Results of the WMT19 Metrics Shared Task

Segment-Level and Strong MT Systems Pose Big Challenges

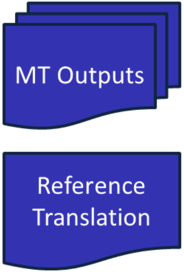
Qingsong Ma
Johnny Tian-Zheng Wei
Ondřej Bojar
Yvette Graham

Overview

- Overview of Metrics Task
- Updates to Metric Task in 2019
- Results in 2019

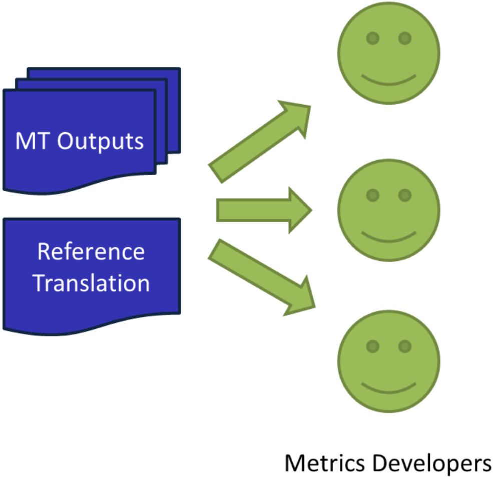
Metrics Task in a Nutshell

Metrics Task in a Nutshell


MT Outputs

Reference Translation

Metrics Task in a Nutshell

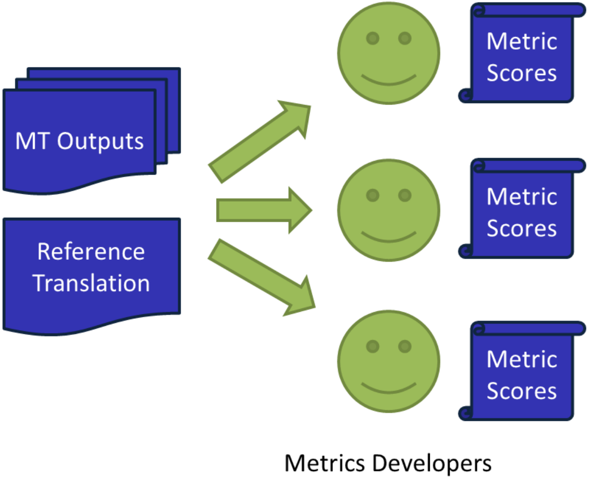


MT Outputs

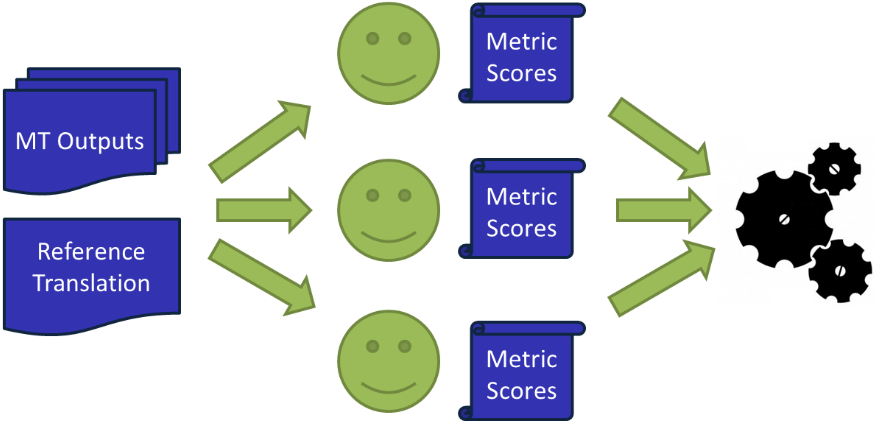
Reference Translation

Metrics Developers

Metrics Task in a Nutshell



Metrics Task in a Nutshell



MT Outputs

Reference Translation

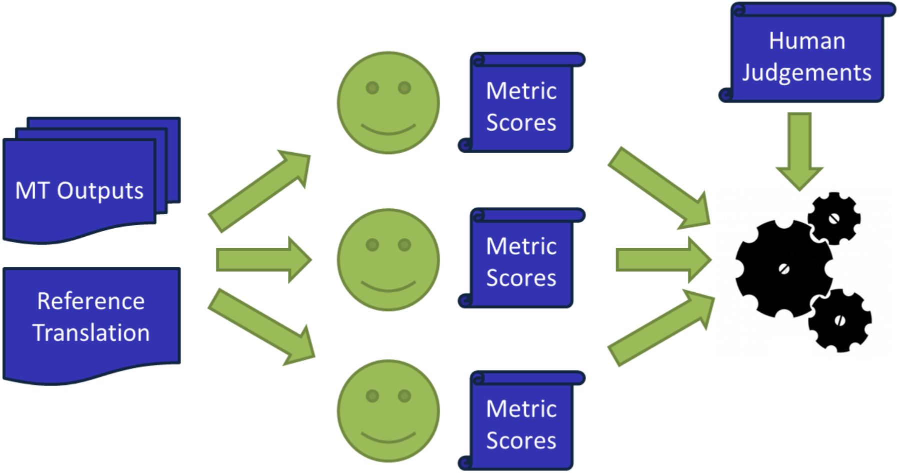
Metric Scores

Metric Scores

Metric Scores

Metrics Developers

Metrics Task in a Nutshell



MT Outputs

Reference Translation

Metric Scores

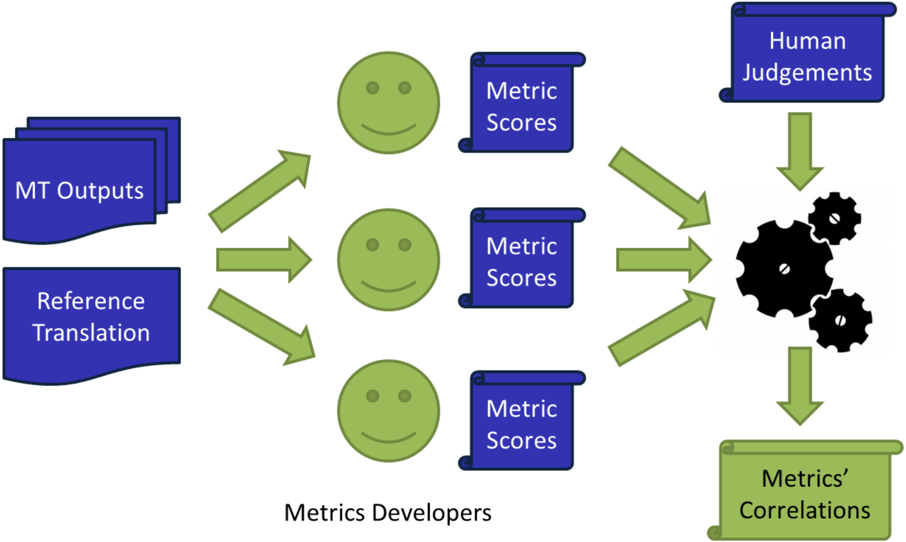
Metric Scores

Metric Scores

Human Judgements

Metrics Developers

Metrics Task in a Nutshell



MT Outputs

Reference Translation

Metric Scores

Metric Scores

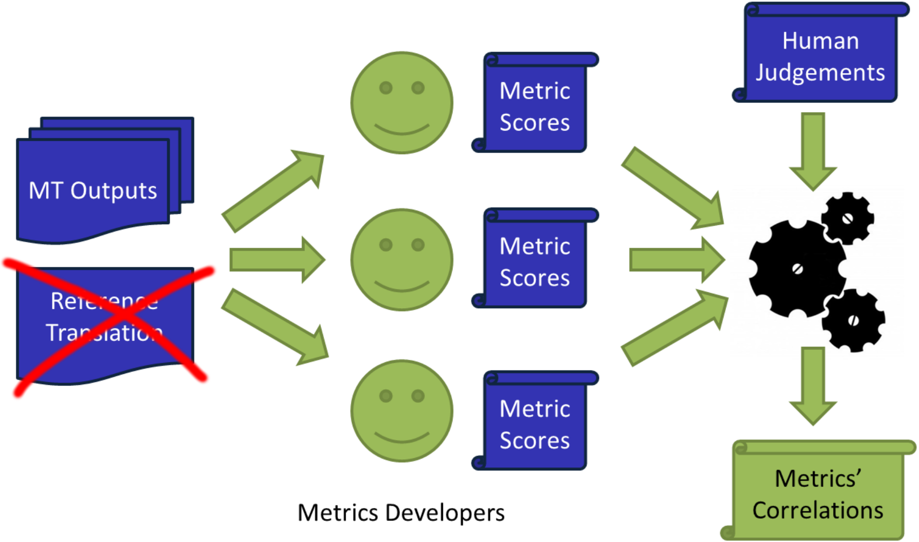
Metric Scores

Human Judgements

Metrics' Correlations

Metrics Developers

