ESTIME: Estimation of Summary-to-Text Inconsistency by Mismatched Embeddings

Oleg Vasilyev, John Bohannon
Primer Technologies Inc.
San Francisco, California
oleg,john@primer.ai

Abstract

We propose a new reference-free summary quality evaluation measure, with emphasis on the faithfulness. The measure is based on finding and counting all probable potential inconsistencies of the summary with respect to the source document. The proposed ESTIME, Estimator of Summary-to-Text Inconsistency by Mismatched Embeddings, correlates with expert scores in summary-level SummEval dataset stronger than other common evaluation measures not only in Consistency but also in Fluency. We also introduce a method of generating subtle factual errors in human summaries. We show that ESTIME is more sensitive to subtle errors than other common evaluation measures.

1 Introduction

Summarization must preserve the factual consistency of the summary with the text. Human annotation of factual consistency can be accompanied with detailed classification of factual errors, thus giving a hope that the annotation scores are reasonably objective (Kryscinski et al., 2020; Huang et al., 2020; Vasilyev et al., 2020b; Gabriel et al., 2020; Maynez et al., 2020).

Factual consistency of a summary is one of several summary qualities; for the purpose of human annotation these qualities can be specified in different ways (Xenouleas et al., 2019; Kryscinski et al., 2020; Fan et al., 2018; Vasilyev et al., 2020b; Fabbri et al., 2020). Summarization models nowadays create satisfactorily fluent, coherent and informative summaries, but the factual consistency suffers from hallucinations, entity swaps and other errors. Some factual errors are easily noticeable; other factual errors could be hardly noticeable even for annotators (Lux et al., 2020; Vasilyev et al., 2020b) - which is arguably even worse.

Existing summary evaluation measures are based on several approaches, which may be sensitive to some qualities more than to others. A question-answering based evaluation estimates how helpful is the summary in answering questions about the source text (Xenouleas et al., 2019; Eyal et al., 2019; Scialom et al., 2019; Deutsch et al., 2020; Durmus et al., 2020; Wang et al., 2020). A text reconstruction approach estimates how helpful is the summary in guessing parts of the source text (Vasilyev et al., 2020a,b; Egan et al., 2021). Evaluation measures that use some kind of text similarity can estimate how similar is the summary to special human-written reference summaries (Zhang et al., 2020; Zhao et al., 2019; Lin, 2004), or, more realistically, how similar is the summary to the source text (Gao et al., 2020; Louis and Nenkova, 2009).

In order to assess how well an evaluation measure works for factual consistency, it is necessary either to have a dataset of human-annotated imperfect machine-generated summaries (Bhandari et al., 2020; Fabbri et al., 2020), or to have a dataset of artificially introduced factual errors in originally factually correct human-written summaries (Kryscinski et al., 2020).

In this paper we focus on presenting a new evaluation measure with emphasis on factual consistency. Our contribution:

1. We introduce **ESTIME**: Estimator of Summary-to-Text Inconsistency by Mismatched Embeddings[1]. Using human-annotated machine-generated summaries of SummEval (Fabbri et al., 2020), we compare ESTIME with other evaluation measures.
2. We introduce a natural method of generating subtle factual errors. We use it here to compare the performance of ESTIME with other measures on human-written summaries with generated subtle errors.

[1] https://github.com/PrimerAI/blanc/tree/master/estime

2 Methods

The motivations for our estimator:

1. Any location in a summary has a context that loosely corresponds to a context in one or more locations in the text.
2. In the most similar context, the summary would normally use the same word that was used in the text.
3. Summary generation models produce very few new (not from the text) words per summary.
4. Transformer-made token embeddings are highly contextual (Ethayarajh, 2019).

