Reevaluating Summarisation Evaluation

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

Shoulders of Giants ...

- In the context of machine translation, Yorick Wilks famously quipped (Wilks, 2009):

  the evaluation of MT systems is almost certainly more developed than MT itself

Shoulders of Giants ...

- In the context of machine translation, Yorick Wilks famously quipped (Wilks, 2009):

  the evaluation of MT systems is almost certainly more developed than MT itself

- In the case of summarisation, the actuality is perhaps more like:

  the evaluation of summarisation metrics is almost certainly more developed than the summarisation metrics themselves

  with recent papers on the evaluation of summarisation metrics and datasets including Bommasani and Cardie (2020); Bhandari et al. (2020a,b); Fabbri et al. (2020); Pagnoni et al. (2021)

Areas of Focus in this Talk

- **RQ1:** What are the complexities of summarisation evaluation?

Source(s): Koto et al. (2021a,b)

Areas of Focus in this Talk

- **RQ1:** What are the complexities of summarisation evaluation?
- **RQ2:** What is current practice in terms of (English) summarisation evaluation?

Source(s): Koto et al. (2021a,b)

Areas of Focus in this Talk

- **RQ1:** What are the complexities of summarisation evaluation?
- **RQ2:** What is current practice in terms of (English) summarisation evaluation?
- **RQ3:** How can we improve on the shortcomings in current practice?

Source(s): Koto et al. (2021a,b)

Areas of Focus in this Talk

- **RQ1:** What are the complexities of summarisation evaluation?
- **RQ2:** What is current practice in terms of (English) summarisation evaluation?
- **RQ3:** How can we improve on the shortcomings in current practice?
- **RQ4:** What has the impact of current evaluation practice been on English summarisation research?

Source(s): Koto et al. (2021a,b)

Areas of Focus in this Talk

- **RQ1:** What are the complexities of summarisation evaluation?
- **RQ2:** What is current practice in terms of (English) summarisation evaluation?
- **RQ3:** How can we improve on the shortcomings in current practice?
- **RQ4:** What has the impact of current evaluation practice been on English summarisation research?
- **RQ5:** How well do existing automatic metrics perform over languages other than English?

Source(s): Koto et al. (2021a,b)

Talk Outline

Contents

(Very!) Potted History of Summarisation

- Early work focused on multi-document summarisation = given a cluster of documents, generate a combined summary
- More recently, research has focused heavily on single-document summarisation, in large part because of data availability
- Early methods were largely "extractive" (= extract *n*-grams from documents, and combine them in the summary) and struggled in terms of coherence; modern methods are largely "abstractive" (= generate a summary from the source document(s)), but also some hybrid methods

Contents

ROUGE (Lin and Hovy, 2003)

- ROUGE ("Recall-Oriented Understudy for Gisting Evaluation") methodology:
  - ROUGE-n (n-gram overlap)

    For the summary S and each reference R_i, generate the multiset of n-grams via $gram_n(T) = \{\!\{\langle w^{(1)} \ldots w^{(n)} \rangle, \langle w^{(2)} \ldots w^{(n+1)} \rangle, \ldots, \langle w^{(l-n+1)} \ldots w^{(n)} \rangle\}\!\}$ (where $l = \text{len}(T)$)

$$\mathcal{P}_n = \arg\max_i \frac{|gram_n(R_i) \cap gram_n(S)|}{|gram_n(S)|}$$

$$\mathcal{R}_n = \arg\max_i \frac{|gram_n(R_i) \cap gram_n(S)|}{|gram_n(R_i)|}$$

$$\text{ROUGE-}n = \frac{2\mathcal{P}_n\mathcal{R}_n}{\mathcal{P}_n + \mathcal{R}_n}$$

ROUGE (Lin and Hovy, 2003)

- Example:

R: Bayern Munich beat Porto 6 - 1 in the Champions League on Tuesday

S: Bayern Munich wins in the Champions League

For $N = 2$:

$$\mathcal{P}_2 = \frac{4}{7} \quad \mathcal{R}_2 = \tfrac{4}{13} \quad \mathrm{ROUGE\text{-}2} = 0.4$$

