From COMET to COMES
Can Summary Evaluation Benefit from Translation Evaluation?

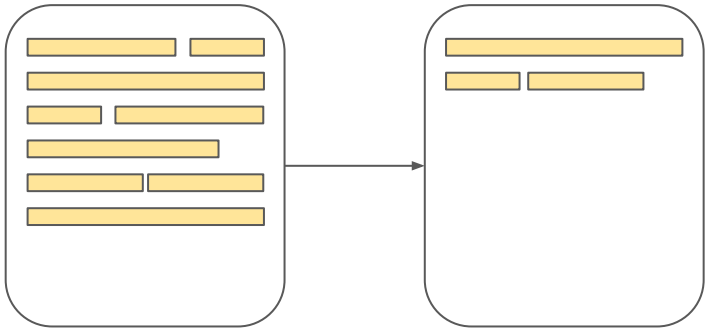
Mateusz Krubiński and Pavel Pecina

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

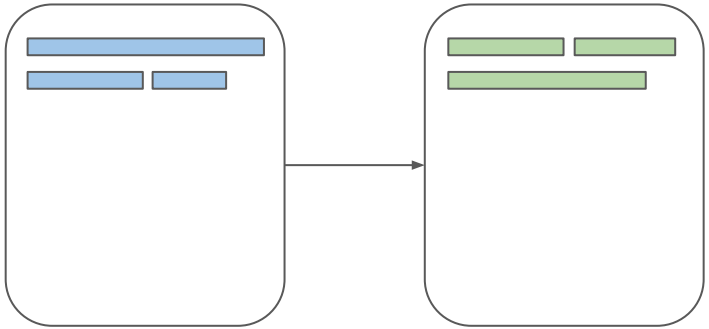
Motivations

Text Summarization vs Machine Translation

Summarization



Machine Translation

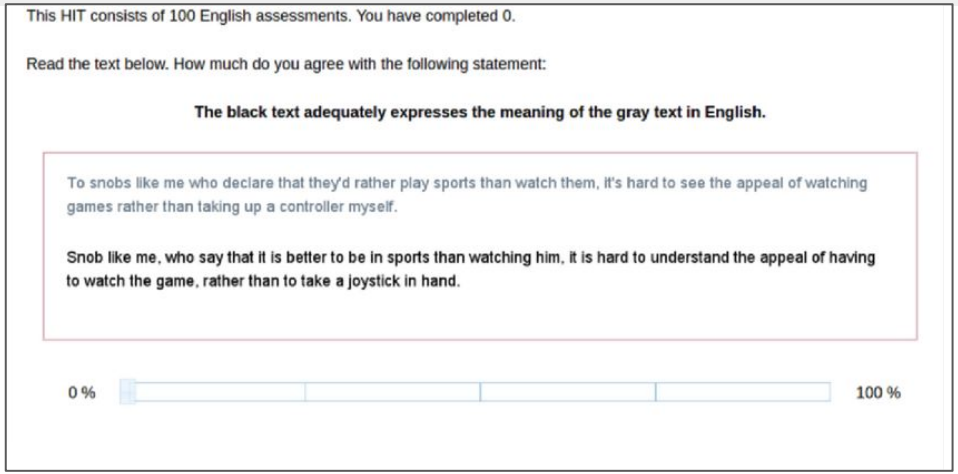


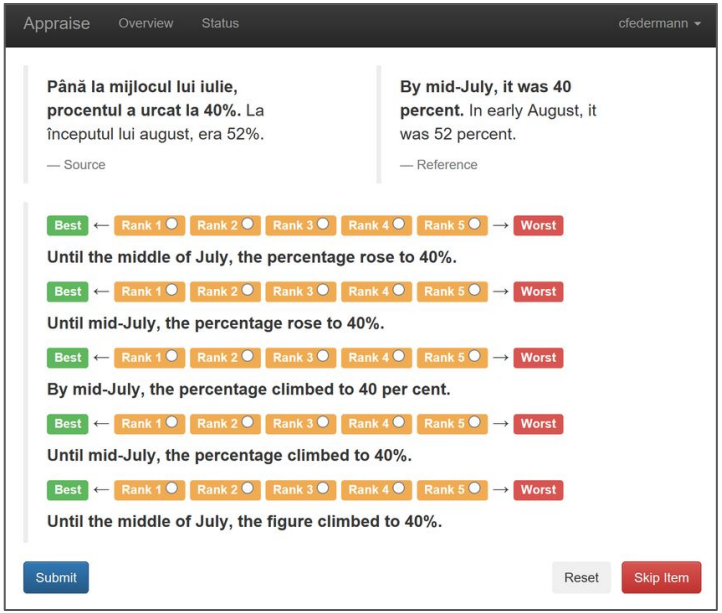
	Summarization	Machine Translation
Same Language	✓	✗
Comparable Length	✗	✓
Evaluation	?	?

Human evaluation for MT

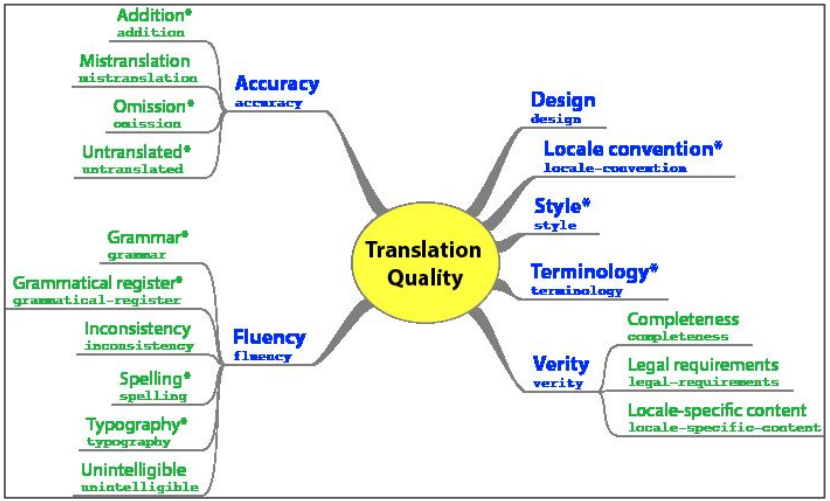
Human evaluation for MT

DA





RR

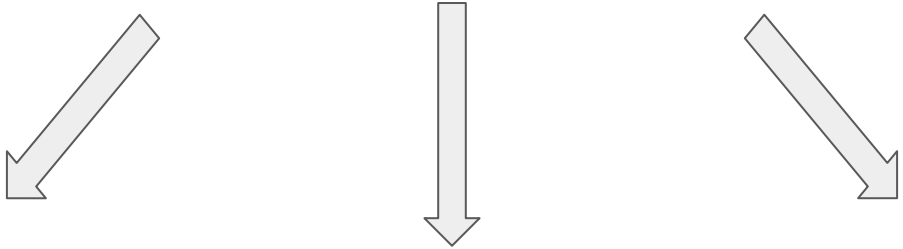


MQM

Fretag et al., 2021, Bojar et al., 2016

Human evaluation for Summarization

...



