MENLI: Robust Evaluation Metrics from Natural Language Inference

Yanran Chen[1,2] and **Steffen Eger**[2]
[1]Technische Universität Darmstadt, Germany
[2]Natural Language Learning Group (NLLG), https://nl2g.github.io/
Faculty of Technology, Universität Bielefeld, Germany
yanran.chen@stud.tu-darmstadt.de, steffen.eger@uni-bielefeld.de

Transactions of the Association for Computational Linguistics, vol. 11, pp. 804–825, 2023. https://doi.org/10.1162/tacl_a_00576
Action Editor: Benjamin Van Durme. Submission batch: 10/2022; Revision batch: 1/2023; Published 7/2023.

Natural Language Inference

- **Definition:** NLI is the task of determining the relationship between two sentences (Premise versus Hypothesis):
  - **Entailment:** Hypothesis is true based on the premise.
  - **Contradiction:** Hypothesis contradicts the premise.
  - **Neutral:** Hypothesis is neither entailed nor contradicted.
- **Applications:** QA, dialogue systems, summarization, etc.
- **Datasets:** SNLI, ANLI, MultiNLI
- **Evaluation:** BERT, RoBERTa, and DeBERTa

Task		Metrics
MT	ref-based	MoverScore (Zhao et al., 2019), BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), SentSim (Song et al., 2021), COMET (Rei et al., 2020b), BLEURT (Sellam et al., 2020)
	ref-free	COMET, SentSim, XMoverScore (Zhao et al., 2020)
Summarization	ref-based	BARTScore, DiscoScore (Zhao et al., 2023), MoverScore, BERTScore
	ref-free	BARTScore, SUPERT (Gao et al., 2020)

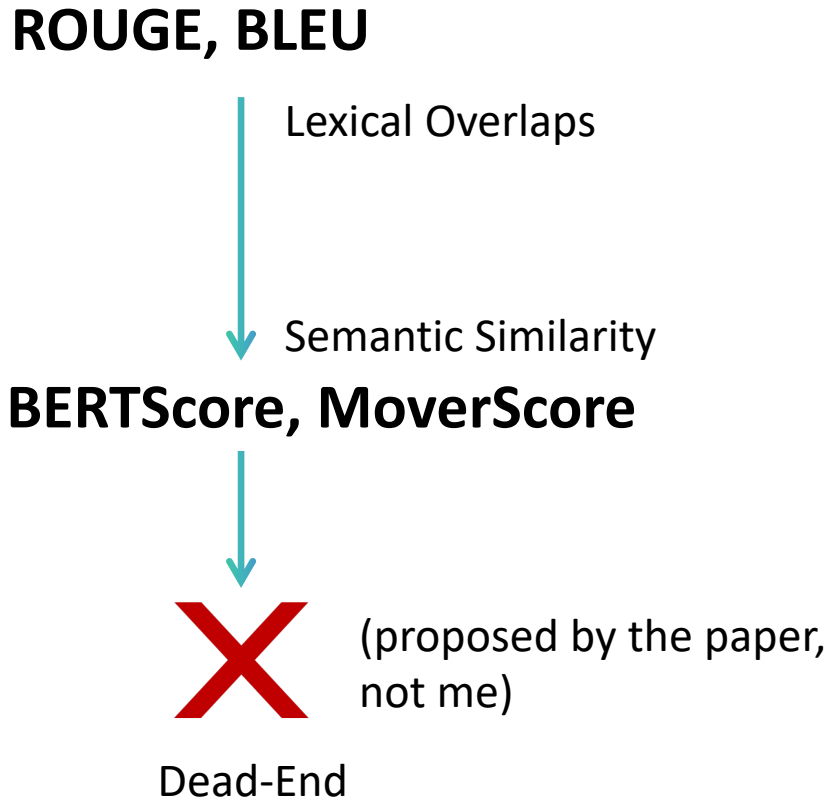
Table 5: Evaluation metrics explored in this work.

- **Lexical Overlaps:** BLEU (n-gram), ROUGE (n-gram), Meteor (n-gram&order), CIDEr (word frequency)
- **Semantic Similarity** (word embedding)**:** BERTScore, MoverScore, SentSim
- **Fine-tuned Models:** BARTScore (propobility), BLEURT, COMET (framwork based on XLM)
- **Semantic Similarity & NLI:** DiscoScore (entity similarity & logical consistent, document-level)

