Microsoft

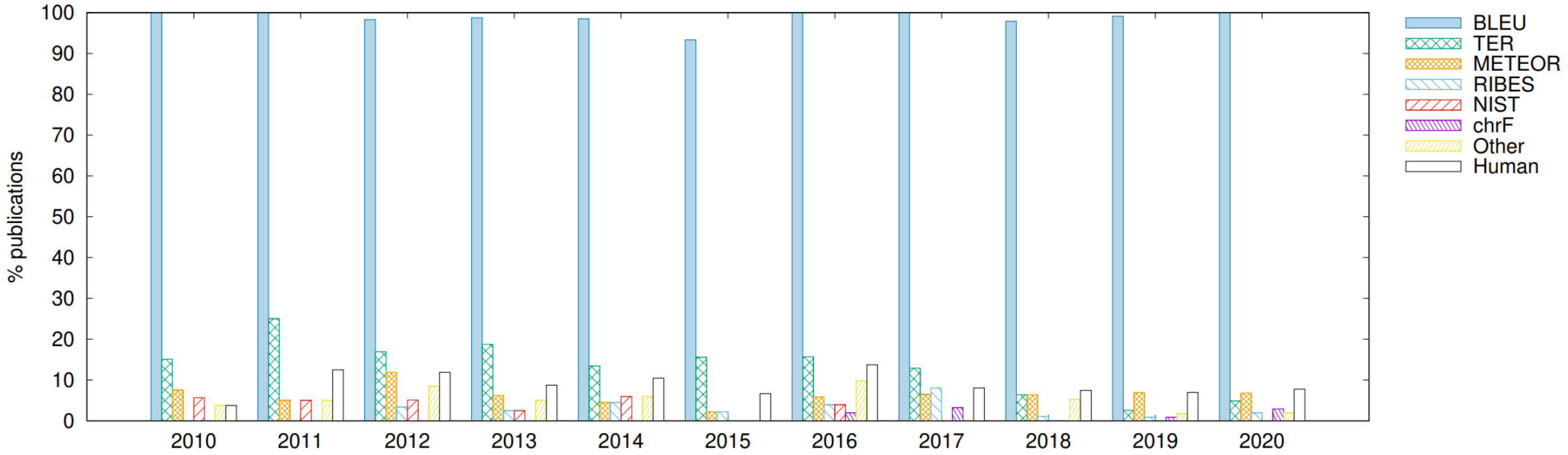
MT Evaluation:

Is your model really better?

Tom Kocmi

2022-09-09 at MT Marathon

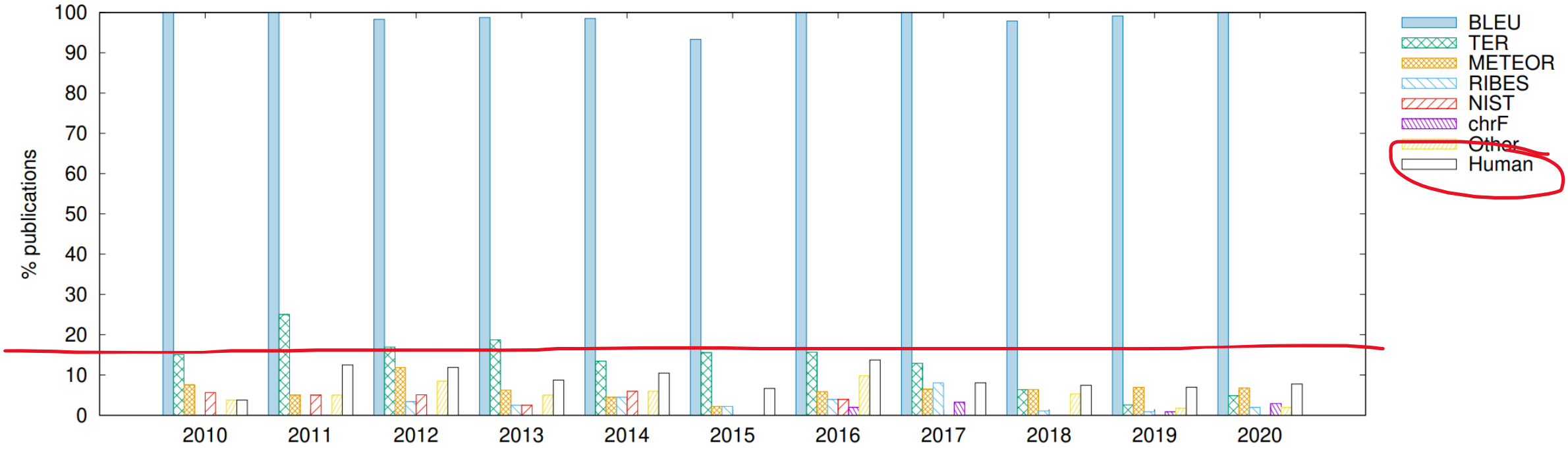
Motivation



Marie et al. (2021) reviewed 769 *ACL MT papers for ways how they evaluate models.

Motivation

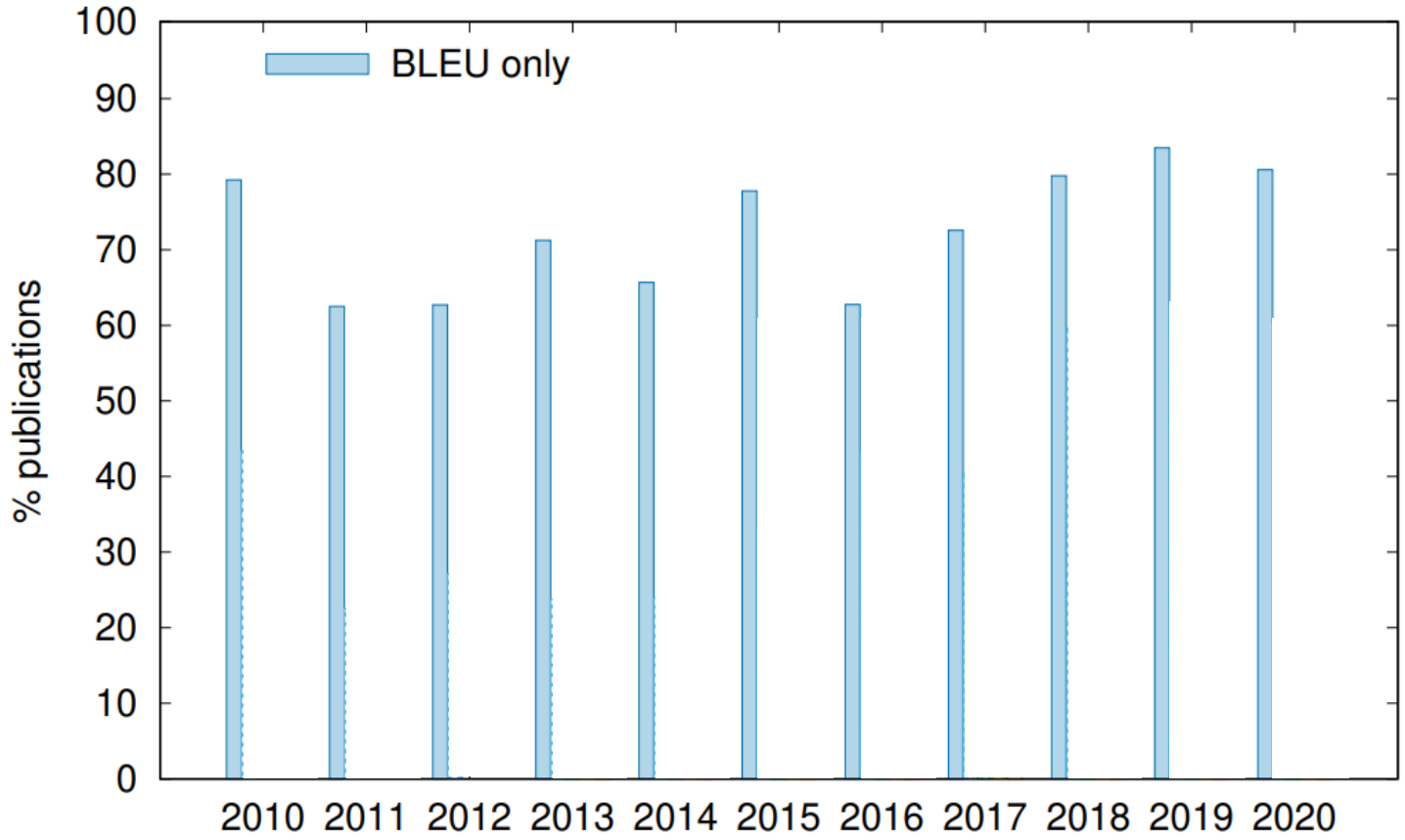
Improvements are often claimed in papers via automatic metrics only



Less than 10% of MT papers conduct human evaluation (Marie et al., 2021)

Motivation

Majority of papers rely solely on BLEU



Marie et al., 2021

Motivation

BLEU is not the best performing metric

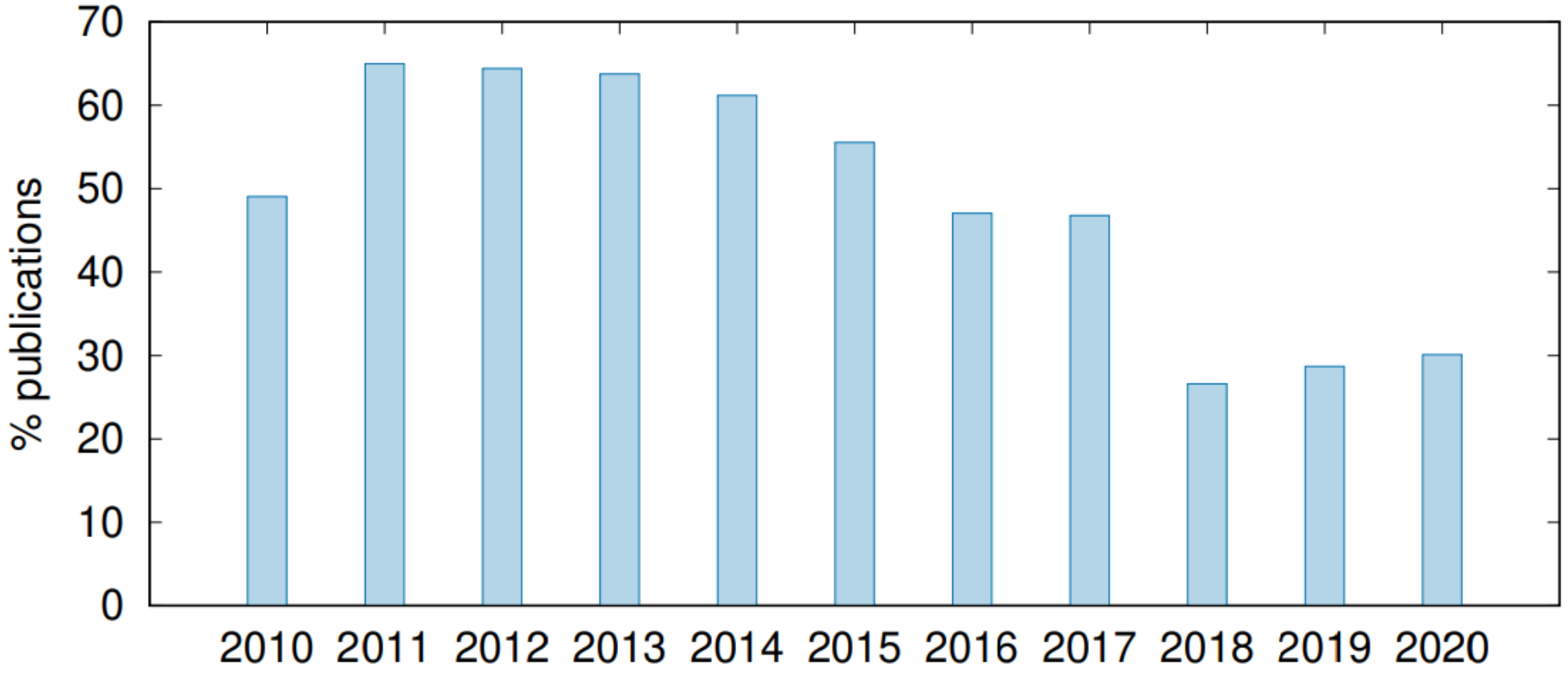
			2020 w/o ZH	2019	2018	2017	2016	2015
string-based	BLEU		0.832	0.906	0.955	0.91	0.873	0.841
	CharacTER		0.871	0.942	0.964	0.932	0.938	
	ChrF		0.864	0.948	0.959	0.942	0.911	0.908
	EED		0.885	0.951				
pretrained	BEER			0.942	0.973	0.938	0.925	0.942
	BLEURT		0.9					
	COMET		0.908					
	ESIM		0.902					
	Prism		0.886					
	YiSi-1		0.89	0.967	0.973			