“QE as a Metric”

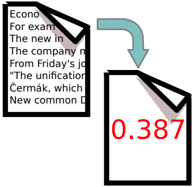


Updates in WMT19

- Golden truth
  - reference-based human evaluation – “monolingual”
  - reference-free human evaluation – “bilingual”
- Metrics
  - standard reference-based metrics
  - reference-less “metrics” – “QE as a Metric”
- “Hybrid” supersampling was not needed for sys-level:
  - Sufficiently large numbers of MT systems serve as datapoints.

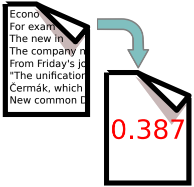
System- and Segment-Level Evaluation

- System Level
  - Participants compute one score for the whole test set, as translated by each of the systems

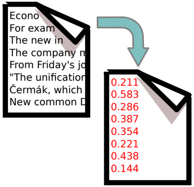


System- and Segment-Level Evaluation

- System Level
  - Participants compute one score for the whole test set, as translated by each of the systems



- Segment Level
  - Participants compute one score for each sentence of each system's translation



Past Metrics Tasks

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19
Participating Teams	-	6	8	14	9	8	12	12	11	9	8	8	**13**
Evaluated Metrics	11	16	38	26	21	12	16	23	46	16	14	10	**24**
Baseline Metrics							5	6	7	7	7	9	**11**
System-level													