Adversarial Example

Ref: 5 Ukrainian soldiers wounded in Russia

Gen: 50000 Russian soldiers killed in Ukraine



Intuitively Idea based on NLI

Ref: 5 Ukrainian soldiers wounded in Russia

Bi-Implication

Gen: 50000 Russian soldiers killed in Ukraine

Adversarial Setup

- **src**: Source Text
- **ref**: Reference Text
- **r**: Reference Text (generated by Google Translation and src)
- **cond$_{para}$**: maximally dissimilar to ref/src but meaning-equivalent (mainly by PAWS dataset)
- **cond$_{adv}$**: maximally similar to ref/src but with key error (advasarial attack)

	Number error	Negation error
src	Der bilaterale Handel wurde auf über **100 Milliarden Dollar** im Jahr gesteigert.	Die Wirtschaft der Entwicklungs- und Schwellenländer **wird schwach bleiben**.
ref	Bilateral trade has increased to more than **$100 billion** a year.	Emerging economies **will remain weak**.
r (google translation of *src*)	Bilateral trade has increased to over **$100 billion** a year.	The economies of developing and emerging countries **will remain weak**.
$cand_{para}$	Bilateral trade has increased to more than **one hundred billion dollars** a year.	Emerging markets **will remain weak**.
$cand_{adv}$(ref-based)	Bilateral trade has increased to more than **$814 billion** a year.	Emerging economies **won't remain weak**.
$cand_{adv}$(ref-free)	Bilateral trade has increased to over **$478 billion** a year.	The economies of developing and emerging countries **won't remain weak**.

Table 2: Examples of our adversarial test suite taken from WMT20$_{de}$. Red words indicate specific adversarial perturbations of the words in green. $cand_{adv}$(ref-based) builds on *ref*, whereas $cand_{adv}$(ref-free) builds on *r* (indicated by corresponding coloring in the first column). The preferences we query for are given in Eq. (1).

Adversarial Setup

- src: Source Text
- ref: Reference Text
- r: Reference Text (generated by Google Translation and src)
- $\text{cond}_{\text{para}}$: maximally dissimilar to ref/src but meaning-equivalent
- cond_{adv}: maximally similar to ref/src but with key error (advasary attack)

Expect a good matric:

$$\text{ref-based} : m(ref, cand_{para}) > m(ref, cand_{adv})$$
$$\text{ref-free} : m(src, ref) > m(src, cand_{adv}) \quad (1)$$

Adversarial Setup

- Addition: 'I love dogs' → 'I love dogs and cats.'
- Omission: Randomly drop ~1%–20% words in the sentence.
- Mismatch: Consider mismatching nouns, verbs, and adjectives
- Negation: Add/remove negations to/from the verb
- Number error: Replace all numbers (except for those related to dates) in the sentence with random numbers in the same format
- Pronoun error: 'he' to 'she' and 'us' to 'them'.
- Name error: Replace exactly one name with a random one of the same gender.
- Fluency:
  - Jumbling word order: Randomly shuffle the word order in a sentence.
  - Spelling error: Add a typo to a word in a sentence.
  - Subject-verb disagreement: Make the subject and verb disagree (e.g., ''He like dogs.'').

Adversarial Setup

Error	Source	MT hypothesis
Mismatch/verb	关注苏宁易购服务号	**Pay attention to** (**Follow**) Suning.com service account
Mismatch/adj.	还不错，玩游戏的画质是真的香	Not bad, the picture quality of playing games is really **fragrant** (**good**)
Pronoun/Addition	买给儿子的，他说很好。	Bought it for **his** (**my**) son, he said it was good.
Name	当天，美国运输部长赵小兰、美联邦众议员孟昭文以及国际领袖基金会创会会长董继玲等分别在会上发言。	On the same day, US Secretary of Transportation **Zhao Xiaolan** (**Elaine Lan Chao**), US Congressman **Meng Zhaowen** (**Grace Meng**) and **Dong Jiling** (**Chiling Tong**), founding president of the International Leaders Foundation, spoke at the meeting respectively.
Omission	I'll **review** your account, one moment, please.	Ich werde Ihr Konto **[...]** (**überprüfen**), einen Moment bitte.
Mismatch/noun	Listen, I don't want to make **my people** mad," she said.	„Hör zu, ich will **mein Volk** (**meine Leute**) nicht verrückt machen‘‘, sagte sie.
Pronoun	Williams wasn't the only one who received a fine at this year's Wimbledon, though **hers** was the most costly.	Williams war nicht die einzige, die beim diesjährigen Wimbledon eine Geldstrafe erhielt, obwohl **sie** (**ihre**) die teuerste war.

Table 4: Examples of errors in WMT MQM annotations for Chinese-to-English and English-to-German. Red texts are the annotated errors (‘‘[...]’’ indicates the missing translation) and the green texts in the bracket refer to a more correct translation accordingly; the green texts in source sentences denote the parts being mistranslated or omitted.