Meta analysis of WMT 2015-2020 (Pearson's r)

*Simplification: column 2020 is without Chinese-English systems due a single problematic system (see paper for details)

Motivation

Disappearing statistical significance testing



Outline

1. Which metric to use?
2. Weaknesses of metrics
3. Statistical significance testing
4. Are we overfitting on BLEU?
5. Recommendations for MT evaluation

Which automatic metric to use?

Automatic metrics groups

Source	2015 verließ Demandt Borussia, wechselte nach Wehen.
Reference	In 2015, Demandt left Borussia for Wehen.
System Output	In 2015, Demandt left Borussia and moved to Wehen.

String-based metrics

Compares reference with system output on a string level

Metrics: BLEU, ChrF, TER

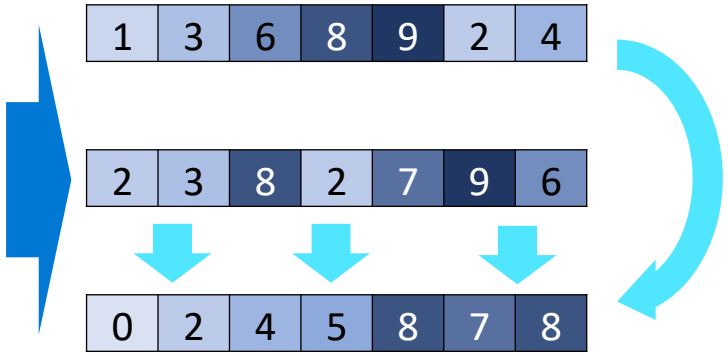
Source	2015 verließ Demandt Borussia, wechselte nach Wehen.
Reference	In 2015, Demandt left Borussia for Wehen.
System Output	In 2015, Demandt left Borussia and moved to Wehen.

Pretrained metrics

Contains pretrained model (e.g. language model: BERT, XLM, ...)

Metrics: COMET, BERTScore, Prism

Source	2015 verließ Demandt Borussia, wechselte nach Wehen.
Reference	In 2015, Demandt left Borussia for Wehen.
System Output	In 2015, Demandt left Borussia and moved to Wehen.

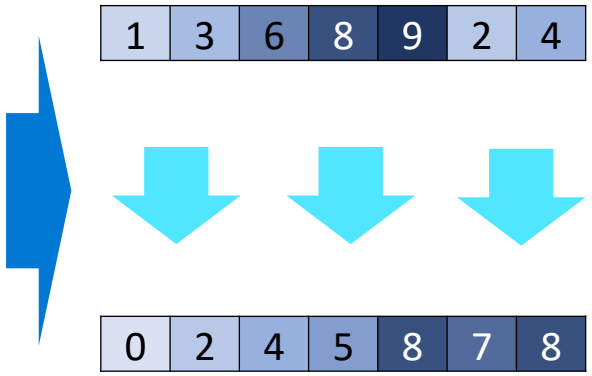


Pretrained source-based metrics

Pretrained metrics that do not use reference (quality estimation)

Metrics: COMET-QE, Prism

Source	2015 verließ Demandt Borussia, wechselte nach Wehen.
Reference	In 2015, Demandt left Borussia for Wehen.
System Output	In 2015, Demandt left Borussia and moved to Wehen.



Main differences

String-based metrics	Pretrained metrics
Work for any language	Support only pretrained languages
Scoring decisions are trackable	Black box behaviour • Possible biases from training data • Possible poor quality for low-resource languages or domains
Need high quality references and cannot score paraphrases	Can score paraphrases and are less sensitive to references
Lower correlation with humans	Higher correlation with humans

Which metric correlates the most with humans?

	en-cs		en-de		en-ja		en-pl		en-ru	en-ta		en-zh	en-iu$_{FULL}$		en-iu$_{NEWS\#}$	
	All	-out	All	-out	All	-out	All	-out	All	All	-out	All	All	-out	All	-out
	12	10	14	11	11	9	14	11	9	15	12	12	11	8	11	8
SENTBLEU	0.840	0.436	0.934	0.823	0.946	**0.976**	0.950	**0.772**	**0.981**	0.881	0.852	0.927	0.129	0.047	0.075	**0.172**
BLEU	0.825	0.390	0.928	0.825	0.945	**0.980**	0.943	**0.743**	**0.980**	0.880	0.829	0.928	0.163	0.131	0.074	**0.111**
TER	0.814	0.339	0.941	**0.848**	0.297	0.801	0.893	0.553	0.064	0.870	**0.883**	-0.213	0.384	**0.133**	0.357	**0.083**
CHRF++	0.833	0.349	0.958	**0.850**	0.952	0.945	**0.956**	**0.783**	**0.983**	0.929	**0.880**	0.878	0.328	0.128	0.315	**0.098**
CHRF	0.826	0.313	**0.962**	**0.862**	0.951	0.964	**0.957**	**0.793**	**0.982**	0.937	**0.890**	0.923	0.350	**0.122**	0.336	**0.091**
PARBLEU	0.870	0.543	0.910	0.774	0.869	0.813	0.948	**0.760**	**0.959**	0.871	**0.849**	0.962	0.194	**0.464**	0.126	**0.306**
PARCHRF++	0.860	0.438	0.957	**0.845**	0.955	0.951	**0.953**	**0.818**	**0.975**	–		0.948	–		–	
CHARACTER	0.807	0.269	**0.961**	**0.868**	0.951	**0.936**	0.935	**0.726**	0.961	0.957	0.851	0.905	0.503	0.008	0.515	**0.121**
EED	0.817	0.271	**0.965**	**0.869**	0.955	0.965	**0.962**	**0.789**	**0.980**	**0.959**	**0.913**	0.928	0.519	0.043	0.483	**0.122**
MEE	0.875	0.495	0.954	0.820	–		**0.952**	**0.733**	0.724	0.906	0.861	–	0.287	0.094	0.242	**0.113**
YISI-0	0.797	0.270	0.953	**0.889**	0.967	**0.972**	**0.953**	**0.783**	**0.971**	0.929	**0.897**	0.362	0.525	0.015	0.505	**0.095**
PRISM	0.949	0.805	0.958	**0.851**	0.932	0.921	**0.958**	**0.742**	0.724	0.863	0.452	0.221	**0.957**	**-0.043**	**0.945**	**0.088**
YISI-1	0.922	0.664	**0.971**	**0.887**	0.969	0.967	**0.964**	**0.714**	0.926	**0.973**	**0.909**	0.959	0.554	-0.217	0.523	-0.014
YISI-COMBI	–		0.971	0.868	–		–		–	–		–	–		–	
BLEURT-YISI-COMBI	–		0.971	0.868	–		–		–	–		–	–		–	
MBERT-L2	0.946	0.782	**0.970**	**0.861**	0.977	**0.969**	**0.976**	**0.775**	0.946	**0.944**	0.834	0.934	–		–	
BLEURT-EXTENDED	**0.989**	**0.960**	**0.969**	**0.870**	0.944	0.953	**0.982**	**0.828**	**0.980**	0.940	0.814	0.928	0.823	**0.122**	0.762	**0.155**
ESIM	0.908	0.575	**0.979**	**0.894**	**0.993**	**0.981**	**0.969**	0.698	**0.967**	0.937	0.833	0.972	0.814	**0.365**	0.760	0.418
PARESIM-1	0.919	0.635	**0.974**	**0.886**	**0.989**	**0.971**	**0.968**	0.705	**0.964**	0.937	0.833	**0.983**	0.814	**0.365**	0.760	0.418
COMET	0.978	0.926	**0.972**	**0.863**	0.974	**0.969**	**0.981**	**0.800**	0.925	0.944	0.798	0.007	0.860	0.028	**0.858**	0.152
COMET-2R	**0.983**	0.942	**0.972**	**0.869**	**0.986**	**0.978**	**0.982**	**0.803**	0.872	**0.959**	0.852	-0.066	0.848	-0.008	**0.867**	**0.177**
COMET-HTER	0.976	0.917	0.951	**0.852**	**0.989**	**0.974**	**0.974**	**0.763**	0.803	0.925	0.681	-0.073	**0.900**	0.142	**0.888**	0.092
COMET-MQM	0.974	0.910	0.881	0.840	0.974	**0.965**	0.967	**0.766**	0.788	0.910	0.641	0.084	0.870	0.129	**0.867**	0.172
COMET-RANK	0.959	0.868	0.877	**0.860**	0.931	0.928	0.957	**0.760**	0.676	0.876	0.511	0.540	0.283	0.099	0.392	**0.252**
BAQ_DYN	–		–		–		–		–	–		0.904	–		–	
BAQ_STATIC	–		–		–		–		–	–		**0.958**	–		–	
EQ_DYN	–		–		–		–		–	–		0.948	–		–	
EQ_STATIC	–		–		–		–		–	–		**0.976**	–		–	
COMET-QE	**0.989**	**0.974**	0.903	0.831	0.953	**0.955**	0.969	**0.804**	0.807	0.887	0.622	0.375	**0.905**	**0.578**	**0.928**	**0.651**
OPENKIWI-BERT	0.920	0.830	0.852	**0.829**	0.363	0.783	0.903	0.450	0.834	0.846	0.370	0.551	0.573	-0.602	0.808	**0.194**
OPENKIWI-XLMR	0.972	0.911	**0.968**	0.814	**0.992**	**0.976**	0.957	**0.638**	0.875	0.910	0.676	-0.010	0.513	-0.668	0.680	-0.358
YISI-2	0.714	0.353	0.899	0.552	0.854	0.646	0.470	-0.107	0.584	0.922	**0.923**	-0.215	0.802	**-0.257**	**0.830**	**0.065**

Mathur et al. (2020)

What do we expect from automatic metrics?

