# 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
{krubinski,pecina}@ufal.mff.cuni.cz

## Abstract

COMET is a recently proposed trainable neural-based evaluation metric developed to assess the quality of Machine Translation systems. In this paper, we explore the usage of COMET for evaluating Text Summarization systems – despite being trained on multilingual MT outputs, it performs remarkably well in monolingual settings, when predicting summarization output quality. We introduce a variant of the model – COMES – trained on the annotated summarization outputs that uses MT data for pre-training. We examine its performance on several datasets with human judgments collected for different notions of summary quality, covering several domains and languages.

## 1 Introduction

Since manual annotation for any generative task is costly and time consuming, automatic metrics are commonly used to both measure the progress during training and compare outputs from independent systems. Thanks to the Metrics Shared Task (Freitag et al., 2021b; Mathur et al., 2020; Ma et al., 2019) collocated with the WMT workshop since 2008 (Callison-Burch et al., 2008), advances in the MT models performance are accompanied by a continuous development of new automatic metrics (Lo, 2019; Kepler et al., 2019; Rei et al., 2020; Sellam et al., 2020) that improve correlation with human judgment and are robust to both domain shifts and changes in annotation style (Freitag et al., 2021a).

In contrary, for the task of text summarization remarkable advances in modeling techniques (Koto et al., 2022) are not followed by corresponding research on evaluation methods – a number of recent studies (Lewis et al., 2020a; Li et al., 2020; Raffel et al., 2020) keep relying mostly on ROUGE (Lin, 2004), a string-overlap metric measuring the n-gram correspondence with the reference summary.

One of the issues making research on summary evaluation metrics difficult is lack of standardized framework for collecting human judgments. They are collected not only along several dimensions (Table 1) but also using different methods – based on Likert scale (Fabbri et al., 2021; Stiennon et al., 2020), Direct Assessment (Koto et al., 2021) or methods that output numerical score indirectly (Maynez et al., 2020; Bhandari et al., 2020) by e.g. counting number of spans highlighted in the model output by annotators. The other issue is the amount of available annotated data. Even the largest datasets (Fabbri et al., 2021; Bhandari et al., 2020; Maynez et al., 2020) have no more than tens of thousands of annotated instances. This is by far less than the amount of available data for machine translation, with roughly 800k ⟨⟨source, hypothesis, reference⟩⟩ annotated triplets available from the evaluation campaigns of the previous editions of WMT News Translation shared task[1].

The question we ask is: *Can we use this resource to improve summary evaluation?* While the tasks of Machine Translation and Text Summarization are different, we believe that the problem of evaluating the quality of generated output is closely related.

To address this question, we examine the applicability of the COMET metric by Rei et al. (2020) (Section 2.2) that is trained on the annotated MT data and capable of directly regressing a quality score. We propose (Section 3) a variant of the model – COMES[2] – that uses the annotated MT data for pre-training and is capable of predicting several aspects of summary quality. We evaluate our approach (Section 4) on selected datasets with various annotation styles.

## 2 Related Work

### 2.1 Automatic Summary Evaluation

Historically, the quality of summary was measured by comparing n-gram overlap between reference

[1] https://wmt-metrics-task.github.io/

[2] **C**rosslingual **O**ptimized **M**etric for **E**valuation of **S**ummarization

*Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 21 - 31
November 20, 2022 ©2022 Association for Computational Linguistics

| | 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) | | | | | | | ✓ | ✓ | |

Table 1: Comparison of the types of annotations in the summary evaluation datasets used in our experiments. For a comprehensive survey on the summary evaluation resources see Koto et al. (2022).

and system output (Papineni et al., 2002; Lin, 2004). Over the years, a variety of metrics were proposed for this task – based on question answering (Eyal et al., 2019; Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020), similarity between summary and reference embeddings (Zhao et al., 2019; Zhang et al., 2020) or the usefulness of summary for language modeling on the source document (Colombo et al., 2022; Liu et al., 2022).

### 2.2 COMET

COMET is a trained metric that, based on semantic similarities between the translated and reference texts, learns to output a score that resembles the human perception of translation quality. In the default settings, input to the model is a ⟨⟨source, hypothesis, reference⟩⟩ triple, but a reference-less variant for Quality Estimation (COMET_QE) that operates on ⟨⟨source, hypothesis⟩⟩ pairs was also proposed.