Datasets

Task		Datasets
MT	segment-level	WMT15-17, WMT20-21
	system-level	WMT20-21
	adversary	*ref-based*: $\text{PAWS}_{\text{ori/back}}$, WMT20_{de}, XPAWS_{de}; *ref-free*: $\text{XPAWS}_{\text{de/fr/zh/ja}}$, WMT20_{de}
Summarization	summary-level	RealSum (Bhandari et al., 2020)
	system-level	RealSum, SummEval
	adversary	SE_{adv}, Rank19 (Falke et al., 2019) (ref-free only)

Table 6: We use the to-English language pairs in WMT15-17 datasets (Stanojević et al., 2015; Bojar et al., 2016, 2017). In segment-level evaluation on WMT20-21 (Mathur et al., 2020b; Freitag et al., 2021a,b), we use the data with MQM scores for zh-en, while in system-level evaluation, we correlate the metrics with DA scores for all to-English language pairs. The datasets for system-level evaluation before WMT20 are skipped, as all metrics mostly get very high correlations on them.

On **MT** standard benchmarks, we evaluate the metrics on both *segment-level* (where we correlate metrics scores to human judgments for individual sentences/segments in the datasets) and *system-level* (where we correlate the average metric scores to the average human scores over the segments generated by each system), using Pearson correlation as the performance indicator. On SummEval for **summarization**, we compute Kendall correlation with system-level human judgements on four criteria: *coherence*, *consistency*, *fluency* and *relevance* (we apply two aggregation methods for the multi-reference setting, *max* and *mean*). We calculate Pearson correlation with both summary-level (analogous to segment-level in MT) and system-level *LitePyramids* (Shapira et al., 2019) human ratings in RealSumm.

NLI as a Metric

3 Scores:

- e
- n
- c

5 Formulas:

- e
- -c
- e-n
- e-c
- e-n-2c

3 Directions:

- ref/src → cand (implication)
- ref/src ← cand (reverse implication)
- ref/src ↔️ cand (bi-implication)

NLI as a Metric

3 Scores:

- e
- n
- c

5 Formulas:

- e
- -c
- e-n
- e-c
- e-n-2c

3 Directions:

- ref/src → cand (implication)
- ref/src ← cand (reverse implication)
- ref/src ↔️ cand (bi-implication)

NLI as a Metric

3 Scores:

- e
- n
- c

5 Formulas:

- e
- -c
- e-n
- e-c
- e-n-2c

3 Directions:

- ref/src → cand (implication)
- ref/src ← cand (reverse implication)
- ref/src ↔ cand (bi-implication)

Pooling Strategy Selection
(on MT evaluation)

(a) Reference-based

	e	-c	e-n	e-c	e-n-2c
ref→cand	3+0	3+0		2+0	
ref←cand					
ref↔cand	0+4		0+3	0+1	0+2

(b) Reference-free

	e	-c	e-n	e-c	e-n-2c
src→cand		2+0			
src←cand	0+1		0+2		
src↔cand	0+1		4+6	4+0	

Table 7: Winning frequency of different pooling strategies for NLI metrics on adversarial (first entry) and MT datasets (second entry). We only show non-zero entries.

Left: Accuracy on advasarial dataset.
Right: Correlation with human judgements on MT datasets

NLI as a Metric

Best Strategy

- MT: e
- Summarization:
  - (1) e-c from direction ref←cand performs best for ref-based NLI metrics
  - (2) -c from direction src→cand is the best strategy for ref-free NLI metrics.

Pooling Strategy Selection
(on MT evaluation)

(a) Reference-based

	e	-c	e-n	e-c	e-n-2c
ref→cand	3+0	3+0		2+0	
ref←cand					
ref↔cand	0+4		0+3	0+1	0+2

(b) Reference-free

	e	-c	e-n	e-c	e-n-2c
src→cand		2+0			
src←cand	0+1		0+2		
src↔cand	0+1		4+6	4+0	

Table 7: Winning frequency of different pooling strategies for NLI metrics on adversarial (first entry) and MT datasets (second entry). We only show non-zero entries.

Left: Accuracy on advasarial dataset.
Right: Correlation with human judgements on MT datasets

NLI Systems We explore both monolingual and cross-lingual NLI-based metrics. For each setup, we choose two NLI models, which are obtained from Hugging Face or fine-tuning by ourselves.