Wishful thinking:

- High correlation with oracle humans
- Meaningful absolute values
- Sentence-level correlations

What do we expect from automatic metrics?

Wishful thinking:

- High correlation with oracle humans
  - **Humans have low inter-annotator agreement**
- Meaningful absolute values
- Sentence-level correlations

What do we expect from automatic metrics?

Wishful thinking:

- High correlation with oracle humans
- Meaningful absolute values
  - **What does 24 BLEU mean without context?**
- Sentence-level correlations

What do we expect from automatic metrics?

Wishful thinking:

- High correlation with oracle humans
- Meaningful absolute values
- Sentence-level correlations
  - System-level scores are more useful

The main setting

Which automatic metric correlates the most with humans in the pairwise system-level setting?

The main setting

Which automatic metric correlates the most with humans in the pairwise system-level setting?

Automatic metrics are mostly used to rank systems.

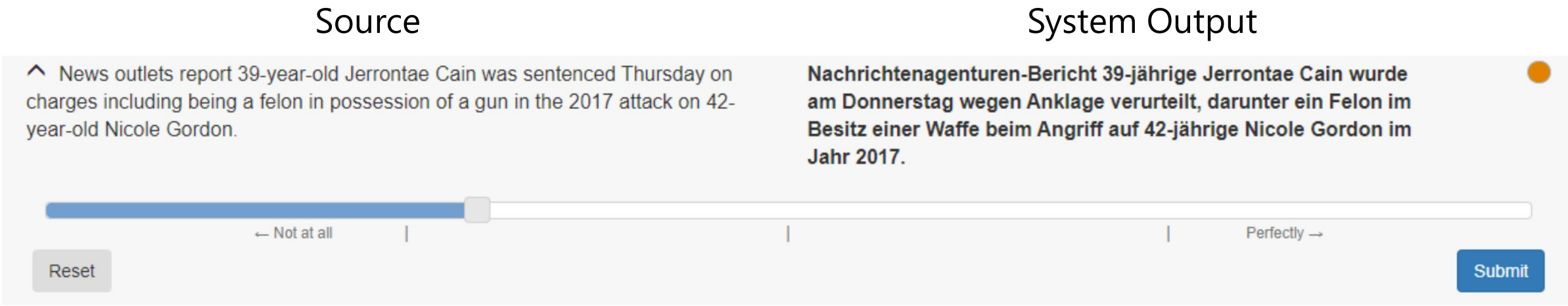
- claiming a new state-of-the-art
- comparing different model architectures
- deciding whether to deploy new production systems.

Best performing metric in pairwise system level setting?

To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation
Kocmi et al. (2021)

Human judgement collection

Bilingual annotators score each translated sentence without reference.



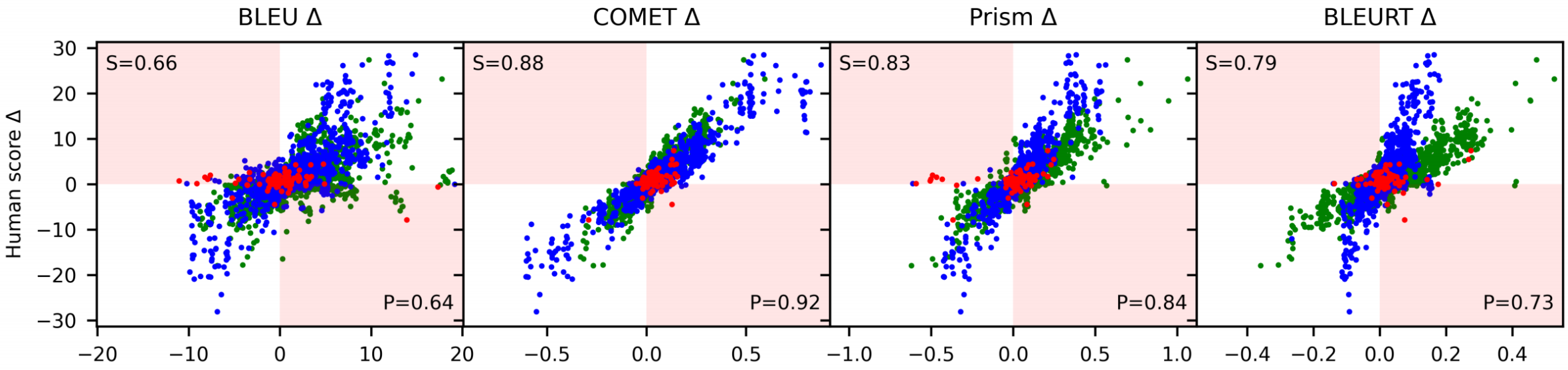
*Limitation: we assume human judgement as a gold standard

Pairwise evaluation

The main unit is the score difference (Δ) for system pair:

$$\Delta_{metric} = score_{metric}(System\ A) - score_{metric}(System\ B)$$

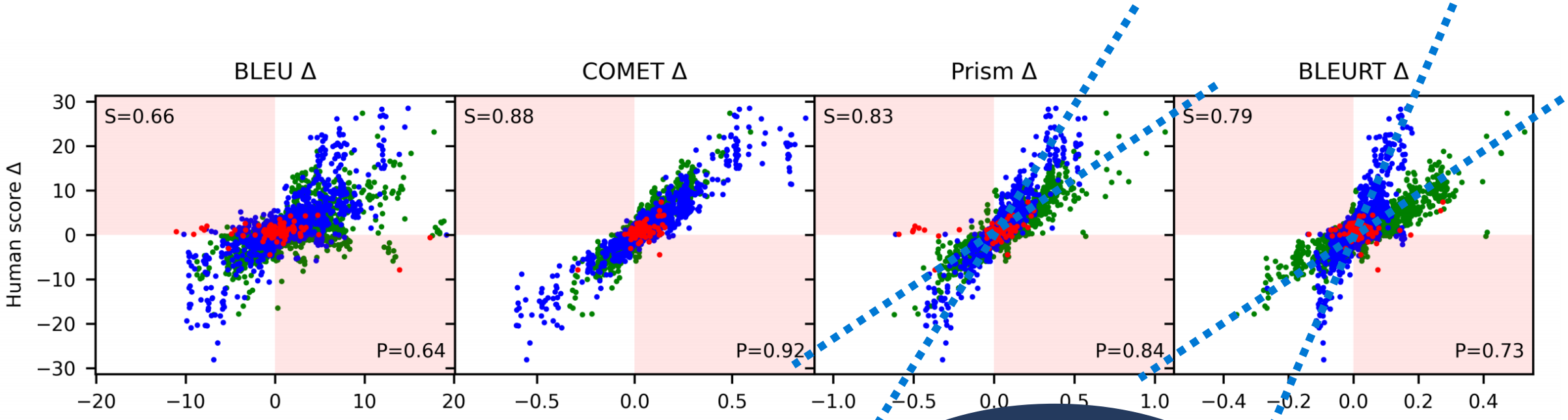
Human judgement vs. metric score differences



Each point represent score differences for a system pair

Blue points are system pairs translating from English
Green points are into English
Red points are non-English systems (there is only few of those, painted on top)

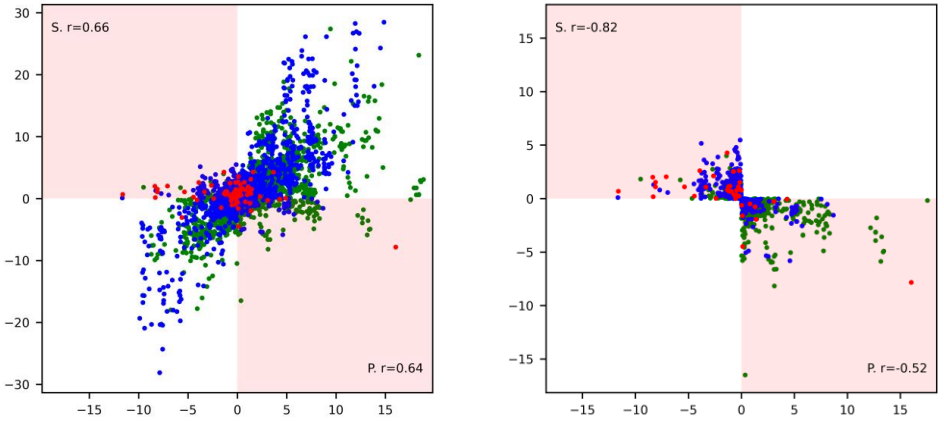
Human judgement vs. metric score differences



Pearson's correlation cannot be used

Accuracy

$$\text{Accuracy} = \frac{|sign(\Delta_{metric}) = sign(\Delta_{human})|}{|all\ system\ pairs|}$$



How accurately metric gives the binary decision: System A is better/worse than System B.

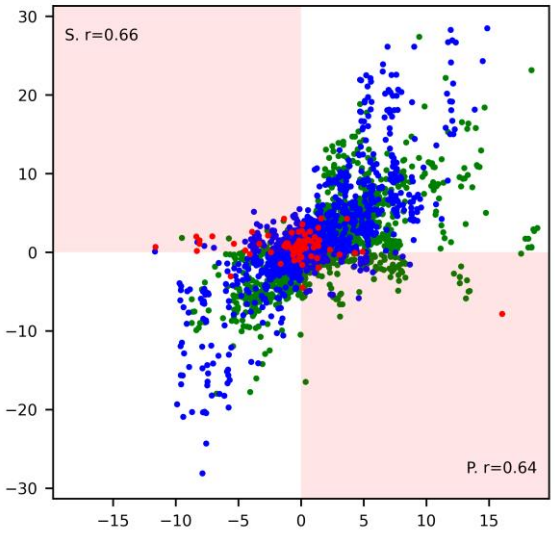
Equally performing systems

What about the magnitude of the difference?

We identify equally performing systems by statistical testing over human judgement.