ROUGE (Lin and Hovy, 2003)

- $\mathrm{ROUGE\text{-}LCS}$ (LCS overlap)

  For the summary S and each reference R_i, calculate the longest common (n-gram) subsequence $\mathsf{LCS}(S, R_i)$

$$\mathcal{P}_{LCS} = \arg\max_i \frac{\mathsf{len}(\mathsf{LCS}(S, R - i)}{\mathsf{len}(S)}$$

$$\mathcal{R}_{LCS} = \arg\max_i \frac{\mathsf{len}(\mathsf{LCS}(S, R - i)}{\mathsf{len}(R_i)}$$

$$\mathrm{ROUGE\text{-}LCS} = \frac{2\mathcal{P}_{LCS}\mathcal{R}_{LCS}}{\mathcal{P}_{LCS} + \mathcal{R}_{LCS}}$$

ROUGE (Lin and Hovy, 2003)

- Example:

R: Bayern Munich beat Porto 6 - 1 in the Champions League on Tuesday

S: Bayern Munich wins in the Champions League

$$\mathcal{P}_{LCS} = \frac{4}{7} \quad \mathcal{R}_{LCS} = \frac{4}{13} \quad \text{ROUGE-LCS} = 0.4$$

- Many, many other variants (Graham, 2015), but these generally perform the best

Repurposed MT Evaluation Metrics

- Given that ROUGE is based on string overlap, an obvious alternative is MT evaluation metrics such as:
  - BLEU (Papineni et al., 2002)

$$P_k(S, R_i) = \frac{|gram_k(S) \cap gram_k(R_i)|}{|gram_k(R_i)|}$$

$$\mathrm{BP}(S, R_i) = \begin{cases} 1 & \text{if } \mathrm{len}(S) > \mathrm{len}(R_i) \\ e^{(1-\mathrm{len}(R_i)/\,\mathrm{len}(*))} & \text{if } \mathrm{len}(S) \leq \mathrm{len}(R_i) \end{cases}$$

$$\mathrm{BLEU}(S, R_i) = \mathrm{BP}(S, R_i) \cdot \left((P_1(S, R_i) \cdot P_2(S, R_i) \cdot P_3(S, R_i) \cdot P_4(S, R_i))^{1/4} \right)$$

  - METEOR (Banerjee and Lavie, 2005): weighted F-score of n-gram overlap, with stemming and synonym matching, and "chunk" penalty

Pyramid (Nenkova and Passonneau, 2004)

- Methodology:
  1. Translate each reference summary into “semantic content units” (SCUs)
  2. Merge SCUs across reference summaries, weighting them according to the number of references they occur in
  3. (Weighted) score a summary based on how many SCUs can be inferred from it

Pyramid (Nenkova and Passonneau, 2004)

- Example:

R: Bayern Munich beat Porto 6 - 1 in the Champions League on Tuesday

S: Bayern Munich wins in the Champions League

SCUs with evaluations:

- Bayern Munich beat Porto ✗
- Bayern Munich won 6 - 1 ✗
- Bayern Munich won in Ch. Lg. ✓
- Bayern Munich won on Tuesday ✗

BERTScore (Zhang et al., 2020b)

- For summarisation evaluation, BERTScore is calculated as:

$$\mathcal{P}_{BERT} = \frac{1}{|S|} \sum_{t_j \in S} \max_{s_k \in R_i} t_j^T s_k$$

$$\mathcal{R}_{BERT} = \frac{1}{|R_i|} \sum_{s_k \in R_i} \max_{t_j \in S} t_j^T s_k$$

$$\mathcal{F}_{BERT} = 2 \frac{\mathcal{P}_{BERT} \cdot \mathcal{R}_{BERT}}{\mathcal{P}_{BERT} + \mathcal{R}_{BERT}}$$

where s_k and t_j are (contextualised) token embeddings of R_i and S.