For **monolingual NLI metrics**, we choose (1) a RoBERTa-large model (Liu et al., 2019) fine-tuned on SNLI (Bowman et al., 2015), MNLI, Fever (Nie et al., 2019) and ANLI (Nie et al., 2020) by Nie et al. (2020) and (2) a DeBERTa-large model fine-tuned by He et al. (2021), using MNLI. We denote the NLI metrics induced from these two models as `NLI-R` and `NLI-D`. They will be used for ref-based MT evaluation, and both ref-based and -free summarization evaluation tasks. Note that, while NLI-R has been fine-tuned on adversarial NLI (ANLI), which has been shown to increase robustness on (for example) negation and numerical reasoning, NLI-D has not been trained on ANLI. **Cross-lingual NLI metrics** should handle premises and hypotheses in different languages, so we select the multilingual versions of the underlying models of `NLI-R`/`NLI-D`. (1) We fine-tune a XLM-RoBERTa-base model (Conneau et al., 2019), using the datasets for fine-tuning `NLI-R` as well as XNLI dataset (Conneau et al., 2018). (2) We select an mDeBERTa-base model fine-tuned on MNLI and XNLI. We denote the corresponding cross-lingual NLI metrics as `XNLI-R` and `XNLI-D`.

MT Results

	Adv.				MT			
	ref-based		ref-free		ref-based		ref-free	
	all	adeq.	all	adeq.	seg	sys	seg	sys
Supervised								
COMET	67.4	67.0	76.8	74.5	0.676	0.808	0.620	0.698
BLEURT	74.8	79.8	–	–	0.708	0.807	–	–
Unsupervised								
sentBLEU	32.9	27.2	–	–	0.380	0.757	–	–
Rouge	34.3	28.7	–	–	0.425	0.774	–	–
MoverScore	48.3	46.9	–	–	0.567	**0.806**	–	–
XMoverS(UMD)			74.5	71.7	–	–	0.400	0.672
XMoverS(CLP)	–	–	73.8	70.9	–	–	0.422	**0.673**
BERTS	65.3	60.9	–	–	**0.620**	0.799	–	–
BARTS-P	67.4	64.2	–	–	0.587	0.761		
BARTS-F	78.4	77.8	–	–	0.593	0.802	–	–
SentS(BERTS)	68.1	67.8	62.7	65.5	0.612	0.401	0.421	−0.021
SentS(WMD)	62.1	61.9	63.0	65.8	0.607	–	**0.427**	–
NLI-based								
X(NLI)-R	85.0	92.1	70.5	75.8	0.451	0.756	0.221	0.335
X(NLI)-D	**86.6**	**92.3**	**79.3**	**85.8**	0.439	0.770	0.149	0.581

Table 8: Pearson correlation with human judgments in WMT and accuracy (%) on our adversarial datasets, averaged over datasets. The performance of ref-based COMET is averaged over $WMT20_{de}$ and $XPAWS_{de}$, since it also requires source texts as input. In bold: best results among all unsupervised metrics including the NLI-based metrics.