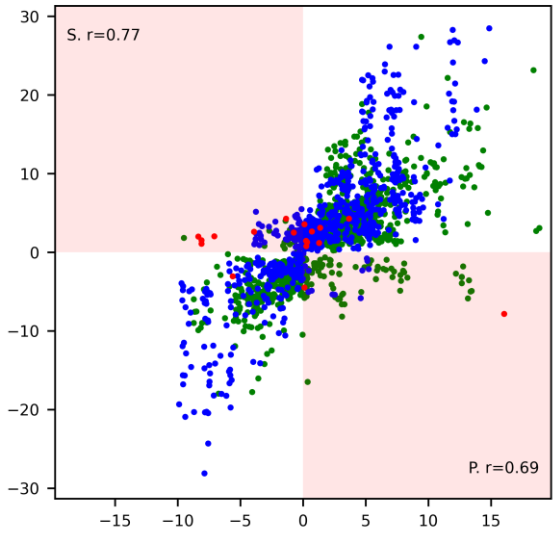
Equally performing systems

All systems



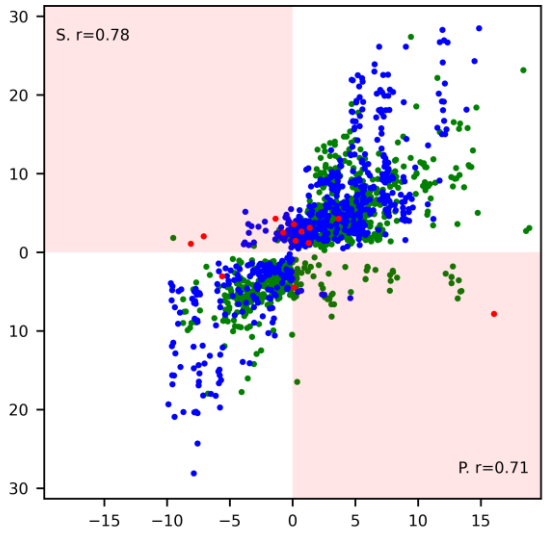
Accuracy = 74.0%

p-value < 0.05



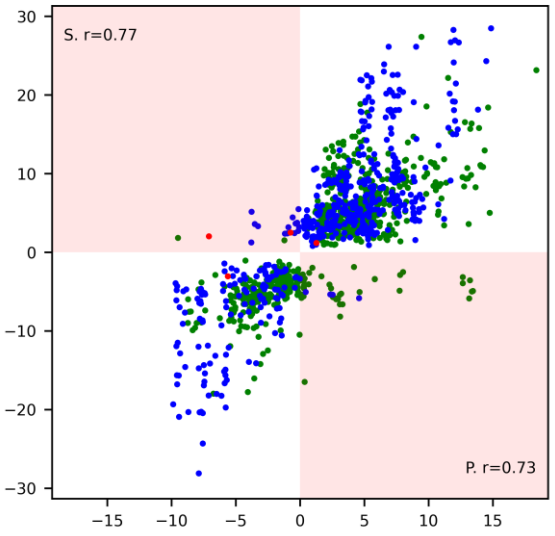
Accuracy = 87.3%

p-value < 0.01



Accuracy = 89.6%

p-value < 0.001



Accuracy = 92.1%

Collection of human judgements

	Our collection*	WMT 2020
Evaluated systems	4380	184
Human annotations	2.3 M	0.4 M
Covered languages	101	12
Annotators	Qualified bilingual annotators	Crowd workers; Paid annotators; Researchers

* two years of MT research at Microsoft

Covered language pairs

Language pair	Sys.
English - French	145
English - German	139
French - English	131
German - English	122
Japanese - English	78
Chinese - English	74
Italian - English	71
English - Portuguese	70
English - Japanese	67
English - Swedish	66
English - Chinese	65
English - Danish	64
English - Italian	64
English - Polish	64
Spanish - English	64
Dutch - English	63
English - Dutch	61
English - Indonesian	61
Indonesian - English	60
English - Czech	59
Arabic - English	59
English - Spanish	58
Hungarian - English	58

Language pair	Sys.
English - Hindi	58
Polish - English	57
Portuguese - English	57
Swedish - English	57
English - Arabic	56
Korean - English	56
Czech - English	55
English - Hungarian	55
English - Korean	55
English - Turkish	55
English - Thai	54
Hindi - English	54
Turkish - English	54
Danish - English	52
English - Russian	49
Russian - English	44
Thai - English	39
English - Catalan	30
Hebrew - English	28
English - Romanian	27
Romanian - English	27
English - Greek	27
Persian - English	26

Language pair	Sys.
English - Ukrainian	25
English - Slovak	25
English - Irish	24
English - Persian	24
Slovak - English	23
Greek - English	23
English - Croatian	22
English - Welsh	22
English - Norwegian	22
English - Hebrew	22
English - Vietnamese	20
Welsh - English	20
Vietnamese - English	20
Catalan - English	20
English - Urdu	18
English - Finnish	17
Tamil - English	16
English - Lithuanian	16
Lithuanian - English	16
English - Maltese	16
English - Kiswahili	16

Evaluation results

Which metric is best suited for pairwise comparison?

system pairs:	All 3344	p-value < 0.05 1717	p-value < 0.01 1420	p-value < 0.001 1176	Within 0.05 and 0.001 541
COMET	83.4				
COMET-QE	83.2				
Prism	80.6				
BLEURT	80.0				
ESIM	78.7				
BERTScore	78.3				
ChrF	75.6				
TER	75.6				
CharacTER	74.9				
BLEU	74.6				
Prism	73.4				

*Values in large font represent cluster of winning metrics (statistically calculated with alpha=0.05).

Which metric is best suited for pairwise comparison?

system pairs:	All 3344	p-value < 0.05 1717	p-value < 0.01 1420	p-value < 0.001 1176	Within 0.05 and 0.001 541
COMET	83.4	96.5	98.7	99.2	90.6
COMET-QE	83.2	95.3	97.4	98.1	89.1
Prism	80.6	94.5	97.0	98.3	86.3
BLEURT	80.0	93.8	95.6	98.2	84.1
ESIM	78.7	92.9	95.6	97.5	82.8
BERTScore	78.3	92.2	95.2	97.4	81.0
ChrF	75.6	89.5	93.5	96.2	75.0
TER	75.6	89.2	93.0	96.2	73.9
CharacTER	74.9	88.6	91.9	95.2	74.1
BLEU	74.6	88.2	91.7	94.6	74.3
Prism	73.4	85.3	87.6	88.9	77.4

Which metric is best suited for pairwise comparison?

system pairs:	All 3344	p-value < 0.05 1717	p-value < 0.01 1420	p-value < 0.001 1176	Within 0.05 and 0.001 541
COMET	83.4	96.5	98.7	99.2	90.6
COMET-QE	83.2	95.3	97.4	98.1	89.1
Prism	80.6	94.5	97.0	98.3	86.3
BLEURT	80.0	93.8	95.6	98.2	84.1
ESIM	78.7	92.9	95.6	97.5	82.8
BERTScore	78.3	92.2	95.2	97.4	81.0
ChrF	75.6	89.5	93.5	96.2	75.0
TER	75.6	89.2	93.0	96.2	73.9
CharacTER	74.9	88.6	91.9	95.2	74.1
BLEU	74.6	88.2	91.7	94.6	74.3
Prism	73.4	85.3	87.6	88.9	77.4

Pretrained metrics

String-based metrics

Which metric is best suited for pairwise comparison?

system pairs:	All 3344	p-value < 0.05 1717	p-value < 0.01 1420	p-value < 0.001 1176	Within 0.05 and 0.001 541	
COMET	83.4	96.5	98.7	99.2	90.6	
COMET-QE	83.2	95.3	97.4	98.1	89.1	Reference-less (QE)
Prism	80.6	94.5	97.0	98.3	86.3	
BLEURT	80.0	93.8	95.6	98.2	84.1	
ESIM	78.7	92.9	95.6	97.5	82.8	
BERTScore	78.3	92.2	95.2	97.4	81.0	
ChrF	75.6	89.5	93.5	96.2	75.0	
TER	75.6	89.2	93.0	96.2	73.9	
CharacTER	74.9	88.6	91.9	95.2	74.1	
BLEU	74.6	88.2	91.7	94.6	74.3	
Prism	73.4	85.3	87.6	88.9	77.4	Reference-less (QE)

Are metrics reliable for various scenarios?

system pairs:	All 1717	Into EN 922	From EN 768	Non latin target 131	Chinese, Japanese, Korean 44	Non WMT langs 484	Speech domain 78
COMET	96.5						
COMET-QE	95.3						
Prism	94.5						
BLEURT	93.8						
ESIM	92.9						
BERTScore	92.2						
ChrF	89.5						
TER	89.2						
CharacTER	88.6						
BLEU	88.2						
Prism	85.3						

Results are for system pairs with p-value smaller than 0.05 over human judgement.

Do metrics correlate equally into and from English?

system pairs:	All 1717	Into EN 922	From EN 768	Non latin target 131	Chinese, Japanese, Korean 44	Non WMT langs 484	Speech domain 78
COMET	96.5	95.3	98.3				
COMET-QE	95.3	93.5	97.7				
Prism	94.5	92.2	98.2				
BLEURT	93.8	93.8	95.1				
ESIM	92.9	90.6	96.6				
BERTScore	92.2	91.2	94.1				
ChrF	89.5	88.7	91				
TER	89.2	87.6	91.7				
CharacTER	88.6	86.4	91.7				
BLEU	88.2	86.9	90.5				
Prism	85.3	80.8	91.4				

Are metrics reliable for non-latin languages?

system pairs:	All 1717	Into EN 922	From EN 768	Non latin target 131	Chinese, Japanese, Korean 44	Non WMT langs 484	Speech domain 78
COMET	96.5	95.3	98.3	96.2	90.9		
COMET-QE	95.3	93.5	97.7	95.4	88.6		
Prism	94.5	92.2	98.2	96.2	90.9		
BLEURT	93.8	93.8	95.1	93.1	84.1		
ESIM	92.9	90.6	96.6	93.9	86.4		
BERTScore	92.2	91.2	94.1	95.4	88.6		
ChrF	89.5	88.7	91	95.4	88.6		
TER	89.2	87.6	91.7	90.1	72.7		
CharacTER	88.6	86.4	91.7	88.5	70.5		
BLEU	88.2	86.9	90.5	92.4	79.5		
Prism	85.3	80.8	91.4	84	65.9		

Can metrics be overfitted?

system pairs:	All 1717	Into EN 922	From EN 768	Non latin target 131	Chinese, Japanese, Korean 44	Non WMT langs 484	Speech domain 78
COMET							
COMET-QE							
Prism							
BLEURT							
ESIM							
BERTScore							
ChrF							
TER							
CharacTER							
BLEU							
Prism							

Some metrics are finetuned on human judgement from WMT News

Can metrics be overfitted?

system pairs:	All 1717	Into EN 922	From EN 768	Non latin target 131	Chinese, Japanese, Korean 44	Non WMT langs 484	Speech domain 78
COMET	96.5	95.3	98.3	96.2	90.9	97.3	93.6
COMET-QE	95.3	93.5	97.7	95.4	88.6	96.7	93.6
Prism	94.5	92.2	98.2	96.2	90.9	96.9	83.3
BLEURT	93.8	93.8	95.1	93.1	84.1	94.6	89.7
ESIM	92.9	90.6	96.6	93.9	86.4	94.8	76.9
BERTScore	92.2	91.2	94.1	95.4	88.6	92.8	71.8
ChrF	89.5	88.7	91	95.4	88.6	89.7	57.7
TER	89.2	87.6	91.7	90.1	72.7	90.9	70.5
CharacTER	88.6	86.4	91.7	88.5	70.5	91.9	69.2
BLEU	88.2	86.9	90.5	92.4	79.5	89.9	61.5
Prism	85.3	80.8	91.4	84	65.9	91.7	84.6

Weaknesses of metrics

Freitag et al. (2021)

Metric	Antonym	W. Omission	Tokenized	Sent. Omission	Punct.	Numbers	Lower.	W. Add.	Spell.
SENT-BLEU	0.792	0.787	-0.617	0.409	0.640	0.715	0.633	0.986	0.954
TER	0.994	0.597	0.966	0.568	0.739	0.996	1.000	1.000	0.997
CHRF	0.887	0.983	-0.516	0.523	0.761	0.899	0.708	0.903	0.981
BERTSCORE	0.986	0.984	0.994	0.909	0.950	0.993	0.799	0.996	0.998
PRISM	0.998	0.995	0.972	0.864	0.990	1.000	0.969	1.000	0.999
BLEURT-20	0.992	0.989	0.983	0.909	0.931	0.993	0.976	0.998	0.997
COMET-MQM_2021	0.996	0.994	0.994	-0.068	0.235	0.993	0.965	0.993	1.000
C-SPECPN	0.991	0.988	0.576	0.409	0.622	0.876	0.922	0.991	0.996
COMET-source (MQM)	0.991	0.983	0.983	-0.318	-0.199	0.989	0.931	0.982	0.998

Kendall's tau-like correlation results for the Corrupted References

Amrhein et al. (2022)

Use Minimum Bayes Risk Decoding to discover weaknesses in COMET.