BERTScore (Zhang et al., 2020b)

- Within the confines of this formulation of BERTScore, we will experiment with alternatives to BERT (Devlin et al., 2019), including RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2019), BART (Lewis et al., 2020), and PEGASUS (Zhang et al., 2020a)

STS-Score

- We can also calculate string similarity for string embeddings using an STS scorer trained on STS data (Agirre et al., 2012):

$$\mathcal{P}_{STS} = \frac{1}{|S|} \sum_{t_j \in S} \max_{s_k \in R_i} \mathrm{STS}(t_j, s_k)$$

$$\mathcal{R}_{STS} = \frac{1}{|R_i|} \sum_{s_k \in R_i} \max_{t_j \in S} \mathrm{STS}(s_k, t_j)$$

$$\mathcal{F}_{STS} = 2 \frac{\mathcal{P}_{STS} \cdot \mathcal{R}_{STS}}{\mathcal{P}_{STS} + \mathcal{R}_{STS}}$$

  based on different segment granularities s_k and t_j, such as sentence or document
- Experiments based on SBERT (Reimers and Gurevych, 2019)

Question Answering-based Evaluation

- In the context of faithfulness evaluation, Wang et al. (2020) proposed the "QAGS" evaluation framework, made up of two components: (1) question generation, and (2) question answering.
    1. Given D, R_i, and S (the source document, reference summary, and system summary, resp.), train a model to generate questions Q from system summary S (based on `bart-large` trained on NewsQA (Trischler et al., 2017))
    2. Given Q, predict answer A based on two terms: $p(A|Q,D)$ and $p(A|Q,S)$ (based on a QA model trained using `bert-large-wwm` over SQuAD 2.0 (Jia et al., 2018))
    3. Measure performance based on the F1 score between the answers generated from D and S

RQ1: What are the Complexities of Summarisation Evaluation?

- What is a good summary anyway ... far from clear that single figure-of-merit evaluation is the right way to go
- Tied to the (often single) reference summary, despite the inherent complexity/diversity in possible summaries for a given document ... moreso than MT
- With the move to abstractive methods, "faithfulness"/hallucination has been identified as a key issue (Maynez et al., 2020; Wang et al., 2020; Durmus et al., 2020; Pagnoni et al., 2021)
- Sensitivity to tokenisation issues (Post, 2018; Deutsch and Roth, 2020; Marie et al., 2021)

RQ1: What are the Complexities of Summarisation Evaluation?

- Metrics have often been "validated" on different datasets/tasks, in many instances based on summarisation methods which are no longer current:

ROUGE	DUC 2001–2003 (MDS)
BERTScore	WMT

👾((((yoav' ()ل)()ل)))
@yoavgo

unpopular (?) take: for publishable modern NLP/ML work, "never look at the test data" should be replaced with something less convenient and more relevant. like "always look at the test errors", and "never evaluate using BLEU or ROUGE".

4:32 AM · Sep 1, 2021 · Twitter Web App

4 Retweets **1** Quote Tweet **55** Likes

Source(s): `https://twitter.com/yoavgo/status/1432773205962469382`

RQ2: What is Current Practice in Terms of (English) Summarisation Evaluation?

- Based on a survey of 111 summarization papers from major NLP conferences over the period 2017–2020 (extending Hardy et al. (2019)):
  - ROUGE-* used by more than 95% of papers
  - Other metrics such as METEOR, BLEU, and BERTScore rarely used
  - 64% of papers included manual evaluation, with major dimensions being:
    1. faithfulness
    2. recall
    3. precision
    4. relevance
    5. coherence
    6. fluency

Source(s): Koto et al. (2021b)

Talk Outline

Contents

RQ3: How can we Improve on the Shortcomings in Current Practice?