On a high level, COMET uses a pre-trained multilingual language model to independently extract representations for each of the input sequences, which are then pooled and concatenated, before being processed with a stack of feed-forward layers that outputs a single numerical value. The choice of COMET for our experiments (as opposed to e.g. BLEURT (Sellam et al., 2020) or YiSi (Lo and Larkin, 2020)) is motivated by a recent metrics study by Kocmi et al. (2021) that shows it's superior performance compared to other (pretrained) metrics and the availability of a well-documented implementation[3].

### 2.3 SummEval

SummEval[4] (Fabbri et al., 2021) is a recently proposed dataset with human annotations for several dimensions of summary quality. It consists of 100 articles randomly sampled from the test split of the CNN/DailyMail corpus (Nallapati et al., 2016), each of them summarized by 17 systems. For each system output, the authors collected 3 expert judgments for *Coherence*, *Consistency*, *Fluency* and *Relevance* on a Likert scale of 1 to 5. In addition to the original reference, for each article, 10 alternative references were created by Kryscinski et al. (2020).

## 3 COMES

In the context of Machine Translation two frameworks for collecting human ratings were employed recently – MQM (Lommel et al., 2014) and DA (Bojar et al., 2017), both producing a single numerical score that indicated the overall translation quality. That is not the case for Text Summarization – content, fluency and clarity are all graded independently (Hardy et al., 2019; Koto et al., 2022). As a result, the COMET metric trained on MT data outputs a single overall score.

In our experiments, when reporting COMET performance, we compare this single overall score to all evaluation dimensions. To enable (independently) predicting several aspects of summary quality at once, we propose a modification that alters the number of outputs in the last feed-forward layer, see Figure 1. We experiment with both training from scratch (COMES) and pre-training on the annotated MT data by initializing the model weights from the COMET checkpoint (COMES_MT). See Appendix A.1 for the training details. In both scenarios, we examine the reference-less variant of the metric (COMES_QE and COMES_QE_MT, respectively).

## 4 Experiments

### 4.1 SummEval experiments

Since, to the best of our knowledge, SummEval is the largest resource for summary evaluation, we

[3] https://github.com/Unbabel/COMET/

[4] https://github.com/Yale-LILY/SummEval

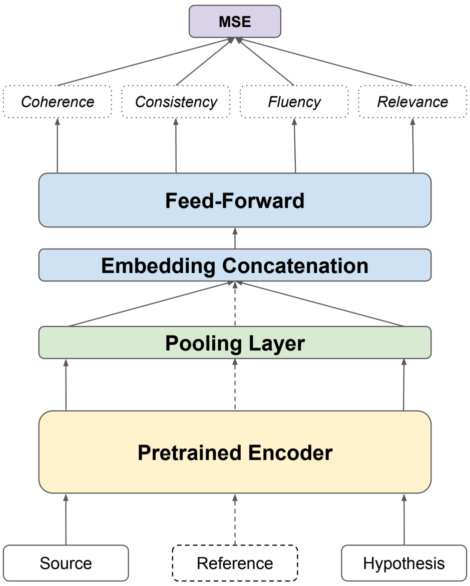

Figure 1: Estimator model architecture used in COMES. Source, reference and hypothesis are all independently encoded with a pre-trained encoder. Pooling layer is used to create sentence embeddings from sequences of token embeddings. In the COMES variant, the last feed-forward layer has 4 outputs, corresponding to different summary evaluation dimensions. Dashed lines are used to indicate the reference-less variant. For the full COMET description see Rei et al. (2020).

would like to use it both for training and evaluation. To achieve this, we rely on cross-validation. We split the data into 10 subsets of 10 articles each, using 80 articles for training, 10 for validation (early stopping) and evaluating on the remaining 10. We train 10 models, use each of them to score 10% of the available (unseen) data and merge the results. That way we can directly compare to other metrics that report correlation on the whole SummEval dataset. During training, we use each reference and each expert annotation[5] to create more training instances (80 articles × 11 references × 17 models × 3 annotations $=$ $44,880$ instances). During evaluation, we handle multiple references by scoring each reference independently and taking the maximum score.