Likert Scale

- Fabbri et al., 2021
- Stiennon et al., 2020
- ...

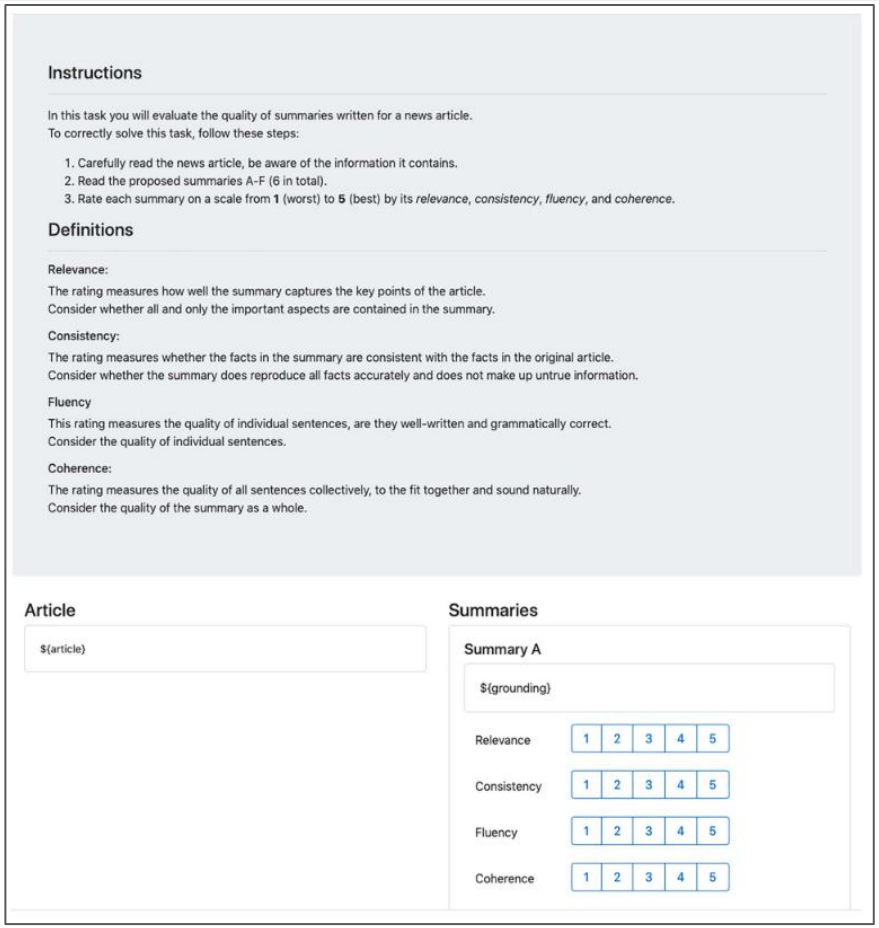
Direct Assessment

- Koto er al., 2021
- ...

Indirect

- Maynez et al., 2020;
- Bhandari et al., 2020
- ...

Human evaluation for Summarization - Likert Scale



Instructions

In this task you will evaluate the quality of summaries written for a news article.
To correctly solve this task, follow these steps:

1. Carefully read the news article, be aware of the information it contains.
2. Read the proposed summaries A-F (6 in total).
3. Rate each summary on a scale from **1** (worst) to **5** (best) by its *relevance*, *consistency*, *fluency*, and *coherence*.

Definitions

Relevance:
The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.

Consistency:
The rating measures whether the facts in the summary are consistent with the facts in the original article.
Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

Fluency
This rating measures the quality of individual sentences, are they well-written and grammatically correct.
Consider the quality of individual sentences.

Coherence:
The rating measures the quality of all sentences collectively, to the fit together and sound naturally.
Consider the quality of the summary as a whole.

Article

${article}

Summaries

Summary A

${grounding}

Relevance	1	2	3	4	5
Consistency	1	2	3	4	5
Fluency	1	2	3	4	5
Coherence	1	2	3	4	5

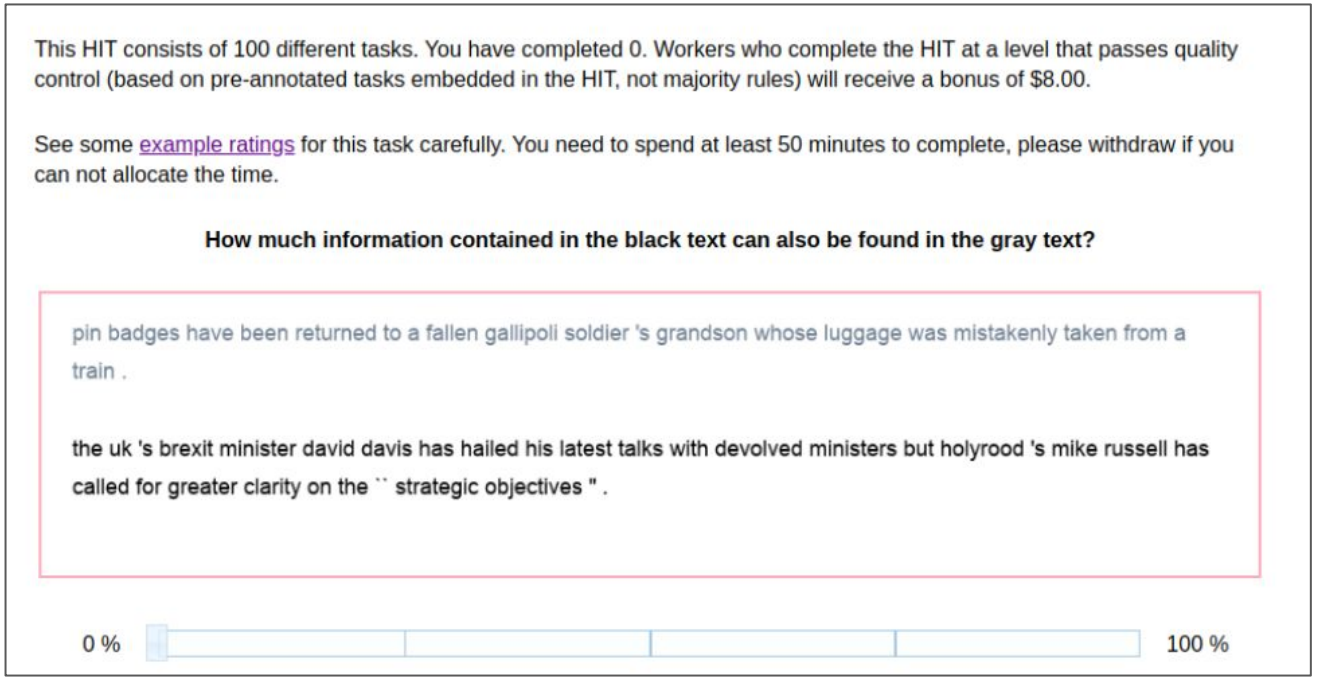
...

Likert Scale

- Fabbri et al., 2021
- Stiennon et al., 2020
- ...

Fabri et al., 2021

Human evaluation for Summarization - Direct Assessment



...

Direct Assessment

- Koto er al., 2021
- ...