- Based on our analysis of how manual evaluation has been carried out, we propose to separate summarisation evaluation across the four dimensions of:
  1. **Faithfulness**: degree of factual consistency with the source
  2. **Focus**: precision of summary content relative to the reference
  3. **Coverage**: recall of summary content relative to the reference
  4. **Inter-sentential Coherence**: document fluency between adjacent sentences

Source(s): Koto et al. (2021b)

Evaluating FAITHFULNESS

- Basic intuition: all content in the generated summary should be factually-consistent with the source document
- Score by comparing summaries with the source document as follows:

$$\mathrm{FA}_{\mathrm{METRIC}} = \frac{1}{|S|} \sum_{t_i \in S} \mathrm{A}(t_i, D, n)$$

$$\mathrm{A}(t_i, D, n) = \operatorname*{AvgTop\text{-}\mathit{n}}_{s_j \in D} \ \mathrm{METRIC}(t_i, s_j)$$

where METRIC $\in$ {ROUGE-*, STS-SCORE, BERTSCORE }, and AvgTop- n matches sentence t_i from the summary with each sentence $s_j \in D$, and returns the average score for the top-n best-matching sentences.

- QAGS can be used directly

Evaluating Focus and Coverage

- For ROUGE-*, BLEU, METEOR, STS-Score, BERTScore, use precision for Focus and recall for Coverage
- For QAGS:
  - in case of Focus, generate questions based on system summary S, and answer the questions based on the system summary S vs. reference summary R_i
  - in case of Coverage, generate questions based on the *reference* summary R_i, and answer those questions based on the system summary S vs. reference summary R_i

Evaluating Inter-sentential Coherence

- Extend Nayeem and Chali (2017) in training a next-sentence-prediction (NSP) classifier as follows:

$$\mathrm{NSP}(S) = \mathrm{mean}_{t_i \in S}\, \mathrm{NSP}(t_i, t_{i+1})$$

where $t_i \in S$, and $\mathrm{NSP}(t_i, t_{i+1})$ returns the probability of t_{i+1} following t_i

Contents

Dataset Outline

- In the absence of a dataset annotated with FFCI, we construct one ourselves:
  - FAITHFULNESS: 2000 samples from Maynez et al. (2020), based on summaries generated by 4 neural models over XSUM (Narayan et al., 2018): pointer generator network ("PG": (See et al., 2017)), Topic-aware convolutional Seq2Seq ("TCONV": (Narayan et al., 2018)), a transformer-based model ("TRANS2S": (Vaswani et al., 2017)), and BERT ("BERT": (Devlin et al., 2019; Liu and Lapata, 2019))
  - FOCUS and COVERAGE: randomly sample 135 articles each from CNN/DailyMail (Hermann et al., 2015) and XSUM, and generate summaries with PG (See et al., 2017) and BERT (Liu and Lapata, 2019), resulting in 540 summaries ($135 \times 2 \times 2$)
  - INTER-SENTENTIAL COHERENCE: used the same 270 system summaries from CNN/DailyMail as for FOCUS and COVERAGE

Dataset Outline

- Focus, Coverage, and Inter-sentential Coherence annotations are based on (customised) Direct Assessment (Graham et al., 2015; Graham et al., 2017)

Focus and Coverage Annotation

How much information contained in the black text can also be found in the gray text?

officials at the famous yellowstone national park in the us have revealed that they had to put down a newborn bison after some tourists put it in the boot of their car .

wildlife rangers in the us state of wyoming have warned visitors to stay away from their herd after they refused a controversial bison .

0 % 100 %

NEXT

Inter-sentential Coherence Annotation

The text below has good inter-sentential coherence (i.e. the flow from one sentence to the next is natural):

sony emails reveal bbc bosses want to turn the hit series starring peter capaldi and is screened in 50 countries , to be turned into a movie to capitalise on its worldwide success .

but the emails show doctor who's creative team are reluctant to rush into making a film that could flop and tarnish its reputation .