The results of our experiments can be found in Table 2. We report the system-level Kendall's Tau correlations with (average) expert annotations. For comparison, we also include metrics which previously (Fabbri et al., 2021) achieved the highest correlation with each of the evaluation dimensions – ROUGE-1 and ROUGE-4, BERTScore (Zhang et al., 2020), CHRF (Popović, 2015) and METEOR (Lavie and Agarwal, 2007). Scoring system outputs with both out-of-the-box variants (COMET and COMET_QE) results in the highest correlation coefficients along all metrics analysed by Fabbri et al. (2021) for *Coherence* and *Relevance* dimensions. The reference-less variant has much higher correlation with the *Consistency* dimension (0.24 → 0.72). Both COMES and COMES_QE variants perform similarly, achieving higher correlations than both COMET (COMET_QE) and traditional metrics. However the effect of pre-training is ambiguous – on average it does not help, but the main cause is the poor performance on predicting the *Consistency* dimension.

### 4.2 Domain and Annotation Style shift

To get a better understanding of the metric performance, we apply it to several other annotated summarization datasets. Since we have trained 10 instances for each variant of the COMES models (Section 4.1), evaluating with each of them allows us to estimate the confidence intervals directly, not having to rely on e.g. bootstrapping (Deutsch et al., 2021).

To examine the performance on non-matching evaluation dimensions, we report results on data[6] from the same domain – subset of the CNN/DailyMail corpus. Bhandari et al. (2020) produced the numerical gold-standard scores by rating

[5] We have tried averaging human ratings during training, the results were comparable but slightly worse.

[6] https://github.com/neulab/REALSumm

| Metric | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|
| 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 |
| 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** |

Table 2: System-level Kendall's Tau correlations with (average) expert annotations for four evaluation dimensions annotated in the SummEval dataset. The three metrics with the highest correlation in each column are bolded. See Table 2 in Fabbri et al. (2021) for results of other metrics.

a system output based on a number of Semantic Content Units (SCUs) that can be inferred from it. LitePyramid (Shapira et al., 2019) method was used to obtain SCUs from reference summaries. On this dataset, the reference-less COMET_QE outperforms any other variant, almost doubling the correlation of COMET (0.46 → 0.75). The *Consistency* head of COMES_QE comes in second (0.59). Considering the recall based nature of annotations, it is not surprising that the best correlation is obtained by the recall variant of ROUGE (0.85).

In an independent work[7], Stiennon et al. (2020) annotated a different subset of the CNN/DailyMail corpus by rating system outputs for *Accuracy*, *Coherence*, *Coverage* and *Overall Quality*. Again, the reference-less variant COMET_QE performs best, obtaining almost a perfect correlation with the *Overall* dimension (0.92). This is by far a better result than any traditional metric considered (0.65 by ROUGE-1 F-score). COMES trained from scratch out-performs the pre-trained variant COMES_MT which may indicate overfitting to the SummEval annotations. Surprisingly, the highest correlation with the *Coherence* dimension (present in the SummEval annotations used for training) is not obtained by the *Coherence* head of COMES. That is however the case for the variant pre-trained on MT data (COMES_MT). For the full, results see Table 5 and Table 6 in Appendix.

To validate the performance on a different domain, we evaluate on the subset of the TL;DR corpus (Völske et al., 2017) annotated in a similar manner by Stiennon et al. (2020), see Table 7 in Appendix. On this dataset COMET achieves the top correlation, with the COMES clearly lagging behind in performance compared to the pre-trained COMES_MT variant.

### 4.3 Non-English data

One of the strengths of the COMET metric is its multilinguality – the model has seen over 30 language pairs during training. To assess its quality as a summary evaluation tool for non-English data, we evaluated it on the Multi_SummEval dataset (Koto et al., 2021). With only two system outputs annotated (along the *Focus* and *Coverage* dimensions), the size of the resource is not sufficient for reporting system-level correlations. Thus, we report the summary-level (segment-level) Pearson correlations.

For a fair comparison, we wanted to train the COMES model variant using the multilingual data. Due to the lack of sufficient resources, we fall back on using automatic machine translation to translate the English annotated data. This approach has proven successful for e.g. Question Answering (Lewis et al., 2020b; Macková and Straka, 2020). We limit our analysis to the subset of languages from Multi_SummEval that originates from the MLSUM (Scialom et al., 2020) corpus. We have translated SummEval into German, French, Russian, Turkish and Spanish using the uni-directional models provided by the Helsinki-NLP group (Tiedemann, 2020) and used the data (together with the original SummEval) to train a multilingual COMES model (COMES_MT_ML).