Koto et al., 2021

Human evaluation for Summarization - Indirect

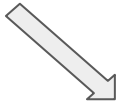
(a) Reference Summary: Bayern Munich beat Porto 6 - 1 in the Champions League on Tuesday. Pep Guardiola's side progressed 7 - 4 on aggregate to reach semi-finals. Thomas Muller scored 27th Champions League goal to pass Mario Gomez. Muller is now the leading German scorer in the competition. After game Muller led the celebrations with supporters using a megaphone.

(b) System Summary (BART, Lewis et al. (2019)): Bayern Munich beat Porto 6 - 1 at the Allianz Arena on Tuesday night. Thomas Muller scored his 27th Champions League goal. The 25 - year - old became the highest - scoring German since the tournament took its current shape in 1992. Bayern players remained on the pitch for some time as they celebrated with supporters.

(c) SCUs with corresponding evaluations:

- Bayern Munich beat Porto. ✓
- Bayern Munich won 6 - 1. ✓
- Bayern Munich won in Champions League. ✓
- Bayern Munich won on Tuesday. ✓
- Bayern Munich is managed by Pep Guardiola. ×
- Bayern Munich progressed in the competition. ✓
- Bayern Munich reached semi-finals. ×
- Bayern Munich progressed 7 - 4 on aggregate. ×
- Thomas Muller scored 27th Champions League goal. ✓
- Thomas Muller passed Mario Gomez in goals. ×
- Thomas Muller is now the leading German scorer in the competition. ✓
- After the game Thomas Muller led the celebrations. ×
- Thomas Muller led the celebrations using a megaphone. ×

...



Indirect

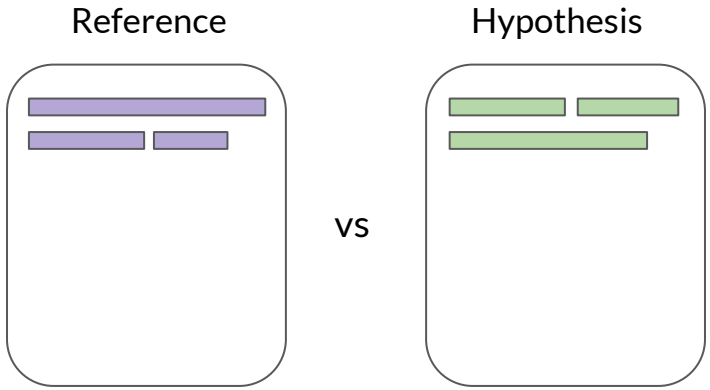
- Maynez et al., 2020;
- Bhandari et al., 2020
- ...

Bhandari et al., 2020

Human evaluation for Summarization

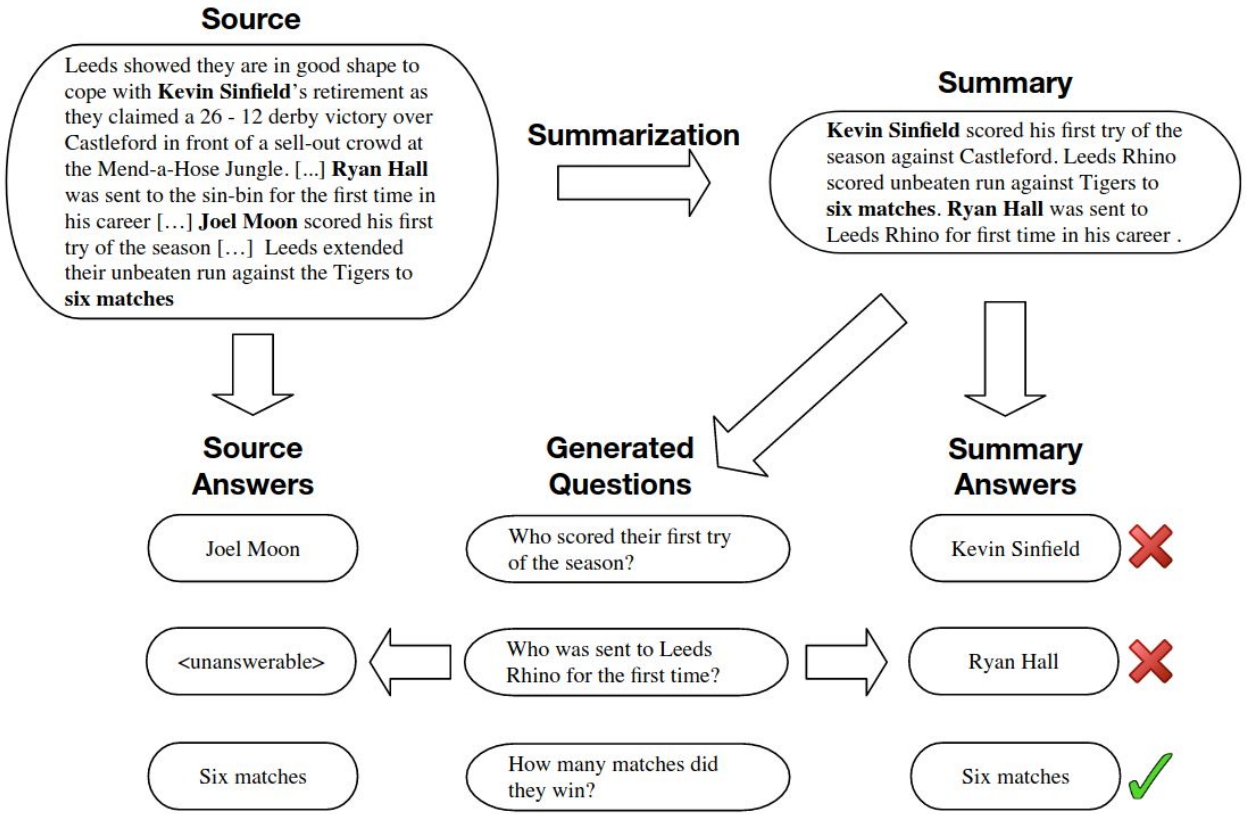
	Coherence	Consistency	Fluency	Relevance	SCU	Accuracy	Coverage	Focus	Overall
SummEval (Fabbri et al., 2021)	✓	✓	✓	✓					
REALSumm (Bhandari et al., 2020)					✓				
Human Feedback (Stiennon et al., 2020)	✓					✓	✓		✓
Multi_SummEval (Koto et al., 2021)							✓	✓	