COMET is relatively insensitive to mistranslated numbers and named entities

src	Schon drei Jahre nach der Gründung verließ Green die Band **1970**.
ref	Green left the band three years after it was formed, in **1970**.
MBR_{chrF++}	Already three years after the foundation, Green left the band in **1970**.
MBR_{COMET}	Three years after the creation, Green left the band in **1980** .

src	[...] **Mahmoud** Guemama's Death - Algeria Loses a Patriot [...], Says President **Tebboune**.
ref	[...] **Mahmoud** Guemamas Tod - Algerien verliert einen Patrioten [...], sagt Präsident **Tebboune**.
MBR_{chrF++}	[...] **Mahmoud** Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident **Tebboune**.
MBR_{COMET}	[...] **Mahmud** Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident **Tebboene** .

Recommendation:

Use COMET as the primary metric
Use ChrF as a secondary metric for noticing pitfalls

Statistical significance testing

Statistical significance of automatic metrics

- Small difference in metric scores (e.g. 0.5 BLEU) could be due to randomness.

- Statistical significance tests
  - Paired bootstrap resampling
  - Approximate randomization
  - Paired Student t-test (for metrics that average sentence scores)

Effect of statistical significance testing

	No test	Bootstrap resampling	Percentage of systems incorrectly rejected by boot.
COMET	83.4		
COMET-source	83.2		
Prism	80.6		
BLEURT	80.0		
BERTScore	78.3		
ChrF	75.6		
BLEU	74.6		
Prism-source	73.4		

Effect of statistical significance testing

	No test	Bootstrap resampling	Percentage of systems incorrectly rejected by boot.
COMET	83.4	95.1	
COMET-source	83.2	94.2	
Prism	80.6	91.3	
BLEURT	80.0	92.0	
BERTScore	78.3	87.9	
ChrF	75.6	85.4	
BLEU	74.6	83.4	
Prism-source	73.4	81.5	

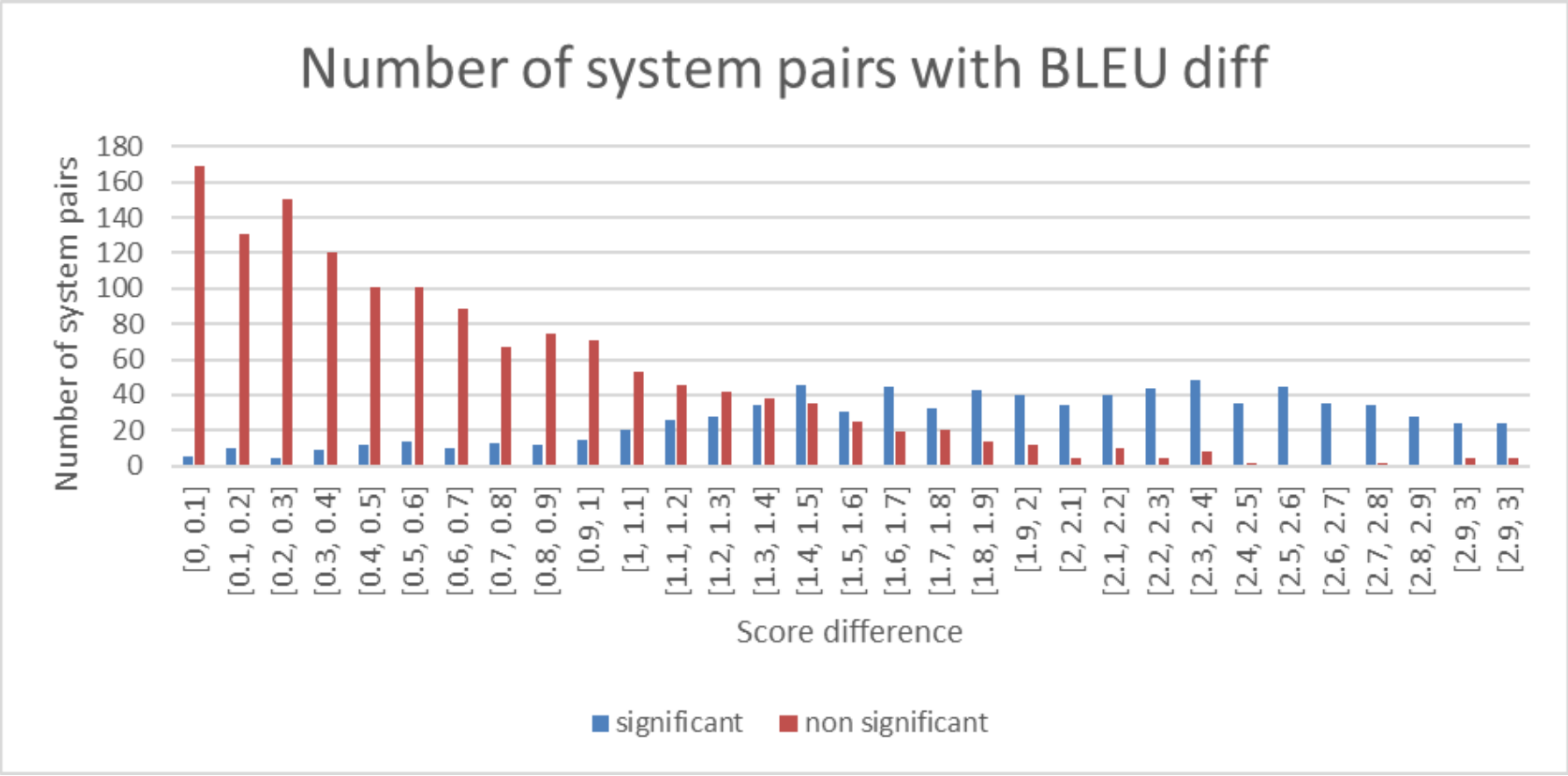
Effect of statistical significance testing

	No test	Bootstrap resampling	Percentage of systems incorrectly rejected by boot.
COMET	83.4	95.1	17.3%
COMET-source	83.2	94.2	19.4%
Prism	80.6	91.3	18.3%
BLEURT	80.0	92.0	25.4%
BERTScore	78.3	87.9	20.9%
ChrF	75.6	85.4	27.3%
BLEU	74.6	83.4	27.4%
Prism-source	73.4	81.5	29.4%

*Limitation: we take humans as gold standard and ignore their type I. and II. errors.

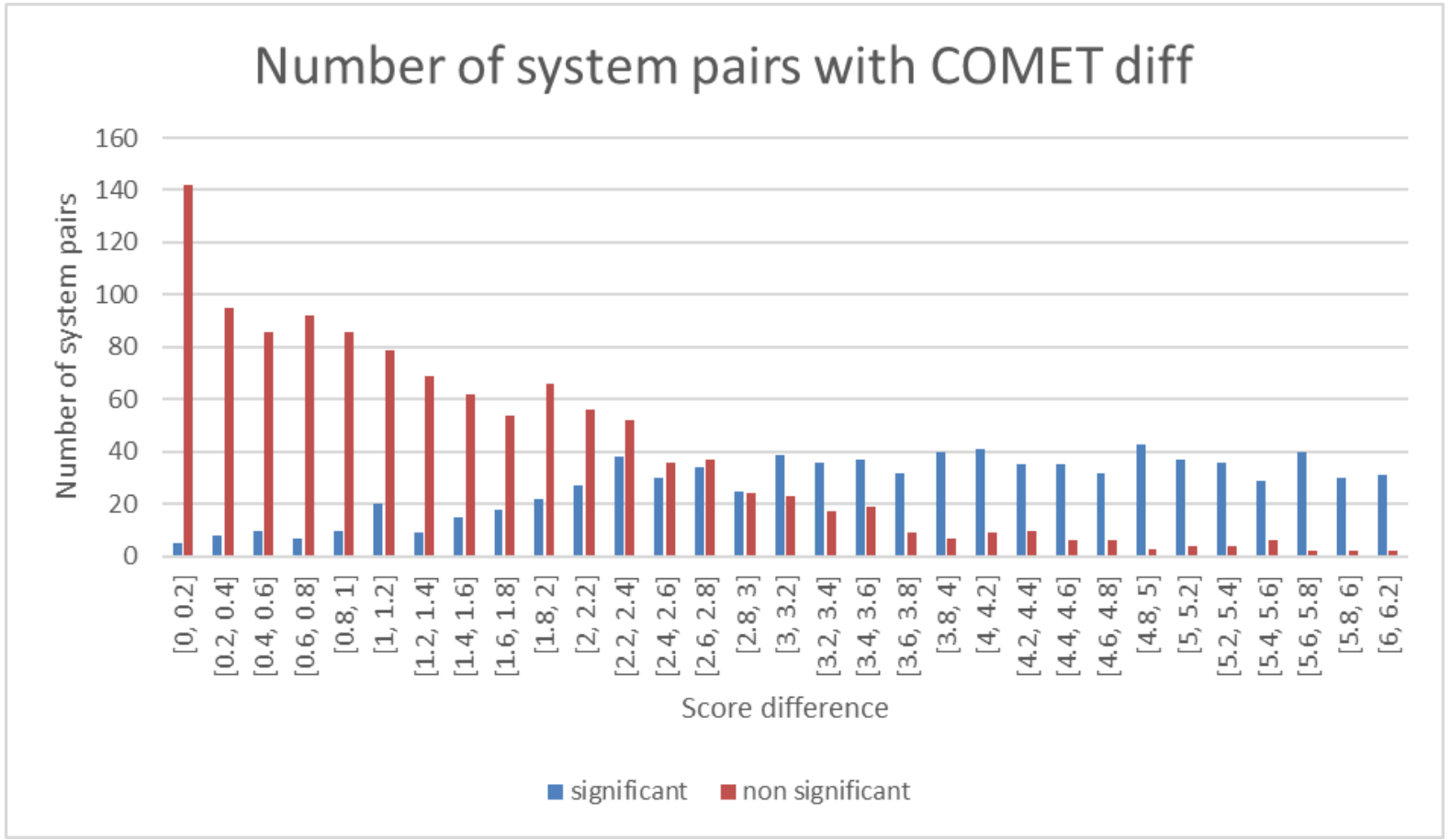
Is small effect size almost always statistically significant?

Is small effect size almost always statistically significant?