Spearman Rank Corr	●	●	●	●	●	●	●	○					
Pearson Corr Coeff							○	●	●	●	●	●	●
Segment-level													
Rat. of Concord. Pairs		●	●										
Kendall's τ				❶	❶	❶	❷	❸	❸	❸	◗	❶	❶
based on		RR	RR	RR	RR	RR	RR	RR	RR	RR	daRR	daRR	daRR
Pearson Corr Coeff										○	◖		
based on										DA	DA		

● main and ○ secondary score reported for the system-level evaluation.

❶, ❷ and ❸ are slightly different variants regarding ties.

RR, DA, daRR are different golden truths.

Past Metrics Tasks

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19
Participating Teams	-	6	8	14	9	8	12	12	11	9	8	8	**13**
Evaluated Metrics	11	16	38	26	21	12	16	23	46	16	14	10	**24**
Baseline Metrics							5	6	7	7	7	9	**11**
System-level													
Spearman Rank Corr	●	●	●	●	●	●	●	○					
Pearson Corr Coeff							○	●	●	●	●	●	●
Segment-level													
Rat. of Concord. Pairs		●	●										
Kendall's τ				❶	❶	❶	❷	❸	❸	❸	◗	❶	❶
based on		RR	RR	RR	RR	RR	RR	RR	RR	RR	daRR	daRR	daRR
Pearson Corr Coeff										○	◖		
based on										DA	DA		

Increase in number of participating teams?

- "Baseline metrics": 9 + 2 reimplementations
  - sacreBLEU-BLEU and sacreBLEU-chrF.
- "Submitted metrics": 10 out of 24 are "QE as a Metric".