Automatic metrics - MT vs Summarization



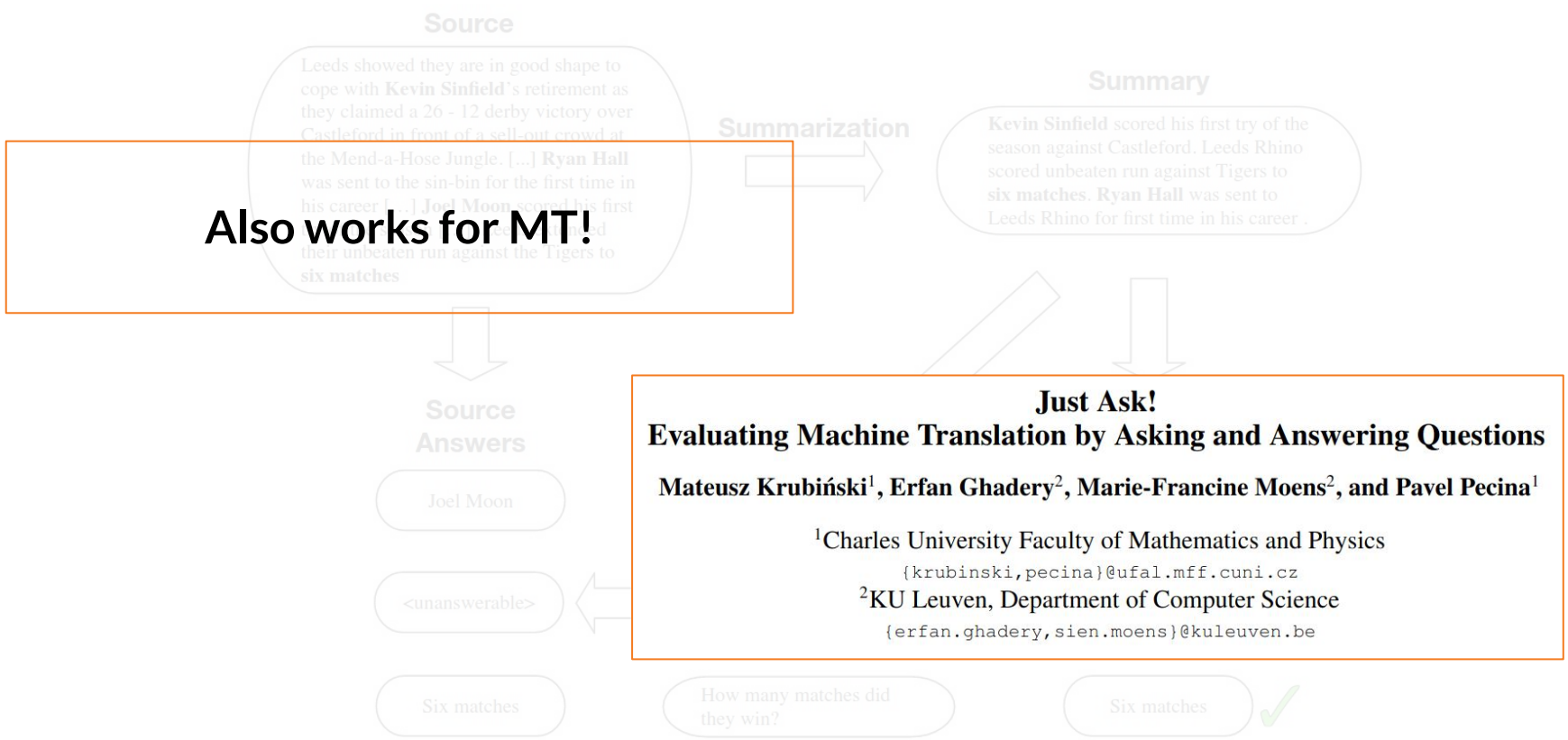
- Surface level metrics: BLEU (Papineni, 2002), ROUGE (Lin, 2004), TER (Snover, 2006) …
- Embedding similarity based metrics: MoverScore (Zhao, 2019), BERTScore (Zhang, 2020) …
- QA based metrics: QAGS (Wang, 2020), QAEval (Deutsch, 2021), MTEQA (Krubiński, 2021) …
- Trainable estimator metrics: BLEURT (Sellam, 2020), COMET (Rei, 2020), SummScore (Lin, 2022) …
- …

