-2
YiSi-2_srl
sacreBLEU-BLEU
sacreBLEU-chrF
1
0
−1
13 12 10 8 6 4