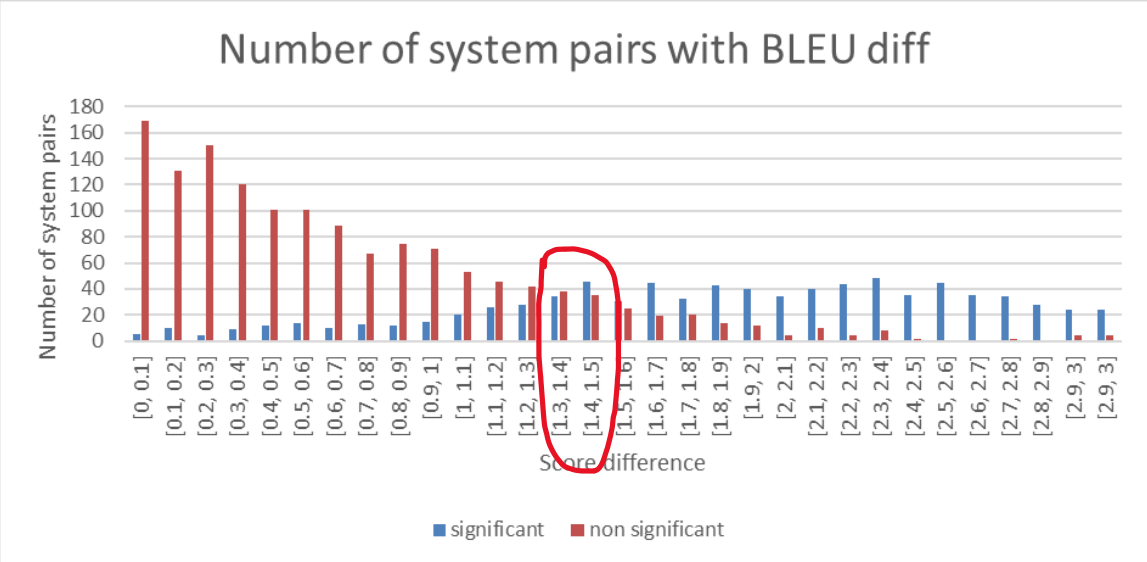
*Limitation: We are comparing various testsets and language pairs with different effect sizes and average testset size is 1000 sents.

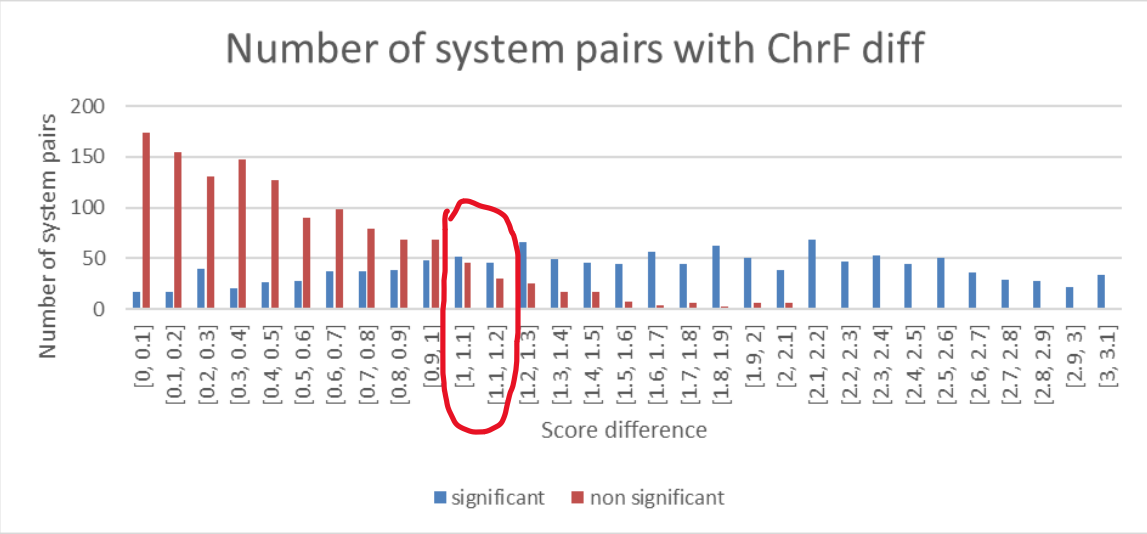
Is small effect size almost always statistically significant?

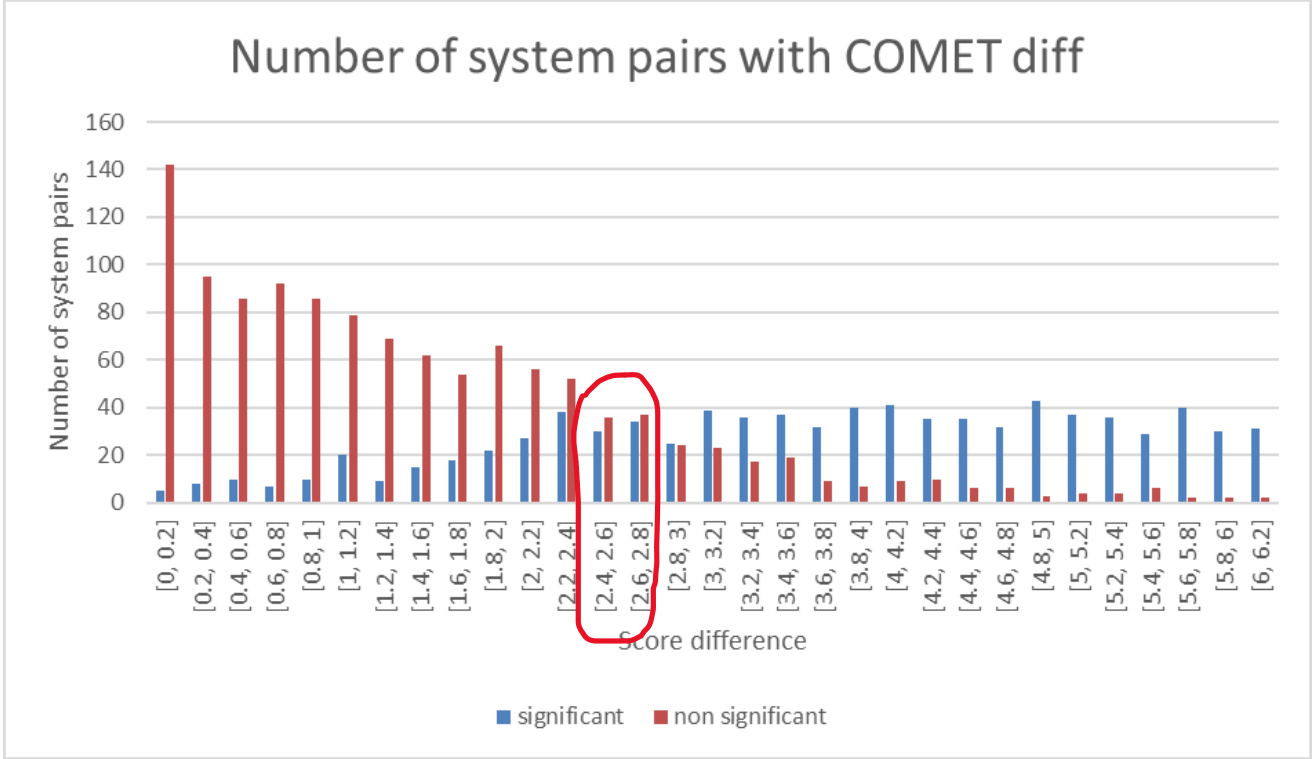


*We use COMET scores multiplied by 100

Rule of thumbs between BLEU, ChrF, COMET effect sizes







COMET needs roughly
2x larger effect sizes

Recommendation:

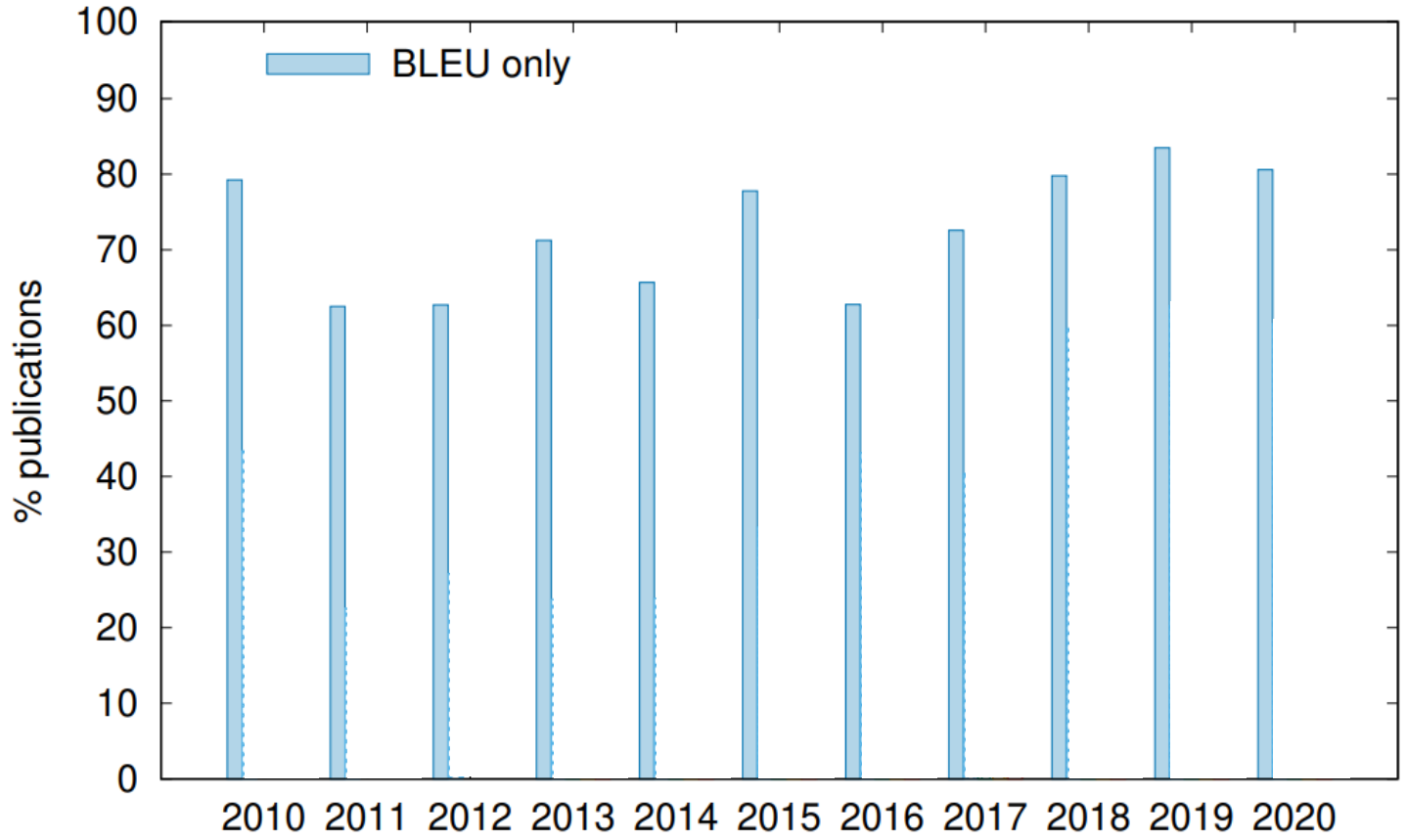
Use statistical significance testing as a source of information.
However, beware of small effect sizes.