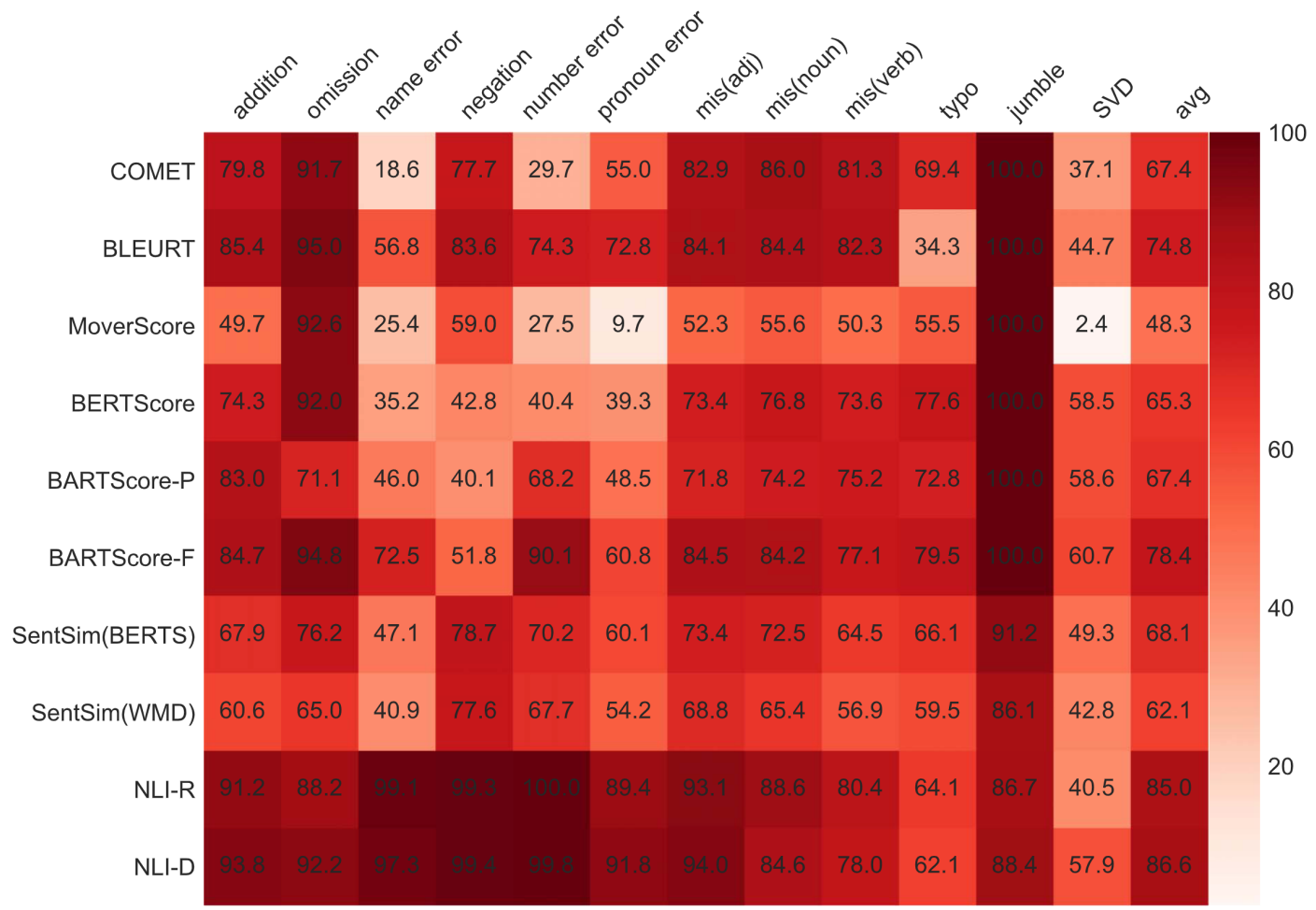


Figure 1: Average accuracy (values in each block) of all metrics per phenomenon over the adversarial datasets for ref-based MT evaluation. Darker color indicates higher accuracy and vice versa.

Summarization Results

(a) Reference-based

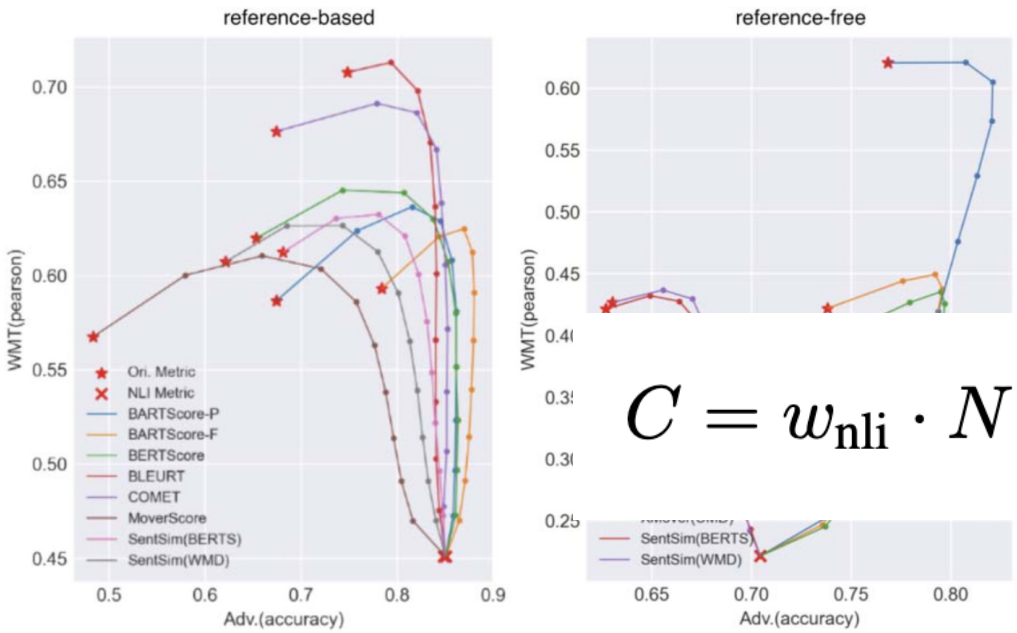
metric	SummEval										RealSumm		Adv.	
	coherence		consistency		fluency		relevance		avg		litePyr		SE_{adv}	
	mean	max	mean	max	mean	max	mean	max	mean	max	sum	sys	all	adeq.
BLEU	0.294	0.279	0.044	−0.029	0.244	0.229	0.397	0.382	0.245	0.215	0.480	0.124	0.182	0.109
Rouge	0.191	0.176	0.088	−0.279	−0.037	−0.081	0.118	0.103	0.090	−0.020	0.540	0.457	0.185	0.117
MoverS	0.206	0.324	0.456	0.103	0.421	0.362	0.368	0.515	0.363	0.326	**0.585**	0.501	0.287	0.251
BERTS	0.618	0.618	0.221	0.044	0.273	0.185	0.603	0.515	0.429	0.340	0.574	0.380	0.598	0.574
BARTS-P	0.485	0.441	0.176	−0.044	0.376	0.185	0.500	0.368	0.385	0.237	0.478	0.531	0.697	0.692
BARTS-F	0.515	**0.647**	0.206	0.250	0.317	0.450	0.529	**0.632**	0.392	0.495	0.583	0.687	0.788	0.792
DiscoS	**0.676**	0.279	0.279	0.676	0.539	0.554	**0.632**	0.353	**0.532**	0.466	−0.199	−0.066	0.334	0.294
NLI-based														
NLI-R	0.147	0.074	0.632	0.676	0.494	0.450	0.279	0.206	0.388	0.352	0.525	**0.856**	**0.864**	**0.905**
NLI-D	0.250	0.265	**0.706**	**0.750**	**0.568**	**0.613**	0.471	0.397	0.499	**0.506**	0.489	0.840	0.806	0.843