Strongly disagree | Strongly agree

NEXT

Contents

Experiments

- We first assess the ability of the different metrics to capture FAITHFULNESS, FOCUS, COVERAGE, and INTER-SENTENTIAL COHERENCE, based on the gold-standard datasets, using Pearson correlation r and Spearman rank correlation ρ

Metric Evaluation: Faithfulness

Metric	r	ρ
Against reference		
ROUGE-1	0.199	0.199
ROUGE-2	0.116	0.161
BLEU-4	0.072	0.133
METEOR	0.131	0.170
BERTScore	0.128	0.131
Against source sentences		
QA (Maynez et al., 2020)	—	0.044
Entailment (Maynez et al., 2020)	—	0.431
QAGS (our implementation)	0.250	0.270
FA_{STS}	0.260	0.258
$\text{FA}_{\text{ROUGE-1}}$	0.361	0.361
$\text{FA}_{\text{ROUGE-2}}$	0.311	0.315
$\text{FA}_{\text{BERTscore}}$	0.178	0.179
$\text{FA}_{\text{BERTscore}}$ (Ours)	**0.476**	**0.474**

Metric Evaluation: Focus

Metric	Focus			
	C-PG	**C-BT**	**X-PG**	**X-BT**
ROUGE-1	0.607	0.623	0.540	0.562
ROUGE-2	0.595	0.552	0.564	0.454
ROUGE-LCS	0.604	0.619	0.528	0.552
METEOR	—	—	—	—
BLEU-4	0.511	0.442	0.526	0.304
QAGS	0.543	0.611	0.541	0.527
STS-Score (sentence)	0.524	0.526	0.444	**0.617**
STS-Score (doc)	0.524	0.569	0.444	**0.617**
BERTScore	0.552	0.519	0.427	0.406
BERTScore (Ours)	**0.665**	**0.625**	**0.577**	0.581

Metric Evaluation: Coverage

Metric	Coverage			
	C-PG	C-BT	X-PG	X-BT
ROUGE-1	0.592	0.641	0.480	0.514
ROUGE-2	0.547	0.569	0.463	0.437
ROUGE-LCS	0.581	0.636	0.482	0.487
METEOR	0.597	0.660	0.523	0.601
BLEU-4	—	—	—	—
QAGS	0.570	0.608	0.452	0.513
STS-Score (sentence)	0.559	0.572	0.559	**0.641**
STS-Score (doc)	0.513	0.508	0.559	**0.641**
BERTScore	0.549	0.579	0.363	0.359
BERTScore (Ours)	**0.680**	**0.695**	**0.617**	**0.623**

Metric Evaluation: Inter-sentential Coherence

Metric	IC	
	C-PG	C-BT
ROUGE-1	0.097	0.138
ROUGE-2	−0.004	0.083
ROUGE-LCS	0.088	0.114
METEOR	0.061	0.143
BLEU-4	−0.030	0.090
STS-Score (doc)	0.124	0.197
BERTScore	0.042	0.152
BERTScore (Ours)	0.055	0.132
Nayeem and Chali (2017)	−0.275	0.166
NSP	**0.388**	**0.351**

Findings

- ROUGE-*, METEOR, and BLEU worse than model-based metrics in all cases
- QAGS performs poorly for Focus and Coverage, and also below (task-optimised) BERTScore for Faithfulness
- Our BERTScore results (based on `gpt2-xl`) better than original due to task-specific model and layer selection
- BERTScore performs the best for Focus (layer 29) and Coverage (layer 4)
- In terms of metric reliability over the four dimensions of FFCI: Coverage $>$ Focus $\gg$ Faithfulness $\gg$ coherence

Contents

RQ4: What has the Impact of Current Evaluation Practice been on English Summarisation Research?

- Given that a lot of “leaderboarding” of summarisation research has been based on ROUGE-*, and ROUGE-* is noisy, where has it led us?