Are we overfitting on BLEU?

4

Motivation

99% of papers use BLEU, majority rely solely on it

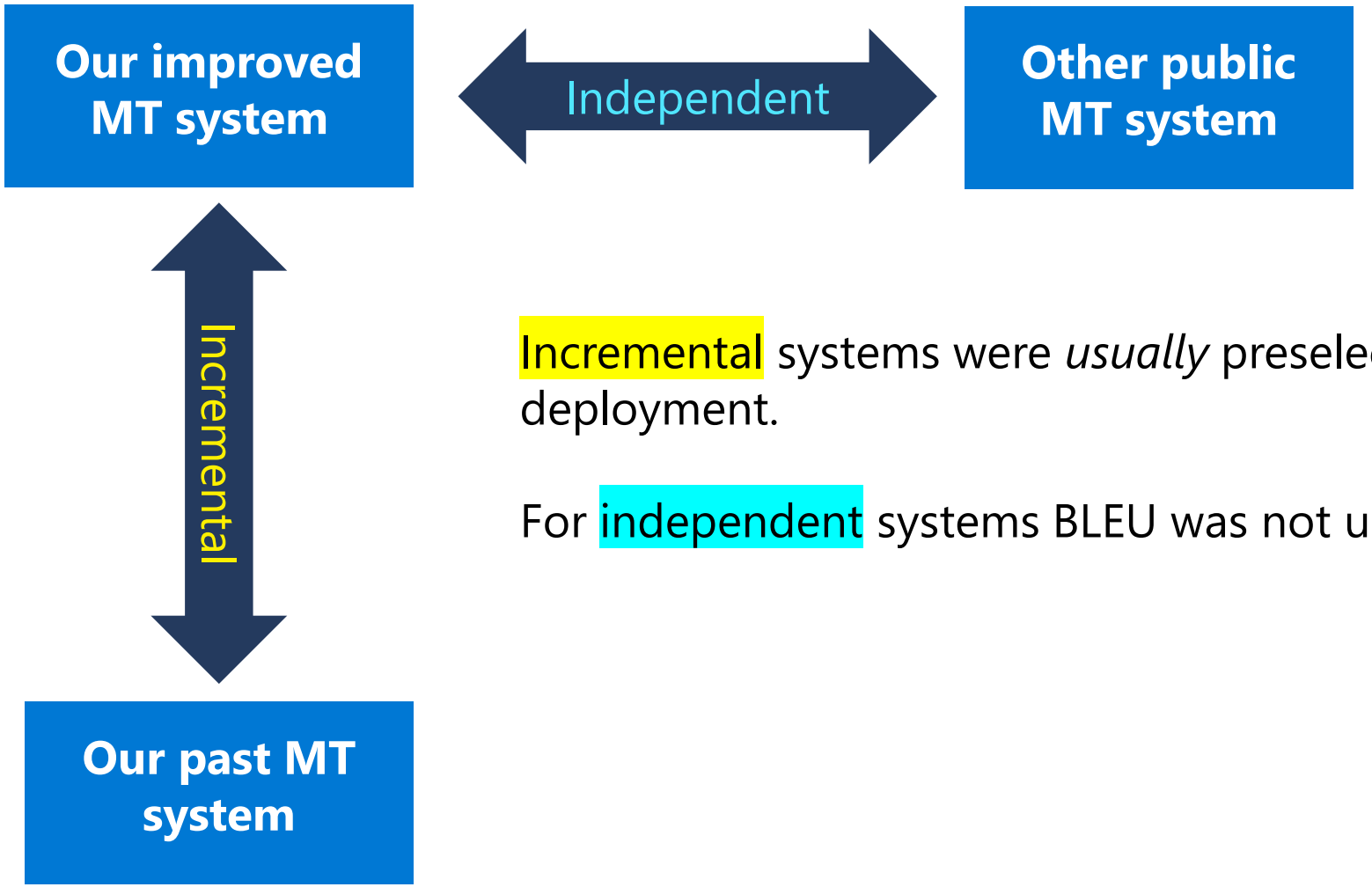


Are we overfitting on BLEU

BLEU has 74.6% accuracy in our evaluation.

Is it possible that it mislead our past research to ignore better models?

Incremental vs. Independent



Incremental systems were *usually* preselected based on BLEU for deployment.

For independent systems BLEU was not used to preselect them.

Are we overfitting on BLEU

system pairs:	Incremental 161	Independent 246
BLEU	99.4	90.7
BERTScore	98.8	91.5
ESIM	98.8	92.3
Prism	98.1	94.3
ChrF	98.1	91.5
COMET	98.1	98.4
COMET-QE	97.5	98.8
CharacTER	97.5	89.8
BLEURT	96.9	93.5
Prism	96.9	92.7
TER	95.7	91.5

An indirect evidence that we rejected improved models due to BLEU degradation.

“When a measure becomes a target, it ceases to be a good measure.”

Marilyn Strathern (1997)
Generalization of Goodhart's law

Recommendations for automatic evaluation

Recommendations for MT Evaluation

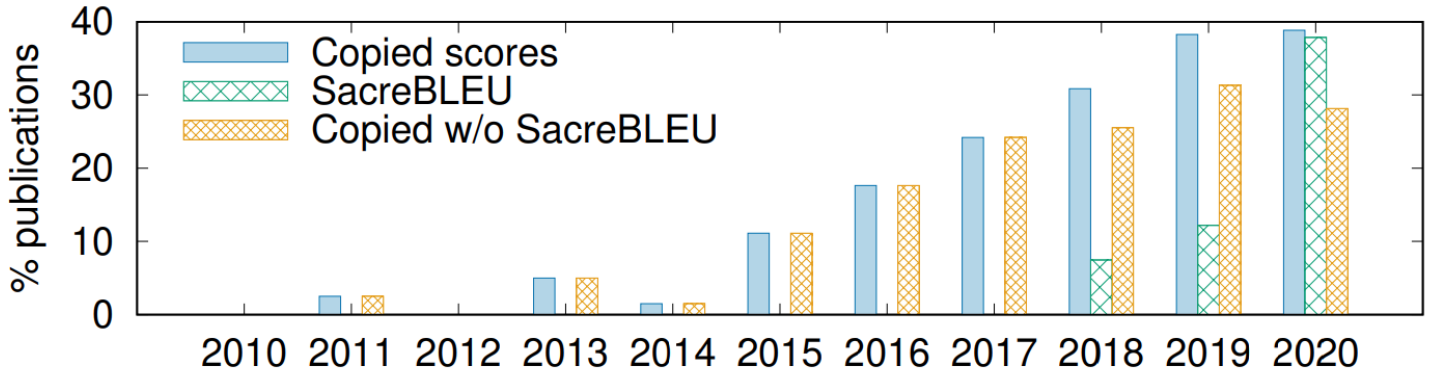
1. Use COMET as the primary metric and ChrF as a secondary metric.

Recommendations for MT Evaluation

2. Run a paired significance test to reduce metric misjudgement.

Recommendations for MT Evaluation

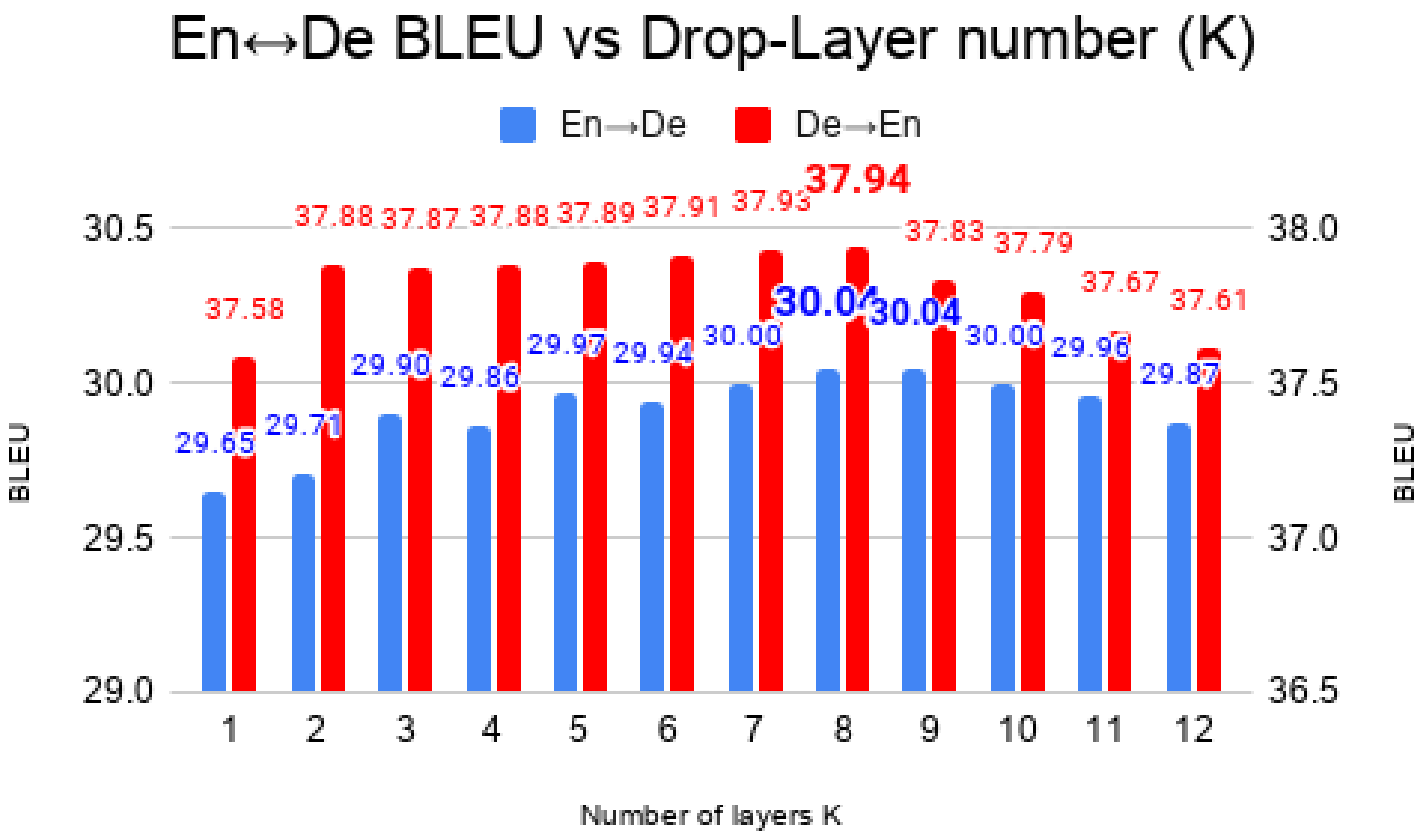
3. Publish system outputs to allow work comparison and recalculation of different metric scores.