(b) Reference-free

metric	SummEval					RealSumm		Adv.			
	coherence	consistency	fluency	relevance	avg	litePyr		SE_{adv}		Rank19	
						summary	system	all	adeq.		avg
BARTS-FN	**0.735**	0.132	0.391	**0.662**	**0.480**	0.178	−0.023	0.427	0.389	0.796	0.612
SUPERT	0.147	0.603	**0.465**	0.279	0.374	**0.522**	0.626	0.296	0.273	0.668	0.482
NLI-based											
NLI-R	0.221	0.235	0.391	0.500	0.337	0.300	**0.688**	**0.720**	**0.722**	0.866	**0.793**
NLI-D	0.162	**0.647**	0.332	0.324	0.366	−0.076	0.568	0.624	0.629	**0.885**	0.755

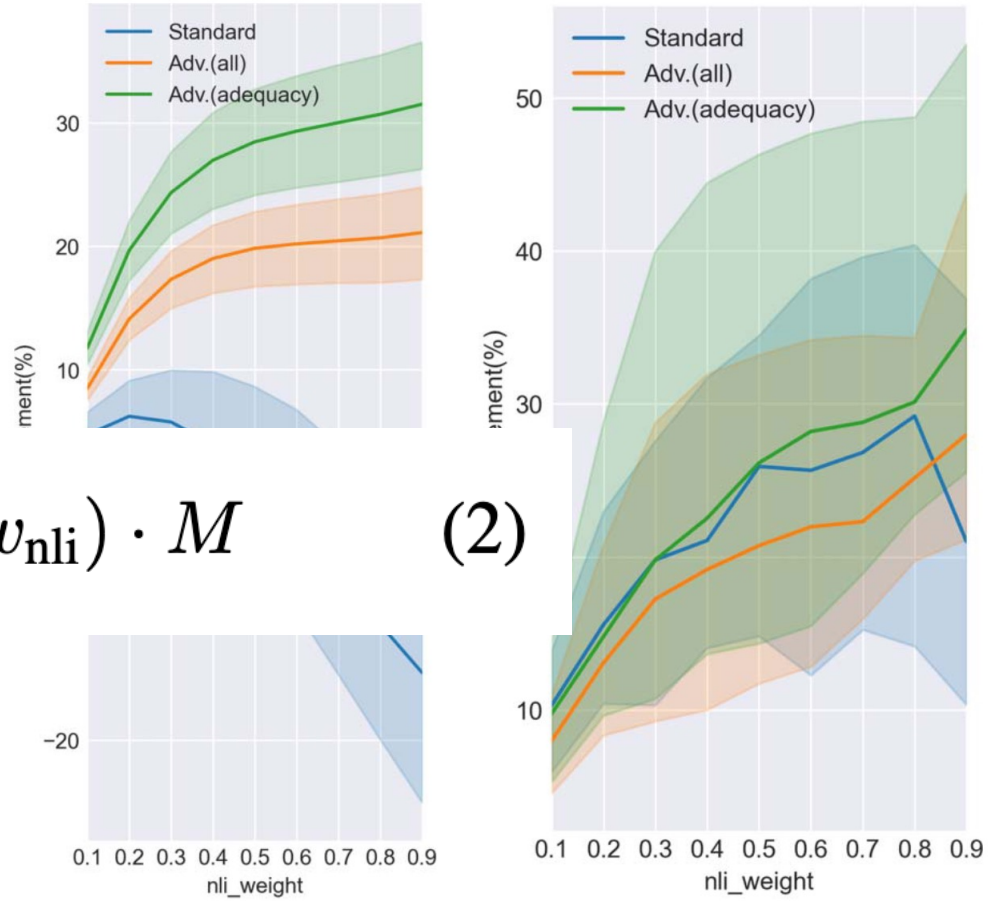
Table 9: Kendall correlation with system-level human judgments in SummEval. Pearson correlation with summary/system-level litePyramid in RealSumm. Accuracy on adversarial benchmarks, averaged over phenomena in SE_{adv}. We bold the best performance on each criterion. ‘‘max/mean’’ denotes the aggregation method used for multi-reference setting in ref-based evaluation on SummEval.