Results for CNN/DM: ROUGE-* vs. FFCI

Method	ROUGE			FFCI			
	R-1	R-2	R-L	Fa	Fo	C	IC
LEAD3	40.1	17.3	36.3	91.2	49.2	70.9	65.3
Abstractive							
PG (See et al., 2017)	36.4	15.7	33.4	90.9	52.1	65.6	52.8
PG+C (See et al., 2017)	39.5	17.3	36.4	**91.1**	52.4	68.6	67.2
rnn+RL+rerank (Chen and Bansal, 2018)	40.9	17.8	38.5	89.6	53.4	70.2	56.4
BOTTOM-UP (Gehrmann et al., 2018)	41.5	18.7	38.6	90.0	55.3	68.5	65.3
BERTSUMEXTABS (Liu and Lapata, 2019)	42.1	19.4	39.1	89.8	51.9	68.7	65.7
BART (Lewis et al., 2020)	44.3	21.1	41.2	89.5	52.6	69.5	69.6
PEGASUS (Zhang et al., 2020a)	44.4	21.5	41.4	89.9	**56.0**	70.8	69.5
PROPHETNET (Yan et al., 2020)	44.4	21.2	41.5	89.9	55.9	**72.0**	**70.0**

Results for CNN/DM: ROUGE-* vs. FFCI

Method	ROUGE			FFCI			
	R-1	R-2	R-L	Fa	Fo	C	IC
LEAD3	40.1	17.3	36.3	91.2	49.2	70.9	65.3
Extractive							
BanditSum (Dong et al., 2018)	41.6	18.7	37.9	91.8	51.5	71.6	61.5
PNBERT (Zhong et al., 2019)	42.7	19.5	38.8	**91.9**	51.9	**73.5**	**66.2**
BERTSUMEXT (Liu and Lapata, 2019)	43.3	20.2	39.7	91.8	52.2	73.0	61.8
MATCHSUM (Zhong et al., 2020)	44.4	20.8	40.6	91.9	**53.3**	72.4	62.5

Results for XSUM: ROUGE-* vs. FFCI

Method	ROUGE			FFCI			
	R-1	R-2	R-L	Fa	Fo	C	IC
LEAD1	16.3	1.6	12.0	90.3	35.3	50.1	—
PG (See et al., 2017)	29.7	9.2	23.2	85.2	45.0	57.1	—
TCONV (Narayan et al., 2018)	31.9	11.5	25.8	85.2	49.4	57.7	—
BERTSUMEXTABS (Liu and Lapata, 2019)	38.8	16.5	31.3	85.6	53.7	62.3	—
BART (Lewis et al., 2020)	45.1	22.3	37.3	**86.6**	61.9	69.0	—
PEGASUS (Zhang et al., 2020a)	47.2	24.6	39.3	86.5	**64.6**	**69.5**	—

Findings

- FAITHFULNESS not a big issue for CNN/DM (although extractive methods unsurprisingly slightly better); more of an issue for XSUM, with slow upward trend for abstractive methods
- In terms of COVERAGE, little improvement over CNN/DM for abstractive methods until very recently; clear progress in FOCUS, with some bumps in the road
- Similarly for XSUM, most improvements in FOCUS
- Clear separation between PEGASUS and BART, esp. in COVERAGE
- Slight improvements in INTER-SENTENTIAL COHERENCE over time (and abstractive $>$ extractive)

Talk Outline

Summarisation Evaluation beyond English

- ROUGE is commonly applied to languages other than English, incl. Chinese, Indonesian, Spanish, Russian, Vietnamese, French, German, Spanish, and Turkish (Hu et al., 2015; Scialom et al., 2020; Ladhak et al., 2020; Koto et al., 2020) ... without any explicit validation of its performance outside English
- Particular concerns:
  - morphology
  - free-er word order languages

Source(s): Koto et al. (2021b)

Summarisation Evaluation beyond English

- ROUGE is commonly applied to languages other than English, incl. Chinese, Indonesian, Spanish, Russian, Vietnamese, French, German, Spanish, and Turkish (Hu et al., 2015; Scialom et al., 2020; Ladhak et al., 2020; Koto et al., 2020) ... without any explicit validation of its performance outside English
- Particular concerns:
  - morphology
  - free-er word order languages
- **RQ5:** How well do existing automatic metrics perform over languages other than English (and can we improve on them)?