Data Overview This Year

- Domains:
  - News
- Golden Truths:
  - Direct Assessment (DA) for sys-level.
  - Derived relative ranking (daRR) for seg-level.
- Multiple languages (18 pairs):
  - English (en) to/from Czech (cs), German (de), Finnish (fi), Gujarati (gu), Kazakh (kk), Lithuanian (lt), Russian (ru), and Chinese (zh), but excluding cs-en.
  - German (de)→Czech (cs) and German (de)↔French (fr).

Baselines

Metric	Features	Seg-L	Sys-L
sentBLEU	n-grams	●	—
BLEU	n-grams	—	●
NIST	n-grams	—	●
WER	Levenshtein distance	—	●
TER	edit distance, edit types	—	●
PER	edit distance, edit types	—	●
CDER	edit distance, edit types	—	●
chrF	character n-grams	●	⊘
chrF+	character n-grams	●	⊘
sacreBLEU-BLEU	n-grams	—	●
sacreBLEU-chrF	n-grams	—	●

We average (⊘) seg-level scores.

Participating Metrics

Metric	Features	Seg-L	Sys-L	Team
BEER	char. n-grams, permutation trees	●	⊘	Univ. of Amsterdam, ILCC
BERTr	contextual word embeddings	●	⊘	Univ. of Melbourne
characTER	char. edit distance, edit types	●	⊘	RWTH Aachen Univ.
EED	char. edit distance, edit types	●	⊘	RWTH Aachen Univ.
ESIM	learned neural representations	●	⊘	Univ. of Melbourne
LEPORa	surface linguistic features	●	⊘	Dublin City University, ADAP
LEPORb	surface linguistic features	●	⊘	Dublin City University, ADAP
Meteor++_2.0 (syntax)	word alignments	●	⊘	Peking University
Meteor++_2.0 (syntax+copy)	word alignments	●	⊘	Peking University
PReP	psuedo-references, paraphrases	●	⊘	Tokyo Metropolitan Univ.
WMDO	word mover distance	●	⊘	Imperial College London
YiSi-0	semantic similarity	●	⊘	NRC
YiSi-1	semantic similarity	●	⊘	NRC
YiSi-1_srl	semantic similarity	●	⊘	NRC

We average (⊘) their seg-level scores.

Participating QE Systems

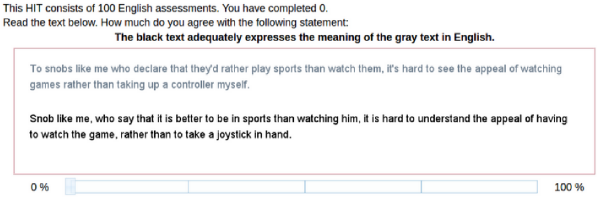
Metric	Features	Seg-L	Sys-L	Team
IBM1-morpheme	LM log probs., ibm1 lexicon	●	⊘	Dublin City University
IBM1-pos4gram	LM log probs., ibm1 lexicon	●	⊘	Dublin City University
LP	contextual word emb., MT log prob.	●	⊘	Univ. of Tartu
LASIM	contextual word embeddings	●	⊘	Univ. of Tartu
UNI	-	●	⊘	-
UNI+	-	●	⊘	-
USFD	-	●	⊘	Univ. of Sheffield
USFD-TL	-	●	⊘	Univ. of Sheffield
YiSi-2	semantic similarity	●	⊘	NRC
YiSi-2_srl	semantic similarity	●	⊘	NRC

We average (⊘) their seg-level scores.

Evaluation of System-Level

Golden Truth for Sys-Level: DA + Pearson

1. You have scored individual sentences: *(Thank you!)*



This HIT consists of 100 English assessments. You have completed 0.

Read the text below. How much do you agree with the following statement:

The black text adequately expresses the meaning of the gray text in English.

To snobs like me who declare that they'd rather play sports than watch them, it's hard to see the appeal of watching games rather than taking up a controller myself.

Snob like me, who say that it is better to be in sports than watching him, it is hard to understand the appeal of having to watch the game, rather than to take a joystick in hand.