Our findings indicate that in the summary-level evaluation, the original COMET metric is superior to any other variant considered, clearly outperforming the reference-less variant COMET_QE.

[7] https://github.com/openai/summarize-from-feedback

| **Metric** | *CV* | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|---|
| COMES | ✓ | 0.6912 | 0.7206 | 0.5830 | 0.7206 |
| COMES | - | 0.9412 | 0.9412 | 0.8340 | 0.9265 |
| COMES_MT | ✓ | 0.6471 | 0.4412 | 0.6273 | 0.7206 |
| COMES_MT | - | 0.8088 | 0.7941 | 0.6864 | 0.8676 |
| COMES_QE | ✓ | 0.6618 | 0.7647 | 0.6126 | 0.7059 |
| COMES_QE | - | 0.9706 | 0.9265 | 0.8782 | 0.9706 |
| COMES_MT_QE | ✓ | 0.6912 | 0.4853 | 0.6126 | 0.6912 |
| COMES_MT_QE | - | 0.8235 | 0.7794 | 0.6568 | 0.8676 |

Table 3: System-level Kendall's Tau correlations with (average) expert annotations for four evaluation dimensions annotated in the SummEval dataset. The *CV* variants correspond to the un-biased cross-validation settings (Section 4.1), the remaining ones are obtained with the over-fitted models, see Section 4.4.

Surprisingly, both the COMES_MT and the COMES variants perform better than the multilingual COMES_MT_ML variant. This is in line with recent findings by Braun et al. (2022), which indicate that summary evaluations do not survive translation. On this dataset, even the best performing COMET is still inferior to both ROUGE and BERTScore. Considering, however, the relatively small size of the dataset (270 instances per language, outputs from two systems) we believe that the question about COMET/COMES usefulness for multilingual and summary-level evaluation is still open. For the full results, see Table 8 in Appendix.

### 4.4 Ablation Study

In Section 4.1, we propose the usage of cross-validation to enable training and un-biased testing on the SummEval dataset – different articles are used for training, validation and testing. To show that the model can over-fit to the data, we have trained a model using all of the available annotations from the SummEval dataset and then applied it to the same articles, already seen during training. Table 3 (rows without the *CV* mark) presents the results. It is clear that the model is able to memorize the annotations proving that the cross-validation approach enables un-biased reporting on the whole SummEval dataset and thus is a fair way of comparing COMES to other metrics.

In Section 2.2 we mention that COMET (and COMES) uses a pre-trained multilingual language model to extract representations from input sequences. In our experiments, it is always the XLM-RoBERTa (Conneau et al., 2020) model. A major difference between Machine Translation and Text Summarization is the length of the typical input. By examining the lengths of the tokenized documents from SummEval, we have realized that only 48% of them fit completely within the model limit of 512 tokens. However, on average, 92% of input tokens are consumed (average input document length in tokens equals 502) so the information lost is hopefully not significant. We leave the detailed analysis for future works.

## 5 Conclusion

In this paper, we showed that the COMET metric trained on (multilingual) MT outputs can be successfully used to evaluate the quality of (monolingual) summaries. We proposed an adaptation that enables scoring several (independent) evaluation dimensions at once. Our results (Table 2) indicate, that the off-the-shelf COMET metric performs comparable to the variants fine-tuned on the annotated summarization outputs. Furthermore, the reference-less variants perform similar to the ones using references, making the metric applicable in settings when the gold-standard summary is not available.

## Acknowledgements

This work was supported by the European Commission via its H2020 Program (contract no. 870930), the Czech Science Foundation (grant no. 19-26934X), and CELSA (project no. 19/018). In this work, we used data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2018101).