Source(s): Koto et al. (2021b)

Summarisation Evaluation beyond English

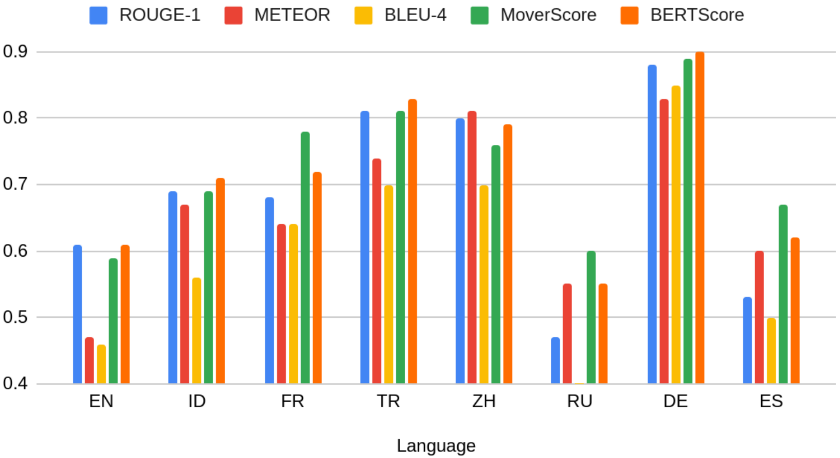
- To explore these questions, we construct a dataset of summaries (Focus + Coverage) for 8 languages:

  English (EN), Indonesian (ID), French (FR), Turkish (TR), Mandarin Chinese (ZH), Russian (RU), German (DE), and Spanish (ES)

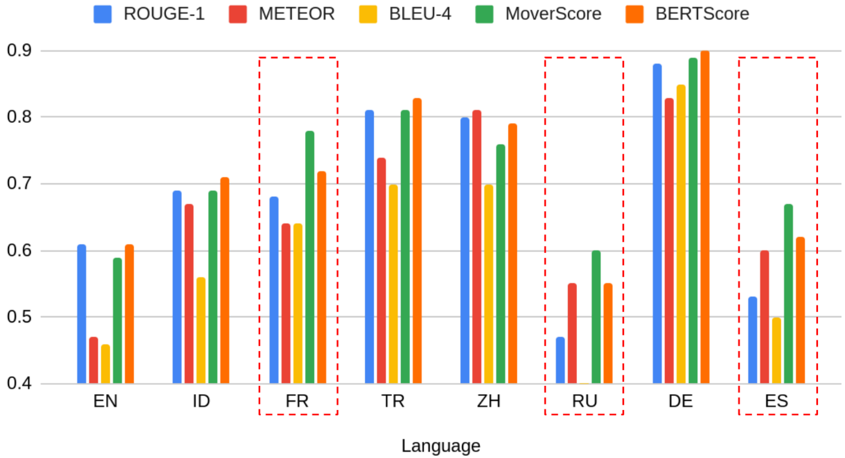
  based on two contemporary abstractive summarisation methods:
  - LSTM-based Pointer Generator Network (See et al., 2017)
  - BERT-based summarisation model (Liu and Lapata, 2019; Dong et al., 2019)
- Total of 8 languages $\times$ 135 documents $\times$ 2 models $\times$ 2 criteria (= Focus and Coverage) $\times$ 3 annotators = 12,960 annotations, based on (localised) DA (Graham et al., 2015; Graham et al., 2017)

Source(s): Koto et al. (2021b)

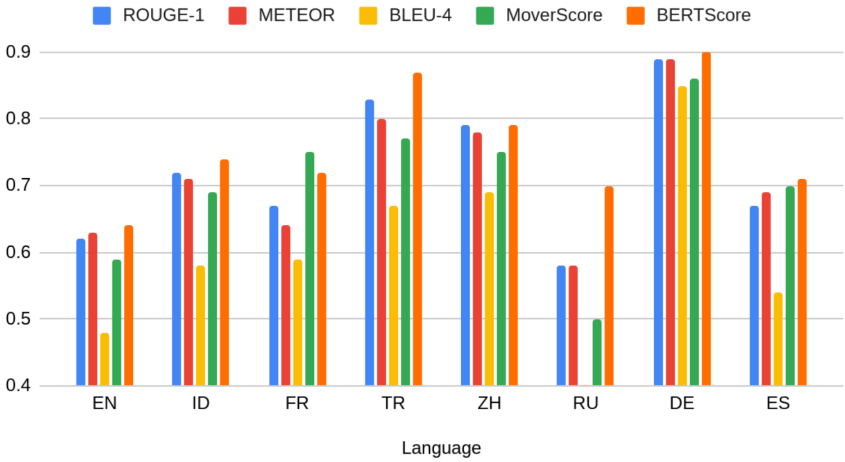
Results: Focus (Pearson's r)