Question-answering based metrics for Summarization

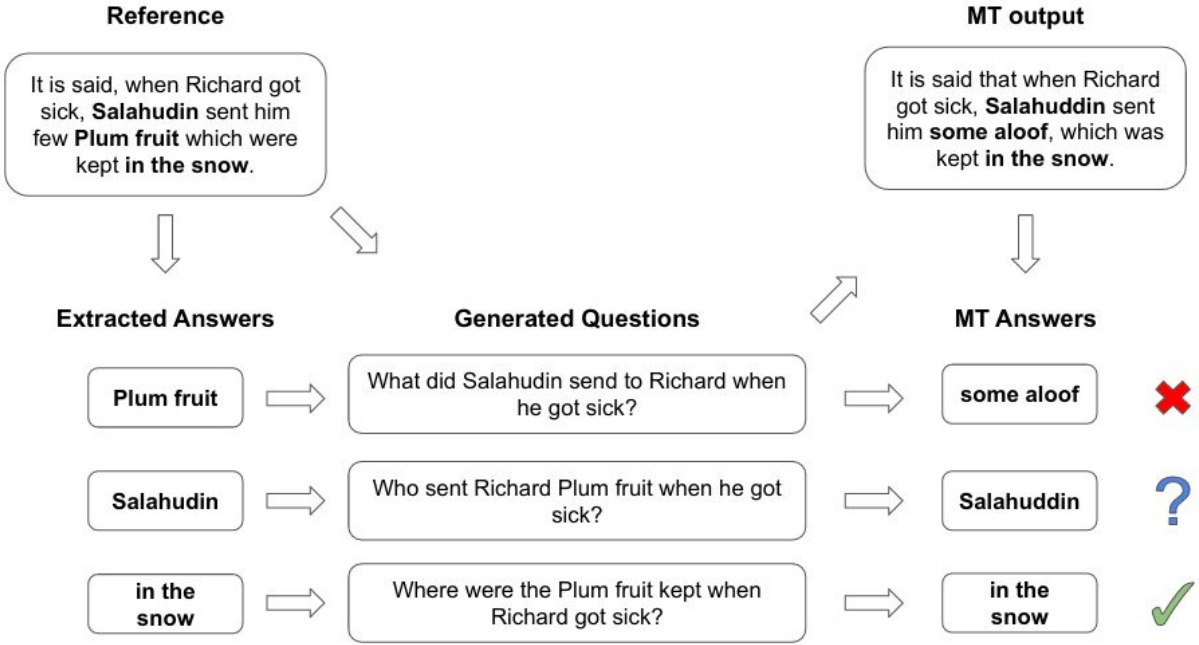


Wang et al., 2020

Question-answering based metric for MT - MTEQA



Question-answering based metric for MT - MTEQA



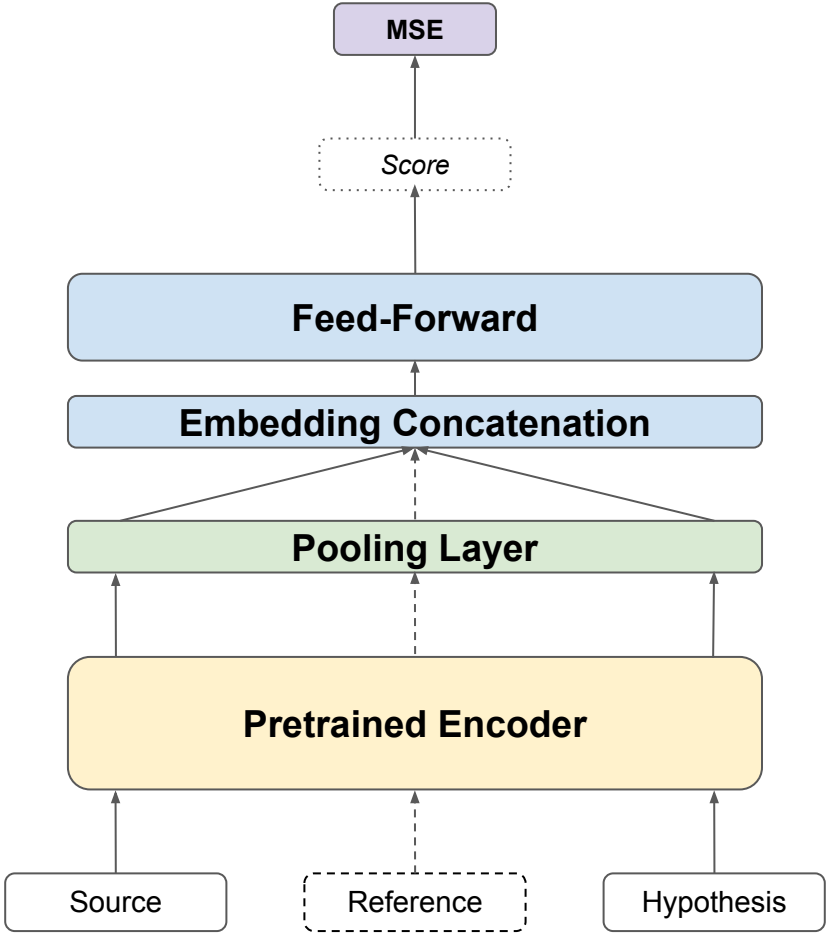
	ref-A	ref-B
MQM	5.52	0.42
MTEQA	0.47 (3)	0.74 (1)
TER	0.40 (9)	0.71 (2)
BERTScore	0.42 (6)	0.69 (3)
bleurt-20	0.45 (5)	0.68 (4)
cushLEPOR (LM)	0.39 (11)	0.68 (5)
Prism	0.46 (4)	0.68 (6)
COMET-MQM_2021	0.40 (8)	0.67 (7)
BLEU	0.30 (13)	0.65 (8)
YiSi-1	0.42 (7)	0.65 (9)
chrF	0.40 (10)	0.62 (10)
MEE2	0.36 (12)	0.60 (11)
C-SPECpn	0.49 (2)	0.54 (12)
tgt-regEMT	0.5 (1)	0.37 (13)
average	0.42	0.64

Table 13: Pairwise accuracy for ranking system pairs for TED Chinese→English using either ref-A (original ref) or ref-B (extra generated ref).

Freitag et al., 2021

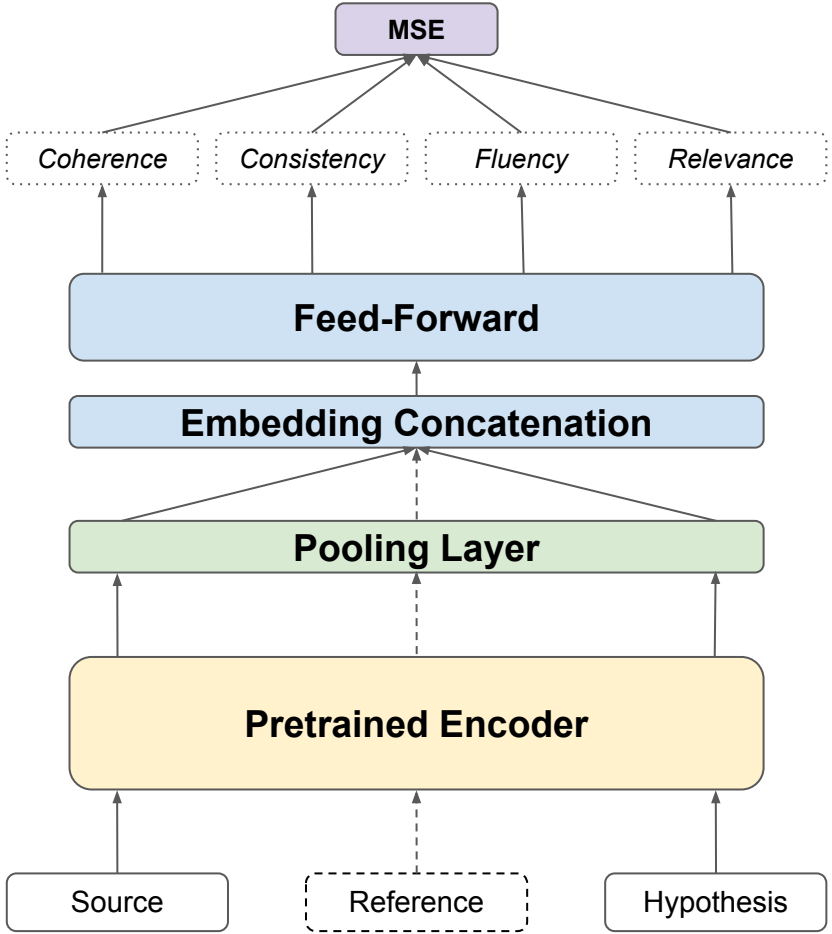
Methodology

COMET (Rei et al. 2020)



COMES (Krubiński and Pecina, 2022)

- Will COMET work for Summary evaluation?
- One score vs multiple scores?
- Reference-based vs QE?
- Fine-tune COMET vs “from scratch”?



Experiments

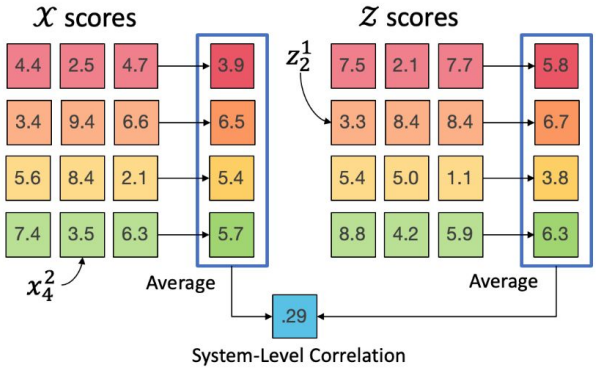
Dataset

- SummEval (Fabbri et al., 2021)
  - 100 articles from CNN/DailyMail corpus (Nallapati et al., 2016) - in English
  - Each summarized by 17 systems
  - 3 expert judgments for *Coherence*, *Consistency*, *Fluency* and *Relevance*
  - 11 references per article (original and 10 alternatives by Kryściński et al., 2020)
  - Evaluate using the System-level Kendall's Tau τ

Dataset

- SummEval (Fabbri et al., 2021)
  - 100 articles from CNN/DailyMail corpus (Nallapati et al., 2016) - in English
  - Each summarized by 17 systems
  - 3 expert judgments for *Coherence*, *Consistency*, *Fluency* and *Relevance*
  - 11 references per article (original and 10 alternatives by Kryściński et al., 2020)
  - Evaluate using the System-level Kendall's Tau τ

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} = 1 - \frac{2(\text{number of discordant pairs})}{\binom{n}{2}}.$$



Deutschet al., 2022

Dataset

- SummEval (Fabbri et al., 2021)
  - 100 articles from CNN/DailyMail corpus (Nallapati et al., 2016) - in English
  - Each summarized by 17 systems
  - 3 expert judgments for *Coherence*, *Consistency*, *Fluency* and *Relevance*
  - 11 references per article (original and 10 alternatives by Kryściński et al., 2020)
  - Evaluate using the System-level Kendall's Tau τ

Q: Largest resource available - how to use for both training and testing?