2. News Task has filtered and standardized this (Ave z).
3. We correlate it with the metric sys-level score.

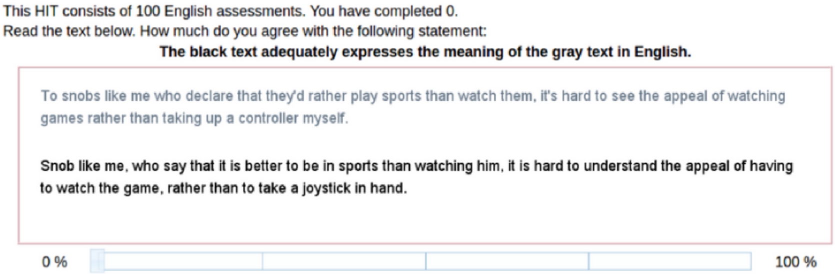
	Ave z	BLEU
CUNI-Transformer	0.594	0.2690
uedin	0.384	0.2438
online-B	0.101	0.2024
online-A	-0.115	0.1688
online-G	-0.246	0.1641

$\Rightarrow$ Pearson $= 0.995$

Evaluation of Segment-Level

Segment-Level News Task Evaluation

1. You scored individual sentences: (Same data as above.)



This HIT consists of 100 English assessments. You have completed 0.

Read the text below. How much do you agree with the following statement:

The black text adequately expresses the meaning of the gray text in English.

To snobs like me who declare that they'd rather play sports than watch them, it's hard to see the appeal of watching games rather than taking up a controller myself.

Snob like me, who say that it is better to be in sports than watching him, it is hard to understand the appeal of having to watch the game, rather than to take a joystick in hand.

2. Standardized, averaged ⇒ seg-level golden truth score.
3. Could be correlated to metric seg-level scores.

   . . . but there are not enough judgements for indiv. sentences.

daRR: Interpreting DA as RR

- If score for candidate A better than B by more than 25 points infer the pairwise comparison: A $>$ B.
  - No ties in golden daRR.
- Evaluate with the known Kendall's τ:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \tag{1}$$

- On average, there are 3–19 of scored outputs per src segm.
- From these, we generate 4k–327k daRR pairs.

Results of News Domain System-Level

Sys-Level into English (“Official”)

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
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						

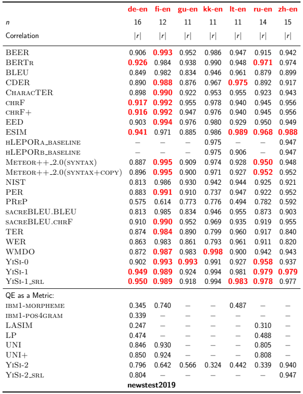
- Top: Baselines and regular metrics. Bottom: QE as a metric.

Sys-Level into English ("Official")

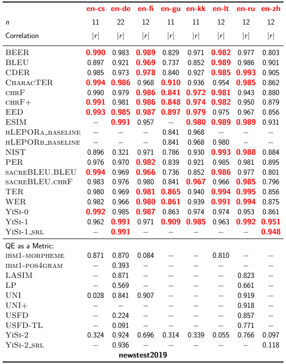
	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
BEER	0.906	**0.993**	0.952	0.986	0.947	0.915	0.942
BERTr	**0.926**	0.984	0.938	0.990	0.948	**0.971**	0.974
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER	0.890	**0.988**	0.876	0.967	**0.975**	0.892	0.917
CharacTER	0.898	**0.990**	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						

- Top: Baselines and regular metrics. Bottom: QE as a metric.
- **Bold**: not significantly outperformed by any others.

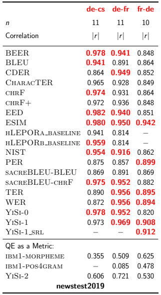
Sys-Level Results: Into, Out-of, Excl EN



	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						



	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					



	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			

- *-EN (except FI-EN) sufficiently discerning.
- EN-* and pair excluding EN somewhat more mixed.