## References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical*

*Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Spencer Braun, Oleg Vasilyev, Neslihan Iskender, and John Bohannon. 2022. Does summary evaluation survive translation to other languages? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2425–2435, Seattle, United States. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Pierre Jean A. Colombo, Chloé Clavel, and Pablo Piantanida. 2022. Infolm: A new metric to evaluate summarization & data2text generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10554–10562. AAAI Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Ffci: A framework for interpretable automatic evaluation of summarization. *J. Artif. Int. Res.*, 73.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yu Lu Liu, Rachel Bawden, Thomas Scaliom, Benoît Sagot, and Jackie Chi Kit Cheung. 2022. Maskeval: Weighted mlm-based evaluation for text summarization and simplification. *CoRR*, abs/2205.12394.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 903–910, Online. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Kateřina Macková and Milan Straka. 2020. Reading comprehension in czech via machine translation and cross-lingual transfer. In *23rd International Conference on Text, Speech and Dialogue*, pages 171–179, Cham, Switzerland. Springer.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

# A Appendix

## A.1 COMES Hyper-Parameters

During COMES training, we mostly follow the training/fine-tuning configuration of Rei et al. (2021), see Table 4. We monitor Pearson correlation on the validation set for early stopping. When fine-tuning the COMET model instead of training from scratch, we decrease the `learning_rate` to 1.0e-05 and load weights from the *wmt21-comet-da* checkpoint. In the reference-less variant, we set the `hidden_sizes` to [2048, 1024] and load weights from the *wmt21-comet-qe-da* checkpoint. We employ gradient accumulation to train with the effective batch size of 40. As a part of pre-processing, we de-tokenize and true-case system outputs with Stanford CoreNLP (Manning et al., 2014) tool.

| | |
|---|---|
| nr_frozen_epochs | 1.0 |
| keep_embeddings_frozen | True |
| optimizer | AdamW |
| encoder_learning_rate | 1.0e-0 |
| learning_rate | 3.1e-05 |
| layerwise_decay | 0.95 |
| encoder | XLM-RoBERTa |
| pretrained_model | xlm-roberta-large |
| pool | avg |
| layer | mix |
| dropout | 0.15 |
| hidden_sizes | [3072, 1024] |
| epochs | 5 |

Table 4: Hyper-parameters used for COMES training.

## A.2 REALSumm results

In Table 5, we report the system-level Kendall's Tau correlations on the REALSumm corpus (100 articles × 25 models), annotated by Bhandari et al. (2020). „Score" column is used for metrics that output a single score, the following ones correspond to outputs from each of the COMES heads. From the analysis, we excluded 2 articles that appear in the SummEval dataset. For the COMES variants that we trained ourselves, we evaluate with models trained on each cross-validation fold, reporting mean and standard deviation, see Section 4.1 for details.

| Metric | LitePyramid SCU | | | | |
|---|---|---|---|---|---|
| | Score | Coherence | Consistency | Fluency | Relevance |
| ROUGE-1 r | **0.779** | | | | |
| ROUGE-2 r | **0.853** | | | | |
| ROUGE-L r | **0.746** | | | | |
| BERTScore r | 0.538 | | | | |
| JS-2 | 0.518 | | | | |
| MoverScore | 0.264 | | | | |
| COMET | 0.457 | | | | |
| COMES | | 0.242 ± 0.05 | 0.561 ± 0.07 | 0.290 ± 0.02 | 0.481 ± 0.05 |
| COMES_MT | | 0.405 ± 0.03 | 0.423 ± 0.02 | 0.434 ± 0.02 | 0.409 ± 0.03 |
| COMET_QE | 0.745 | | | | |
| COMES_QE | | 0.264 ± 0.06 | 0.592 ± 0.04 | 0.309 ± 0.06 | 0.490 ± 0.06 |
| COMES_MT_QE | | 0.457 ± 0.05 | 0.473 ± 0.04 | 0.472 ± 0.04 | 0.460 ± 0.05 |

Table 5: System-level Kendall's Tau correlations on the REALSumm corpus annotated by Bhandari et al. (2020). The three metrics with the highest correlation in each column are bolded.

### A.3 Human Feedback data results

Table 6 presents the system-level Kendall's Tau correlations on the subset of the test split of the CNN/DailyMail corpus annotated by Stiennon et al. (2020). The columns indicate different evaluation dimensions in the annotated (test) data. In the rows, we include outputs from each of the COMES heads, that correspond to evaluation dimensions used in the training data. From the analysis, we excluded 6 articles that appear in the SummEval dataset. In Table 7, we present the corresponding numbers when evaluating on the subset of the TL;DR corpus annotated by Stiennon et al. (2020) in a similar manner. For the COMES variants that we trained ourselves we evaluate with models trained on each cross-validation fold, reporting mean and standard deviation, see Section 4.1 for details.

| **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** |

Table 6: System-level Kendall's Tau correlations on the subset of CNN/DailyMail corpus annotated by Stiennon et al. (2020). The three metrics with the highest correlation in each column are bolded.