Dataset

- SummEval (Fabbri et al., 2021)
  - 100 articles from CNN/DailyMail corpus (Nallapati et al., 2016) - in English
  - Each summarized by 17 systems
  - 3 expert judgments for *Coherence*, *Consistency*, *Fluency* and *Relevance*
  - 11 references per article (original and 10 alternatives by Kryściński et al., 2020)
  - Evaluate using the System-level Kendall's Tau τ

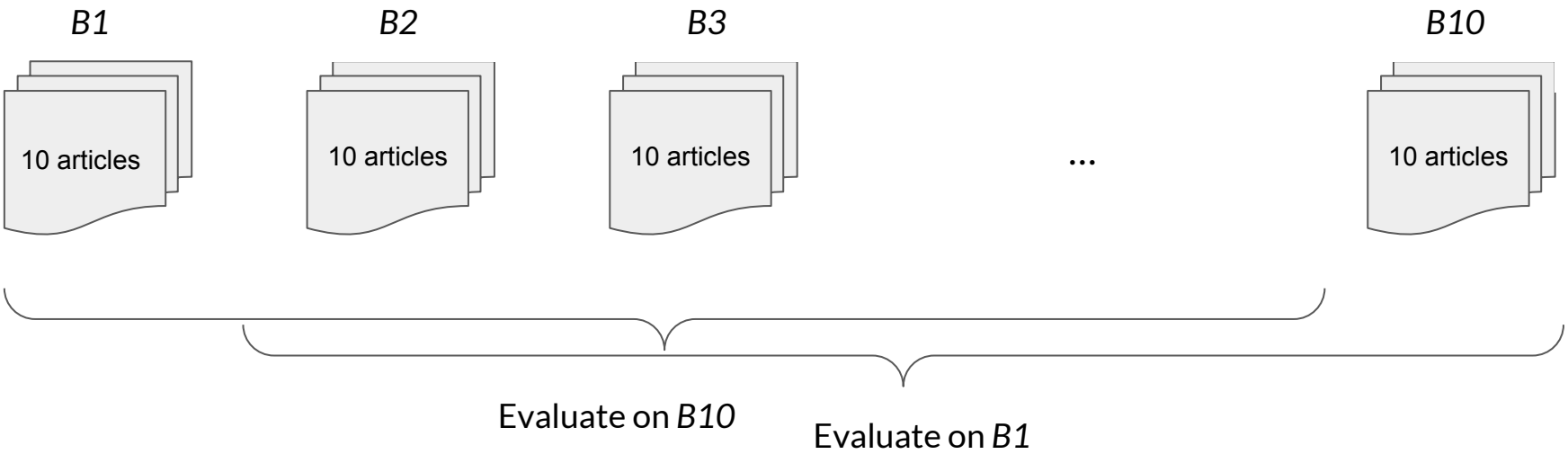
Q: Largest resource available - how to use for both training and testing?

A: Use cross-validation!

Dataset

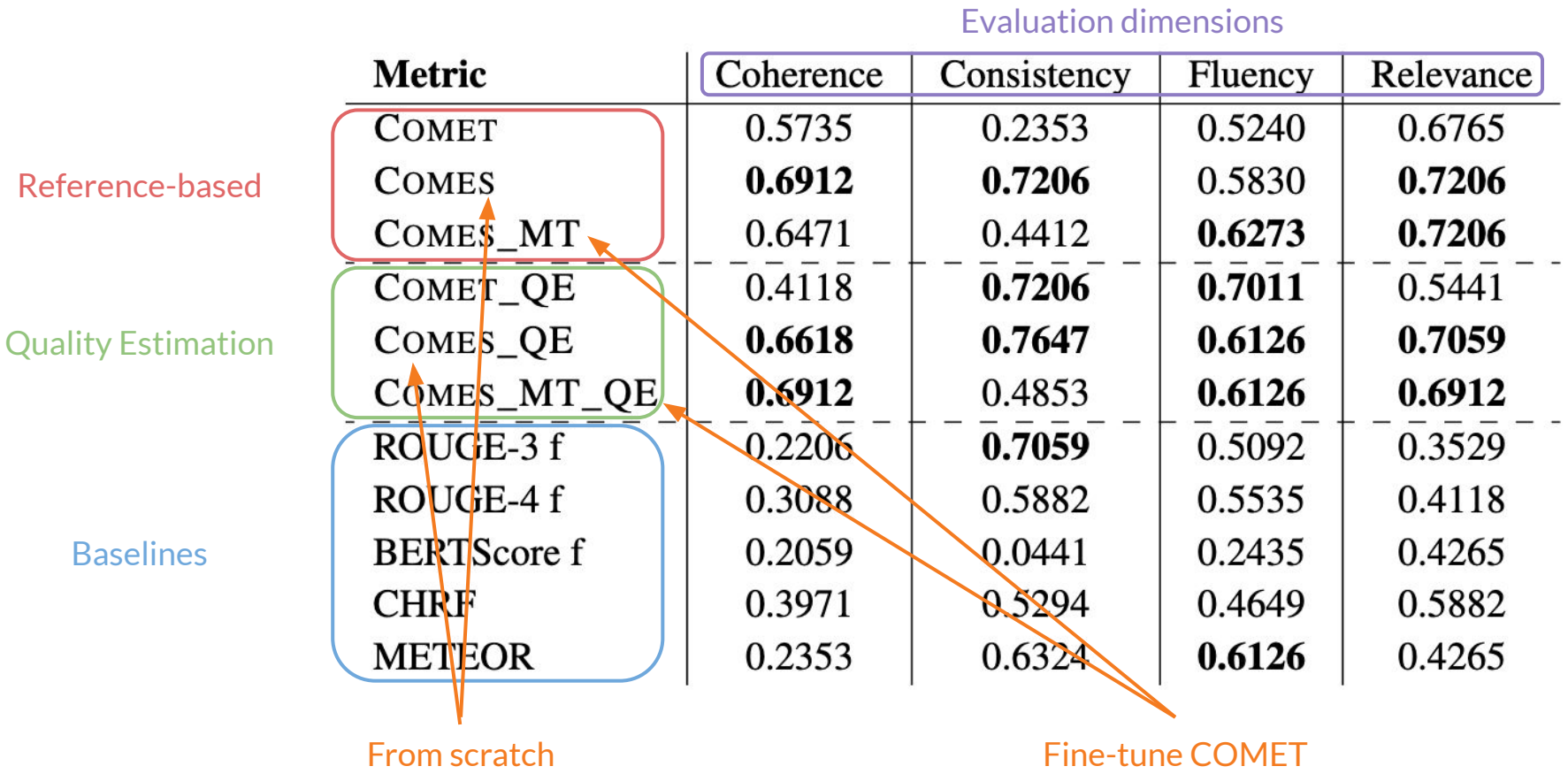
Q: Largest resource available - how to use for both training and testing?