Summary of Sys-Level Wins – Metrics

	Into EN			Out-of EN			Excluding EN			
	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	Overall wins
ESIM	7	0.96	4	6	0.97	4	3	0.96	3	12
YiSi-1	7	0.97	4	8	0.97	5	3	0.95	2	11
EED	7	0.95	1	8	0.95	5	3	0.92	2	8
chrF	7	0.95	2	8	0.95	4	3	0.92	1	7
chrF+	7	0.95	2	8	0.95	5	3	0.92	0	7
TER	7	0.89	1	8	0.95	4	3	0.91	2	7
YiSi-0	7	0.96	3	8	0.95	2	3	0.92	2	7
YiSi-1_srl	7	0.97	4	2	0.97	2	1	0.91	1	7
BEER	7	0.95	1	8	0.94	3	3	0.92	2	6
CDER	7	0.93	2	8	0.95	3	3	0.89	1	6
CharacTER	7	0.94	1	8	0.95	4	3	0.91	0	5
sacreBLEU-chrF	7	0.95	1	8	0.94	2	3	0.94	2	5
NIST	7	0.92	0	8	0.85	2	3	0.91	2	4
BLEU	7	0.91	0	8	0.91	2	3	0.9	1	3
PER	7	0.91	1	8	0.94	1	3	0.88	1	3
sacreBLEU-BLEU	7	0.9	0	8	0.91	3	3	0.88	0	3
BERTr	7	0.96	2	-	-	-	-	-	-	2
Met++_2.0(s.)	7	0.94	2	-	-	-	-	-	-	2
Met++_2.0(s.+copy)	7	0.94	2	-	-	-	-	-	-	2
WMDO	7	0.95	2	-	-	-	-	-	-	2
hLEPORb_baseline	3	0.94	0	3	0.93	0	2	0.89	1	1
PReP	7	0.66	0	-	-	-	-	-	-	0

Summary of Sys-Level Wins – QE

	Into EN			Out-of EN			Excluding EN		
	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins
ibm1-morpheme	3	0.52	0	4	0.66	0	3	0.5	0
ibm1-pos4gram	1	0.34	0	1	0.39	0	2	0.28	0
LASIM	2	0.28	0	2	0.85	0	-	-	-
LP	2	0.48	0	2	0.61	0	-	-	-
UNI+	3	0.86	0	1	0.92	0	-	-	-
UNI	3	0.86	0	4	0.67	0	-	-	-
USFD	-	-	-	2	0.54	0	-	-	-
USFD-TL	-	-	-	2	0.43	0	-	-	-
YiSi-2	7	0.58	0	8	0.44	0	3	0.62	0
YiSi-2_srl	2	0.88	0	2	0.53	0	-	-	-

Results of News Domain Segment-Level

Seg-Level Results: Into, Out-of, Excl EN

	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

	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

	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

- YiSi-1* win across the board and ESIM not far.
- FR-DE is not discerning.