### A.4 Multi_SummEval results

In Table 8, we report the summary-level (segment-level) Pearson correlations on the subset of Multi_SummEval corpus annotated by Koto et al. (2021). Koto et al. (2021) collected human judgments for *Focus* and *Coverage*, using the Direct Assessment method to collect scores on a continuous scale of 1 to 100. For other metrics, see Table 2 in Koto et al. (2021). For readability reasons, we report only the mean COMES scores and do not report variance, see Section 4.1 for details.

| Metric | | Overall | Accuracy | Coverage | Coherence |
|---|---|---|---|---|---|
| ROUGE-1 f | | 0.545 | 0.000 | 0.576 | 0.333 |
| ROUGE-2 f | | 0.576 | 0.091 | 0.606 | **0.424** |
| ROUGE-L f | | 0.606 | 0.061 | 0.636 | **0.394** |
| BERTScore f | | 0.424 | −0.121 | 0.455 | 0.212 |
| COMET | | **0.727** | −0.061 | **0.758** | 0.273 |
| COMES | Coherence | −0.058 ± 0.19 | **0.306 ± 0.15** | −0.052 ± 0.18 | 0.124 ± 0.09 |
| | Consistency | 0.239 ± 0.05 | 0.082 ± 0.01 | 0.209 ± 0.05 | −0.003 ± 0.05 |
| | Fluency | 0.227 ± 0.09 | −0.106 ± 0.04 | 0.258 ± 0.09 | 0.039 ± 0.04 |
| | Relevance | 0.600 ± 0.12 | 0.042 ± 0.08 | 0.630 ± 0.12 | 0.315 ± 0.08 |
| COMES_MT | Coherence | **0.682 ± 0.02** | −0.100 ± 0.03 | **0.712 ± 0.02** | 0.294 ± 0.03 |
| | Consistency | 0.536 ± 0.14 | −0.155 ± 0.05 | 0.567 ± 0.14 | 0.215 ± 0.09 |
| | Fluency | 0.561 ± 0.10 | −0.161 ± 0.07 | 0.591 ± 0.10 | 0.233 ± 0.07 |
| | Relevance | **0.676 ± 0.03** | −0.112 ± 0.03 | **0.706 ± 0.03** | 0.282 ± 0.03 |
| COMET_QE | | 0.545 | **0.121** | 0.576 | **0.394** |
| COMES_QE | Coherence | 0.088 ± 0.27 | **0.258 ± 0.14** | 0.100 ± 0.27 | 0.173 ± 0.15 |
| | Consistency | 0.206 ± 0.11 | 0.085 ± 0.06 | 0.182 ± 0.11 | 0.012 ± 0.08 |
| | Fluency | 0.218 ± 0.11 | −0.073 ± 0.06 | 0.248 ± 0.11 | 0.055 ± 0.06 |
| | Relevance | 0.533 ± 0.09 | 0.085 ± 0.07 | 0.564 ± 0.09 | 0.315 ± 0.07 |
| COMES_MT_QE | Coherence | 0.564 ± 0.04 | 0.048 ± 0.04 | 0.594 ± 0.04 | **0.394 ± 0.02** |
| | Consistency | 0.491 ± 0.11 | 0.012 ± 0.08 | 0.521 ± 0.11 | 0.321 ± 0.09 |
| | Fluency | 0.473 ± 0.11 | 0.000 ± 0.07 | 0.503 ± 0.11 | 0.297 ± 0.10 |
| | Relevance | 0.555 ± 0.05 | 0.058 ± 0.04 | 0.585 ± 0.05 | 0.385 ± 0.03 |

Table 7: System-level Kendall's Tau correlations on the subset of TL;DR corpus annotated by Stiennon et al. (2020). The three metrics with the highest correlation in each column are bolded.

| | | **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 |

Table 8: Summary-level Pearson correlations on the Multi_SummEval corpus annotated by Koto et al. (2021). The three metrics with the highest correlation in each column are bolded.