Combined Matrics



$$C = w_{\text{nli}} \cdot N + (1 - w_{\text{nli}}) \cdot M \qquad (2)$$

Figure 2: Accuracy on adversarial datasets and Pearson correlation with **segment-level** human judgements in WMT datasets of combined metrics with (X)NLI-R, averaged over datasets. The points on each path from the original metric to the NLI metric indicate $w_{\text{nli}} = 0, 0.1, \ldots, 1$. The purple line denoting the combination with ref-based COMET ends at another point since the corresponding adversarial performance is averaged over the 2 adversarial datasets containing source texts.



(a) MT (b) Summarization

Figure 3: Improvements of all metrics on standard benchmarks and adversarial datasets for $w_{\text{nli}} = 0.1$, ... 0.9, averaged over all experiments. We show 95% confidence interval.

Combined Matrics

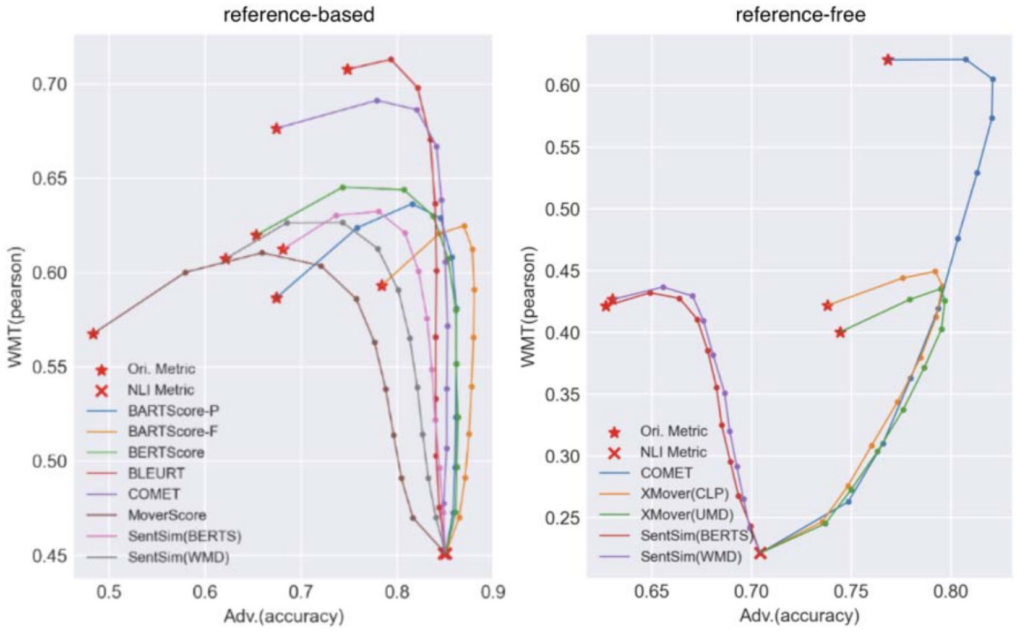
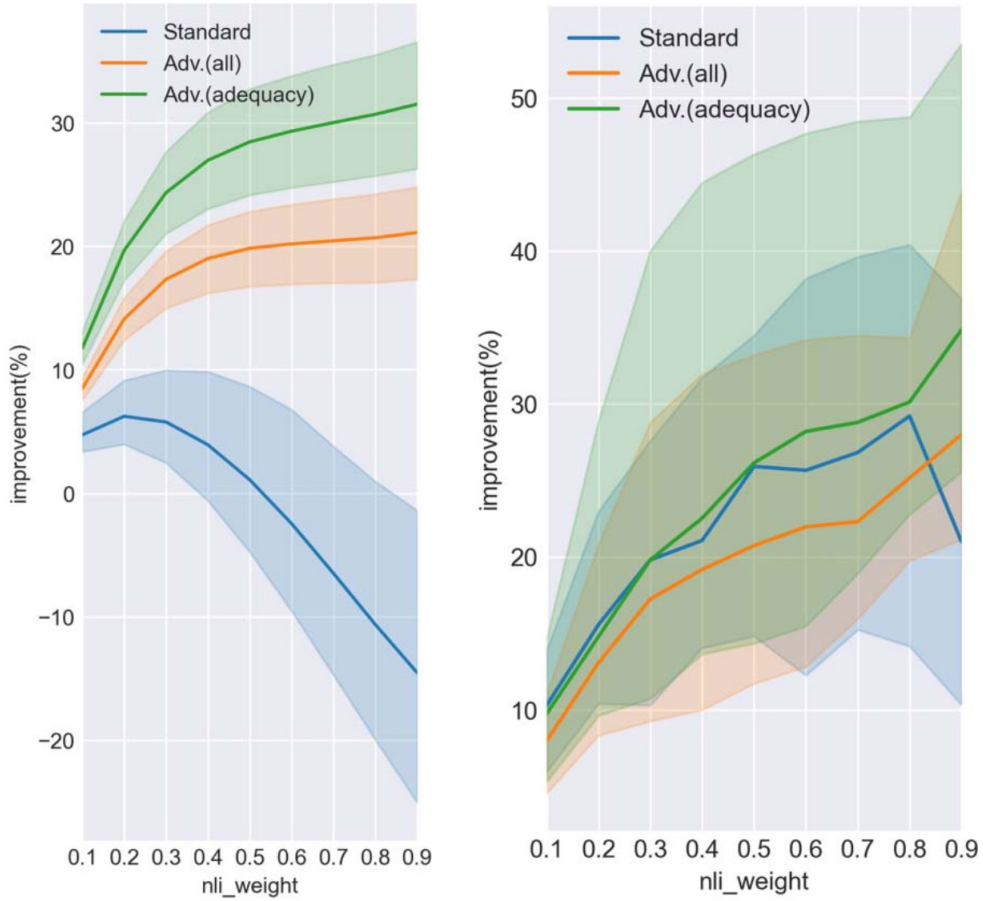