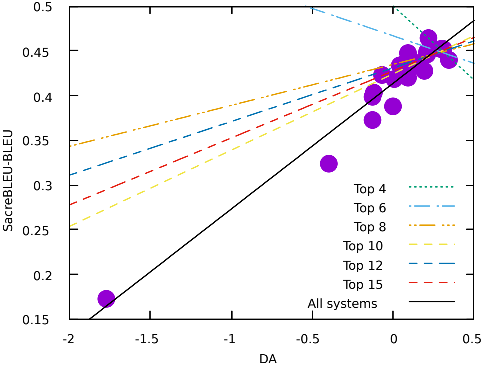
Summary of Seg-Level Wins – Metrics

	Into EN			Out-of EN			Excluding EN			
	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	Tot
YiSi-1	7	0.33	6	8	0.48	7	3	0.34	3	16
YiSi-1_srl	7	0.33	7	2	0.36	2	1	0.3	1	10
ESIM	7	0.31	3	6	0.45	2	3	0.3	1	6
chrF+	7	0.27	0	8	0.45	2	3	0.3	1	3
EED	7	0.27	0	8	0.38	1	3	0.3	1	2
BEER	7	0.28	0	8	0.44	0	3	0.3	1	1
CharacTER	7	0.23	0	8	0.34	0	3	0.24	1	1
chrF	7	0.27	0	8	0.45	0	3	0.29	1	1
YiSi-0	7	0.27	0	8	0.43	0	3	0.3	1	1
BERTr	7	0.3	0	-	-	-	-	-	-	0
hLEPORa_baseline	2	0.36	0	2	0.43	0	2	0.22	0	0
Meteor++_2.0(syntax)	7	0.26	0	-	-	-	-	-	-	0
Meteor++_2.0(syntax+copy)	7	0.26	0	-	-	-	-	-	-	0
PReP	7	0.2	0	-	-	-	-	-	-	0
sentBLEU	7	0.22	0	8	0.37	0	3	0.21	0	0
WMDO	7	0.27	0	-	-	-	-	-	-	0

Summary of Seg-Level Wins – QE

	Into EN			Out-of EN			Excluding EN		
	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins	LPs	⊘Corr	Wins
IBM1-MORPHEME	3	0.0	0	4	-0.08	0	3	-0.01	0
IBM1-POS4GRAM	1	-0.15	0	1	-0.12	0	2	-0.09	0
LASIM	2	0.0	0	2	-0.05	0	-	-	-
LP	2	-0.07	0	2	-0.14	0	-	-	-
UNI	3	0.1	0	4	0.19	0	-	-	-
UNI+	3	0.1	0	1	0.22	0	-	-	-
USFD	-	-	-	2	0.05	0	-	-	-
USFD-TL	-	-	-	2	0.08	0	-	-	-
YiSi-2	7	0.1	0	8	0.09	0	-	-	-
YiSi-2_srl	2	0.16	0	2	0.14	0	3	0.15	0

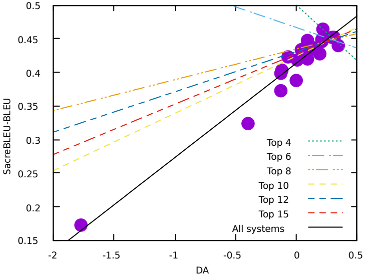
Stability across MT Systems

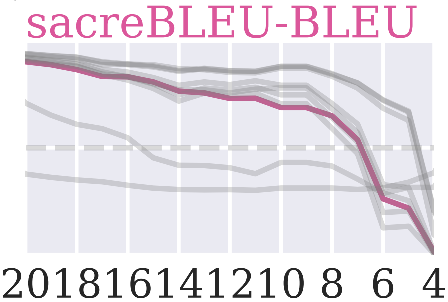


- EN→DE sys-level sacreBLEU-BLEU vs. golden truth.
- One outlier makes the task for metrics too easy.

Stability across MT Systems

- Get correlation when MT systems reduced to top-N ones.





- Baseline metrics are plotted in grey.
- In general, most metrics show a strong degrading pattern with the top-N systems across most language pairs.
    - Some “QE as a metric” have upward correlation trends.

Overall Status of MT Metrics

- Sys-level very good overall:
  - Pearson Correlation >.90 mostly, best reach >95 or >.98
    - Low pearsons exist but not many.
  - Correlations are heavily affected by the underlying set of MT systems.
    - System-level correlations are much worse when based on only the better performing systems.
  - No clear winners, *but have a look at this year's posters*.

Overall Status of MT Metrics

- Seg-level much worse:
  - The top Kendall's τ only .59.
    - standard metrics correlations varies between 0.03 and 0.59.
    - "QE a metric" obtains even negative correlations.
  - Methods using embeddings are better:
    - YISI-*: Word embeddings + other types of available resources.
    - ESIM: Sentence embeddings.

Next Metrics Task

- Yes, we will run the task!
- Big Challenge remains: References possibly worse than MT.

- Yes, we like the "QE as a metric" track.
- We will report the top-N plots.
  - We have to summarize them somehow, though.

- Doc-level golden truth did not seem different from sys-level.
  - This may change $\Rightarrow$ We might run doc-level metrics.

References