A: Use cross-validation!



80 articles x *11* references x *17* models x *3* annotations = *44, 880* training instances

Kendall's Tau correlations - SummEval dataset



Metric	Coherence	Consistency	Fluency	Relevance
COMET	0.5735	0.2353	0.5240	0.6765
COMES	**0.6912**	**0.7206**	0.5830	**0.7206**
COMES_MT	0.6471	0.4412	**0.6273**	**0.7206**
COMET_QE	0.4118	**0.7206**	**0.7011**	0.5441
COMES_QE	**0.6618**	**0.7647**	**0.6126**	**0.7059**
COMES_MT_QE	**0.6912**	0.4853	**0.6126**	**0.6912**
ROUGE-3 f	0.2206	**0.7059**	0.5092	0.3529
ROUGE-4 f	0.3088	0.5882	0.5535	0.4118
BERTScore f	0.2059	0.0441	0.2435	0.4265
CHRF	0.3971	0.5294	0.4649	0.5882
METEOR	0.2353	0.6324	**0.6126**	0.4265

Kendall's Tau correlations - SummEval dataset

Metric	Coherence	Consistency	Fluency	Relevance
COMET	0.5735	0.2353	0.5240	0.6765
COMES	**0.6912**	**0.7206**	0.5830	**0.7206**
COMES_MT	0.6471	0.4412	**0.6273**	**0.7206**
COMET_QE	0.4118	**0.7206**	**0.7011**	0.5441
COMES_QE	**0.6618**	**0.7647**	**0.6126**	**0.7059**
COMES_MT_QE	**0.6912**	0.4853	**0.6126**	**0.6912**
ROUGE-3 f	0.2206	**0.7059**	0.5092	0.3529
ROUGE-4 f	0.3088	0.5882	0.5535	0.4118
BERTScore f	0.2059	0.0441	0.2435	0.4265
CHRF	0.3971	0.5294	0.4649	0.5882
METEOR	0.2353	0.6324	**0.6126**	0.4265

- **Coherence** - the collective quality of all sentences.
- **Consistency** - the factual alignment between the summary and source
- **Fluency** - the quality of individual sentences
- **Relevance** - selection of important content from the source.

- COMET high correlation with Coherence
- COMET_QE much higher with Consistency

Kendall's Tau correlations - SummEval dataset

Metric	Coherence	Consistency	Fluency	Relevance
COMET	0.5735	0.2353	0.5240	0.6765
COMES	**0.6912**	**0.7206**	0.5830	**0.7206**
COMES_MT	0.6471	0.4412	**0.6273**	**0.7206**
COMET_QE	0.4118	**0.7206**	**0.7011**	0.5441
COMES_QE	**0.6618**	**0.7647**	**0.6126**	**0.7059**
COMES_MT_QE	**0.6912**	0.4853	**0.6126**	**0.6912**
ROUGE-3 f	0.2206	**0.7059**	0.5092	0.3529
ROUGE-4 f	0.3088	0.5882	0.5535	0.4118
BERTScore f	0.2059	0.0441	0.2435	0.4265
CHRF	0.3971	0.5294	0.4649	0.5882
METEOR	0.2353	0.6324	**0.6126**	0.4265

- COMES > COMET?
- Fine-tuning COMET vs training from scratch -> mixed results, no clear improvement

Kendall's Tau correlations - Stiennon et al., 2020

Metric		Overall	Accuracy	Coverage	Coherence
ROUGE-1 f		0.647	**0.752**	0.621	0.464
ROUGE-2 f		0.569	0.699	0.542	0.438
ROUGE-L f		0.595	0.699	0.569	0.412
BERTScore f		0.621	**0.725**	0.595	0.464
COMET		**0.843**	0.686	**0.817**	0.425
COMES	Coherence	−0.204 ± 0.05	−0.050 ± 0.04	−0.230 ± 0.05	0.264 ± 0.04
	Consistency	0.722 ± 0.12	0.630 ± 0.06	0.695 ± 0.12	0.565 ± 0.07
	Fluency	0.209 ± 0.10	0.340 ± 0.07	0.186 ± 0.09	0.625 ± 0.07
	Relevance	**0.774 ± 0.03**	**0.703 ± 0.04**	**0.750 ± 0.03**	0.627 ± 0.02
COMES_MT	Coherence	0.366 ± 0.16	0.403 ± 0.12	0.340 ± 0.16	**0.654 ± 0.07**
	Consistency	0.455 ± 0.11	0.418 ± 0.10	0.431 ± 0.12	0.604 ± 0.11
	Fluency	0.433 ± 0.12	0.414 ± 0.11	0.407 ± 0.12	0.634 ± 0.06
	Relevance	0.379 ± 0.16	0.403 ± 0.12	0.353 ± 0.16	**0.654 ± 0.06**
COMET_QE		**0.922**	0.660	**0.895**	0.477
COMES_QE	Coherence	−0.158 ± 0.1	−0.017 ± 0.09	−0.184 ± 0.10	0.305 ± 0.09
	Consistency	0.714 ± 0.05	0.630 ± 0.05	0.688 ± 0.05	0.544 ± 0.06
	Fluency	0.170 ± 0.13	0.272 ± 0.11	0.144 ± 0.13	0.559 ± 0.08
	Relevance	0.695 ± 0.07	0.648 ± 0.06	0.669 ± 0.07	0.646 ± 0.04
COMES_MT_QE	Coherence	0.480 ± 0.11	0.467 ± 0.09	0.454 ± 0.11	**0.668 ± 0.03**
	Consistency	0.528 ± 0.07	0.484 ± 0.08	0.502 ± 0.07	0.638 ± 0.06
	Fluency	0.519 ± 0.07	0.480 ± 0.08	0.493 ± 0.07	0.647 ± 0.05
	Relevance	0.493 ± 0.09	0.477 ± 0.08	0.467 ± 0.09	**0.678 ± 0.02**

Evaluation dimensions

COMES* heads

- Cross-validation for COMES* -> confidence estimation
- COMET correlates great with "Overall" dimension
- Reference-less variant better