Figure 2: Accuracy on adversarial datasets and Pearson correlation with **segment-level** human judgements in WMT datasets of combined metrics with (X)NLI-R, averaged over datasets. The points on each path from the original metric to the NLI metric indicate $w_{\text{nli}} = 0, 0.1, \ldots, 1$. The purple line denoting the combination with ref-based COMET ends at another point since the corresponding adversarial performance is averaged over the 2 adversarial datasets containing source texts.



(a) MT (b) Summarization

Figure 3: Improvements of all metrics on standard benchmarks and adversarial datasets for $w_{\text{nli}} = 0.1$, ... 0.9, averaged over all experiments. We show 95% confidence interval.

Combined Matrics

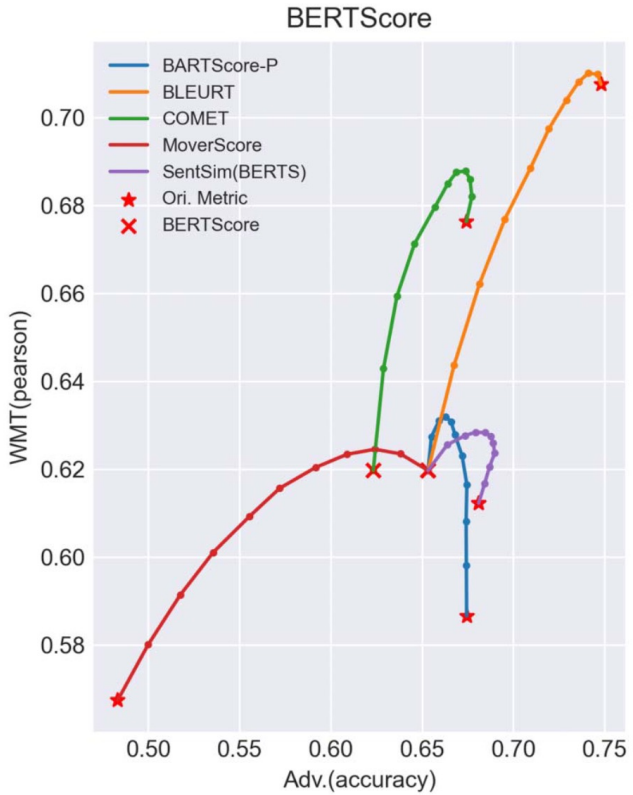


Figure 4: Accuracy on adversarial datasets and Pearson correlation with segment-level human judgements in WMT datasets of combined metrics with BERTScore, averaged over datasets. The green line denoting the combination with COMET ends at another point since the corresponding adversarial performance is only averaged over the 2 adversarial datasets containing source texts.