Marie et al., 2021

Recommendations for MT Evaluation

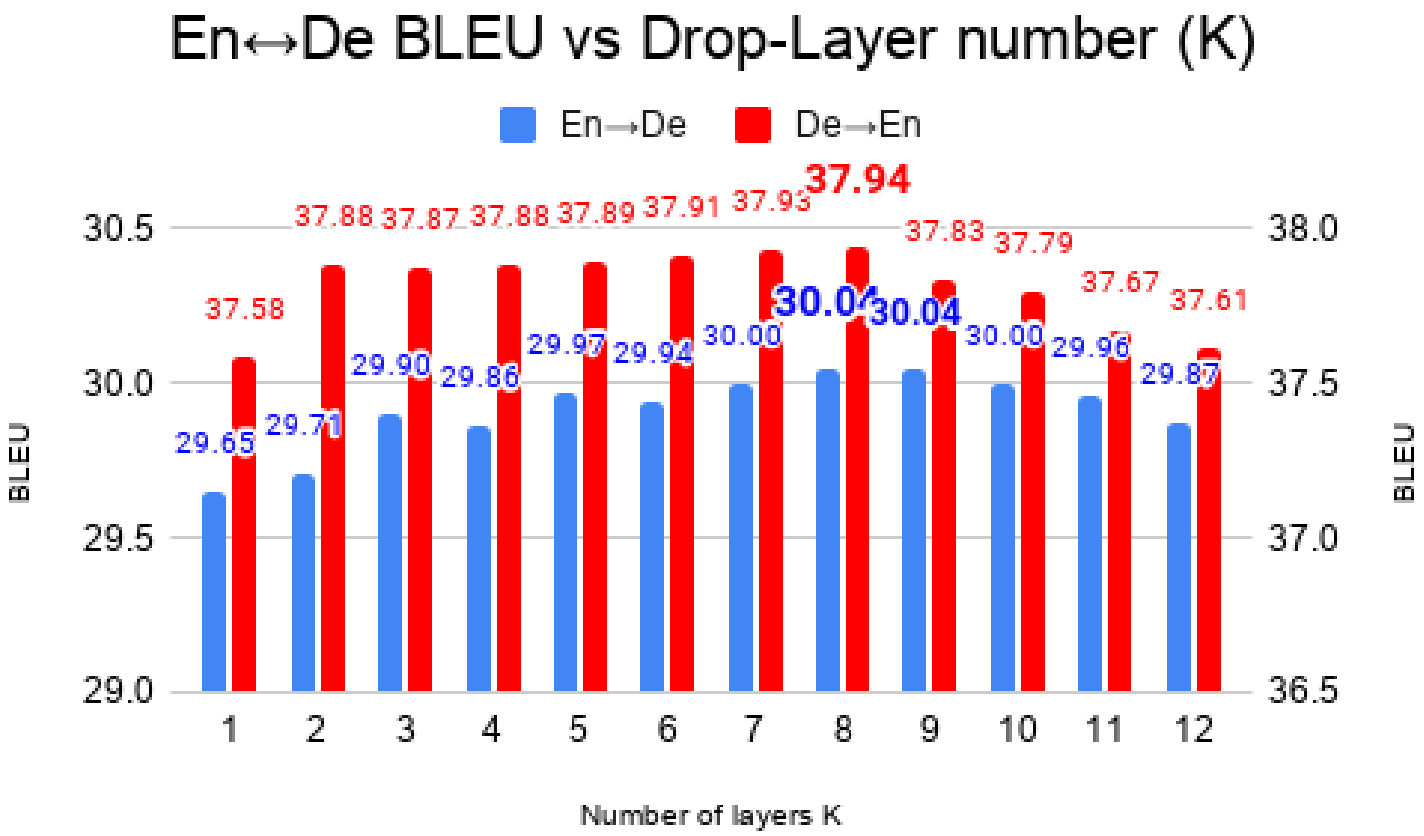
4.



An example of *creative* evaluation from an EMNLP 2021 paper

Recommendations for MT Evaluation

4. Round to **three** significant digits (i.e. single decimal for BLEU, COMET, ...)



An example of *creative* evaluation from an EMNLP 2021 paper

Recommendations for MT Evaluation

1. Use COMET as the primary metric and ChrF as a secondary metric.
2. Run a paired significance test to reduce metric misjudgement.
3. Publish system outputs to allow work comparison and recalculation of different metric scores.
4. Round to three significant digits (i.e. single decimal for BLEU, COMET, ...)
5. If possible, use human evaluation ☺

Microsoft

Amrhein, Chantal, and Rico Sennrich. "**Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET.**" (2022).

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. "**Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain.**" (2021)

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. "**To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.**" (2021)

Marie, Benjamin, Atsushi Fujita, and Raphael Rubino. "**Scientific credibility of machine translation research: A meta-evaluation of 769 papers.**" (2021)

Example of evaluation

Why even automatic evaluation is hard to automatize?

Is new architecture better or worse?

	Diff Comet	Diff ChrF
enu-eti	1.1	1.0
enu-fin	3.5	1.6
enu-heb	0.6	0.9
enu-hun	2.5	0.8
enu-ind	0.6	0.7
enu-lvi	3.8	1.2
enu-msl	-7.4	-0.3
enu-sky	-3.0	0.4
enu-vit	2.2	3.0
eti-enu	-1.4	0.1
fin-enu	1.8	1.0
heb-enu	-0.8	1.3
hun-enu	-1.1	0.0
ind-enu	-1.6	-1.0
lvi-enu	-0.6	0.1
msl-enu	-8.9	-3.4
sky-enu	-2.0	-1.2
vit-enu	-3.1	-1.0

Table show score differences for COMET and ChrF between the previous and the new architecture.

Could the problem be directionality?

	Diff Comet	Diff ChrF	Autentic % sentences
enu-eti	1.1	1.0	100%
enu-fin	3.5	1.6	100%
enu-heb	0.6	0.9	100%
enu-hun	2.5	0.8	100%
enu-ind	0.6	0.7	100%
enu-lvi	3.8	1.2	100%
enu-msl	-7.4	-0.3	100%
enu-sky	-3.0	0.4	100%
enu-vit	2.2	3.0	100%
eti-enu	-1.4	0.1	9%
fin-enu	1.8	1.0	18%
heb-enu	-0.8	1.3	0%
hun-enu	-1.1	0.0	7%
ind-enu	-1.6	-1.0	0%
lvi-enu	-0.6	0.1	23%
msl-enu	-8.9	-3.4	0%
sky-enu	-2.0	-1.2	0%
vit-enu	-3.1	-1.0	0%

From English

Into English

What about the effect size?

	Diff Comet	Diff ChrF	Auth %
enu-eti	1.1	1.0	100%
enu-fin	3.5	1.6	100%
enu-heb	0.6	0.9	100%
enu-hun	2.5	0.8	100%
enu-ind	0.6	0.7	100%
enu-lvi	3.8	1.2	100%
enu-msl	-7.4	-0.3	100%
enu-sky	-3.0	0.4	100%
enu-vit	2.2	3.0	100%
eti-enu	-1.4	0.1	9%
fin-enu	1.8	1.0	18%
heb-enu	-0.8	1.3	0%
hun-enu	-1.1	0.0	7%
ind-enu	-1.6	-1.0	0%
lvi-enu	-0.6	0.1	23%
msl-enu	-8.9	-3.4	0%
sky-enu	-2.0	-1.2	0%
vit-enu	-3.1	-1.0	0%

Most seem to have large enough effect size

Small effect size:
<0.5 for ChrF
<1.0 for Comet

Questionable effect size:
<1.0 for ChrF
<2.0 for Comet

What about statistical significance testing?

	Diff Comet	Diff ChrF	Auth %	p-value Comet	p-value ChrF
enu-eti	1.1	1.0	100%	***	***
enu-fin	3.5	1.6	100%	***	***
enu-heb	0.6	0.9	100%	-	***
enu-hun	2.5	0.8	100%	***	***
enu-ind	0.6	0.7	100%	**	***
enu-lvi	3.8	1.2	100%	***	***
enu-msl	-7.4	-0.3	100%	***	**
enu-sky	-3.0	0.4	100%	***	*
enu-vit	2.2	3.0	100%	***	***
eti-enu	-1.4	0.1	9%	***	-
fin-enu	1.8	1.0	18%	***	***
heb-enu	-0.8	1.3	0%	**	***
hun-enu	-1.1	0.0	7%	***	-
ind-enu	-1.6	-1.0	0%	***	***
lvi-enu	-0.6	0.1	23%	*	-
msl-enu	-8.9	-3.4	0%	***	***
sky-enu	-2.0	-1.2	0%	***	***
vit-enu	-3.1	-1.0	0%	***	***

Most seems to be strongly statistically significant.

* p-value < 0.05
** p-value < 0.01
*** p-value < 0.001

Evaluate authentic monolingual sentences

	Diff Comet	Diff ChrF	Auth %	p-value Comet	p-value ChrF	MonoTest diff	COMET Agreed
enu-eti	1.1	1.0	100%	***	***	1.1	TRUE
enu-fin	3.5	1.6	100%	***	***	1.7	TRUE
enu-heb	0.6	0.9	100%	-	***	0.3	TRUE
enu-hun	2.5	0.8	100%	***	***	2.1	TRUE
enu-ind	0.6	0.7	100%	**	***	0.1	TRUE
enu-lvi	3.8	1.2	100%	***	***		
enu-msl	-7.4	-0.3	100%	***	**	-4.5	TRUE
enu-sky	-3.0	0.4	100%	***	*	-1.2	TRUE
enu-vit	2.2	3.0	100%	***	***	-1.5	FALSE
eti-enu	-1.4	0.1	9%	***	-	-4.8	TRUE
fin-enu	1.8	1.0	18%	***	***	0.6	TRUE
heb-enu	-0.8	1.3	0%	**	***	-0.4	TRUE
hun-enu	-1.1	0.0	7%	***	-	-4.3	TRUE
ind-enu	-1.6	-1.0	0%	***	***	-1.1	TRUE
lvi-enu	-0.6	0.1	23%	*	-	-1.8	TRUE
msl-enu	-8.9	-3.4	0%	***	***	-10.9	TRUE
sky-enu	-2.0	-1.2	0%	***	***	-3.7	TRUE
vit-enu	-3.1	-1.0	0%	***	***	-5.0	TRUE

It differs only for one language pair

Full Picture

	Diff Comet	Diff ChrF	Mono diff	Decision	
enu-eti	1.1	1.0	1.1	Better	small effect size/non significant
enu-fin	3.5	1.6	1.7	Better	
enu-heb		0.9		Better	
enu-hun	2.5	0.8	2.1	Better	
enu-ind		0.7		Better	
enu-lvi	3.8	1.2		Better	
enu-msl	-7.4		-4.5	Worse	
enu-sky	-3.0		-1.2	Worse	
enu-vit	2.2	3.0	-1.5	Ambiguous	
eti-enu	-1.4		-4.8	Worse	
fin-enu	1.8	1.0		Better	
heb-enu		1.3		Better	
hun-enu	-1.1		-4.3	Worse	
ind-enu	-1.6	-1.0	-1.1	Worse	
lvi-enu			-1.8	Worse	
msl-enu	-8.9	-3.4	-10.9	Worse	
sky-enu	-2.0	-1.2	-3.7	Worse	
vit-enu	-3.1	-1.0	-5.0	Worse	

Only a single language pair is ambiguous.

(COMET-QE either has problems with Vietnamese or our monolingual test is broken)