Results: Focus (Pearson's r)



Results: Coverage (Pearson's r)



Overall Findings

- Performance of ROUGE-1 actually remarkably high(!) $\sim$ ROUGE-L $>$ ROUGE-2 $>$ ROUGE-3 (both FOCUS and COVERAGE)
- Best overall results for BERTscore (but MoverScore better for some langs)
- Marginally higher results for monolingual BERT models (and monolingual layer selection) with BERTscore; mBERT $>$ XLM
- **Overall recommendation** $=$ use BERTScore with mBERT uncased (precision $=$ FOCUS; recall $=$ COVERAGE)

Acknowledgements

- Joint work with Fajri Koto and Jey Han Lau
- Supported by Australian Research Council and Australia Awards

References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, USA.

References

Bhandari, M., Gour, P. N., Ashfaq, A., and Liu, P. (2020a). Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bhandari, M., Gour, P. N., Ashfaq, A., Liu, P., and Neubig, G. (2020b). 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.

Bommasani, R. and Cardie, C. (2020). Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

References

Chen, Y.-C. and Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Deutsch, D. and Roth, D. (2020). SacreROUGE: An open-source library for using and developing summarization evaluation metrics. *CoRR*, abs/2007.05374.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32, pages 13063–13075.

Dong, Y., Shen, Y., Crawford, E., van Hoof, H., and Cheung, J. C. K. (2018). Banditsum: Extractive summarization as a contextual bandit. In *EMNLP 2018: 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748.

Durmus, E., He, H., and Diab, M. (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.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2020). Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

References

Gehrmann, S., Deng, Y., and Rush, A. (2018). Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Graham, Y., Baldwin, T., and Mathur, N. (2015). Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

References

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Hardy, H., Narayan, S., and Vlachos, A. (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.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1693–1701.

References

Hu, B., Chen, Q., and Zhu, F. (2015). LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Jia, R., Rajpurkar, P., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *ACL 2018: 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 784–789.

Koto, F., Baldwin, T., and Lau, J. H. (2021a). Ffci: A framework for interpretable automatic evaluation of summarization.

References

Koto, F., Lau, J. H., and Baldwin, T. (2020). Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Koto, F., Lau, J. H., and Baldwin, T. (2021b). Evaluating the efficacy of summarization evaluation across languages. In *Findings of ACL*.

Ladhak, F., Durmus, E., Cardie, C., and McKeown, K. (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

References

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). 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.

Lin, C.-Y. and Hovy, E. H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

References

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marie, B., Fujita, A., and Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (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.

References

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Nayeem, M. T. and Chali, Y. (2017). Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56, Vancouver, Canada. Association for Computational Linguistics.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, USA.

References

Pagnoni, A., Balachandran, V., and Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL and 3rd Annual Meeting of the NAACL (ACL-02)*, pages 311–318, Philadelphia, USA.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

References

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (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.

References

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008.

References

Wang, A., Cho, K., and Lewis, M. (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. Association for Computational Linguistics.

Wilks, Y. (2009). The future of MT in the new millennium. *Machine Translation*, pages 225–236.

Yan, Y., Qi, W., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. (2020). Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 5753–5763.

References

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020a). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML 2020: 37th International Conference on Machine Learning*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with bert. In *ICLR 2020 : Eighth International Conference on Learning Representations*.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020). Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

References

Zhong, M., Liu, P., Wang, D., Qiu, X., and Huang, X. (2019). Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, Florence, Italy. Association for Computational Linguistics.