Kendall's Tau correlations - Stiennon et al., 2020

Metric		Overall	Accuracy	Coverage	Coherence
ROUGE-1 f		0.647	**0.752**	0.621	0.464
ROUGE-2 f		0.569	0.699	0.542	0.438
ROUGE-L f		0.595	0.699	0.569	0.412
BERTScore f		0.621	**0.725**	0.595	0.464
COMET		**0.843**	0.686	**0.817**	0.425
COMES	Coherence	−0.204 ± 0.05	−0.050 ± 0.04	−0.230 ± 0.05	0.264 ± 0.04
	Consistency	0.722 ± 0.12	0.630 ± 0.06	0.695 ± 0.12	0.565 ± 0.07
	Fluency	0.209 ± 0.10	0.340 ± 0.07	0.186 ± 0.09	0.625 ± 0.07
	Relevance	**0.774 ± 0.03**	**0.703 ± 0.04**	**0.750 ± 0.03**	0.627 ± 0.02
COMES_MT	Coherence	0.366 ± 0.16	0.403 ± 0.12	0.340 ± 0.16	**0.654 ± 0.07**
	Consistency	0.455 ± 0.11	0.418 ± 0.10	0.431 ± 0.12	0.604 ± 0.11
	Fluency	0.433 ± 0.12	0.414 ± 0.11	0.407 ± 0.12	0.634 ± 0.06
	Relevance	0.379 ± 0.16	0.403 ± 0.12	0.353 ± 0.16	**0.654 ± 0.06**
COMET_QE		**0.922**	0.660	**0.895**	0.477
COMES_QE	Coherence	−0.158 ± 0.1	−0.017 ± 0.09	−0.184 ± 0.10	0.305 ± 0.09
	Consistency	0.714 ± 0.05	0.630 ± 0.05	0.688 ± 0.05	0.544 ± 0.06
	Fluency	0.170 ± 0.13	0.272 ± 0.11	0.144 ± 0.13	0.559 ± 0.08
	Relevance	0.695 ± 0.07	0.648 ± 0.06	0.669 ± 0.07	0.646 ± 0.04
COMES_MT_QE	Coherence	0.480 ± 0.11	0.467 ± 0.09	0.454 ± 0.11	**0.668 ± 0.03**
	Consistency	0.528 ± 0.07	0.484 ± 0.08	0.502 ± 0.07	0.638 ± 0.06
	Fluency	0.519 ± 0.07	0.480 ± 0.08	0.493 ± 0.07	0.647 ± 0.05
	Relevance	0.493 ± 0.09	0.477 ± 0.08	0.467 ± 0.09	**0.678 ± 0.02**

Evaluation dimensions

COMES* heads

- Coherence head of COMES has the lowest correlation with the Coherence dimension in data
- … but the highest if we use the variant pre-trained on MT!

Segment-level Pearson correlations - Koto et al., 2021

Metric		Focus					Coverage				
		de	es	tr	fr	ru	de	es	tr	fr	ru
COMET		**0.82**	**0.51**	**0.64**	**0.47**	**0.42**	**0.82**	**0.54**	**0.72**	**0.40**	**0.45**
COMET_QE		0.29	0.06	0.03	0.01	0.10	0.31	0.09	0.27	−0.03	**0.24**
COMES	Coherence	0.21	0.03	0.07	**0.16**	−0.01	0.15	−0.01	−0.05	0.08	−0.07
	Consistency	0.33	0.11	**0.21**	0.10	**0.14**	0.35	0.13	0.30	0.07	0.22
	Fluency	0.36	0.05	0.10	0.11	0.08	0.33	0.06	0.10	0.05	0.15
	Relevance	**0.42**	**0.15**	**0.25**	**0.18**	**0.12**	**0.44**	**0.20**	**0.38**	**0.15**	**0.26**
COMES_MT	Coherence	**0.37**	0.13	**0.25**	0.15	0.08	0.36	0.09	**0.31**	0.11	0.14
	Consistency	0.31	0.10	0.20	0.14	0.09	0.30	0.09	0.24	0.09	0.16
	Fluency	0.31	0.10	**0.21**	0.14	0.09	0.30	0.09	0.25	0.09	0.16
	Relevance	0.36	0.12	**0.25**	0.15	0.09	0.35	0.09	0.30	0.10	0.15
COMES_MT_ML	Coherence	0.03	−0.01	−0.03	0.13	−0.09	−0.04	−0.04	−0.17	0.10	−0.14
	Consistency	0.10	0.02	0.01	0.00	0.01	0.10	0.00	0.01	−0.02	0.12
	Fluency	0.23	0.02	0.09	0.07	0.01	0.22	0.03	0.08	−0.01	0.01
	Relevance	0.36	**0.20**	0.16	0.15	0.06	**0.38**	**0.25**	0.27	**0.16**	0.23

- Beyond English -> QE variant of COMES significantly worse
- Lagging behind traditional metrics, i.e. BERTScore
- … but the dataset is small - 135 documents, 2 systems

Segment-level Pearson correlations - Koto et al., 2021

		Focus					Coverage				
Metric		de	es	tr	fr	ru	de	es	tr	fr	ru
COMET		**0.82**	**0.51**	**0.64**	**0.47**	**0.42**	**0.82**	**0.54**	**0.72**	**0.40**	**0.45**
COMET_QE		0.29	0.06	0.03	0.01	0.10	0.31	0.09	0.27	−0.03	**0.24**
COMES	Coherence	0.21	0.03	0.07	**0.16**	−0.01	0.15	−0.01	−0.05	0.08	−0.07
	Consistency	0.33	0.11	**0.21**	0.10	**0.14**	0.35	0.13	0.30	0.07	0.22
	Fluency	0.36	0.05	0.10	0.11	0.08	0.33	0.06	0.10	0.05	0.15
	Relevance	**0.42**	**0.15**	**0.25**	**0.18**	**0.12**	**0.44**	**0.20**	**0.38**	**0.15**	**0.26**
COMES_MT	Coherence	**0.37**	0.13	**0.25**	0.15	0.08	0.36	0.09	**0.31**	0.11	0.14
	Consistency	0.31	0.10	0.20	0.14	0.09	0.30	0.09	0.24	0.09	0.16
	Fluency	0.31	0.10	**0.21**	0.14	0.09	0.30	0.09	0.25	0.09	0.16
	Relevance	0.36	0.12	**0.25**	0.15	0.09	0.35	0.09	0.30	0.10	0.15
COMES_MT_ML	Coherence	0.03	−0.01	−0.03	0.13	−0.09	−0.04	−0.04	−0.17	0.10	−0.14
	Consistency	0.10	0.02	0.01	0.00	0.01	0.10	0.00	0.01	−0.02	0.12
	Fluency	0.23	0.02	0.09	0.07	0.01	0.22	0.03	0.08	−0.01	0.01
	Relevance	0.36	**0.20**	0.16	0.15	0.06	**0.38**	**0.25**	0.27	**0.16**	0.23

COMES trained on machine-translated SummEval

- (multilingual) COMES trained on automatically translated SummEval performs worse than the one trained on SummEval (English only)
- Confirms the finding of Braun et al., (2022) that **summary evaluations do not survive translation**
- … but still behind COMET

Conclusions

Conclusions

- COMET metric trained on (multilingual) annotated MT outputs can be successfully used to evaluate (monolingual) Summarization outputs
  - Applicable in cross-lingual settings
  - Applicable is scenarios when the reference is not available
- Training on labels from one dataset (SummEval) and applying to other datasets is inconsistent
  - Different annotation methods, different annotation dimensions
  - SummEval - correlation between expert and crowd-sourced judges close to 0
- Use COMET_QE when evaluating summarization systems

Thank you!

krubinski@ufal.mff.cuni.cz