In order to estimate the consistency of a summary with the text, we attempt to count all the summary tokens that could be potentially related to a factual error. To this end, we check embeddings of all the tokens of the summary that have one or more occurrences in the text. For each embedding we find its match: the most similar embedding in the text. If the corresponding tokens are not the same, we add up such mismatch into our score of inconsistency. Our goal is not an error correction or precise location of errors, but a score estimating the summary consistency quality. The algorithm is simple:

1. Obtain embeddings for all tokens in the text. In the summary, obtain embeddings only for the tokens that occur at least once in the text. To obtain an embedding of a token, mask the token, run a token-prediction transformer model on the context surrounding the token, and take the embedding of the token from a hidden layer.
2. For each of the obtained embeddings of the summary tokens, find the most similar embedding in the text. If the corresponding tokens do not coincide, count this as a potential inconsistency. The total number of such inconsistencies is our score, ESTIME.

We measure 'similarity' of embeddings by their scalar product. Thus, ESTIME score is the number N_a of 'alarms':

$$N_a = \sum_{i:t_i \in T} H(\max_{\beta:t'_\beta \neq t_i} (e_i e'_\beta) - s(i)) \quad (1)$$

$$s(i) \equiv \max_{\alpha:t'_\alpha = t_i} (e_i e'_\alpha) \quad (2)$$

Here H is Heaviside function; the summary is a sequence of tokens t_i, each having embedding e_i; the text T is a sequence of tokens t'_α, having embeddings e'_α. The summation in Equation 1 is over all the summary tokens t_i that exist in the text T. The count N_a gets added $+1$ whenever the best match to e_i from the embeddings of unequal tokens e'_β exceeds the best match from the embeddings of occurrences of the same token $t_\alpha = t_i$ in the text.

The tokens are obtained by the tokenizer corresponding to the token-prediction transformer model. Notice that we do not verify the summary tokens that do not occur in the text. Such tokens still can influence the context used for embeddings of other tokens. The algorithm is asymmetric with respect to the summary and the text: it is supposed to estimate summary-to-text inconsistency.

This approach is different from matching embeddings for sake of measuring similarity (e.g. similarity between a summary and a reference summary in BERTScore (Zhang et al., 2020)), and from using a model trained to replace wrong tokens with correct ones (Cao et al., 2020; Kryscinski et al., 2020).

The embeddings are taken using the pretrained BERT model (Devlin et al., 2019) bert-large-uncased-whole-word-masking of Transformers library (Wolf et al., 2020). While there is no crucial difference with other varieties of BERT, ALBERT and RoBERTa, this model showed a better overall performance, and we used it for evaluations in the next sections. For the sake of faster processing, we do not run the model separately for each token, but at a single run obtain embeddings for all tokens separated by the distance of 8 tokens. This means that the context for each masked token is a little muddied by masking of a few other tokens, but the distance of 8 tokens is large enough for the effect to be negligible. The results of the next sections are obtained with input size of 450 tokens (close to max BERT input length). Finally, when input window does not touch the beginning or end of the text, we do not mask the tokens too close to the edge of the window: no masking within the margin of 50 tokens at the edges of the input window. The algorithm is simple, but for convenience we provide the code[2].

In the next sections we present results for the versions ESTIME-12 and ESTIME-24, corresponding to the embeddings from the middle (12th layer) and from the top (24th layer) of the large BERT; as explained later we also consider ESTIME-21.

[2]https://github.com/PrimerAI/blanc/tree/master/estime

measure	consistency		relevance		coherence		fluency	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLANC-AXXL	0.200	0.098	0.246	0.179	0.127	0.093	0.115	0.066
BLANC-BLU	0.207	0.102	0.217	0.156	0.116	0.085	0.112	0.065
(-)ESTIME-12	0.374	0.184	0.140	0.100	0.238	0.173	0.343	0.198
(-)ESTIME-21	**0.404**	**0.200**	0.188	0.134	0.300	0.217	**0.399**	**0.232**
(-)ESTIME-24	0.358	0.176	0.117	0.084	0.187	0.134	0.363	0.209
(-)J-Shannon	0.193	0.095	0.406	0.298	0.289	0.213	0.125	0.072
SummaQA-F1	0.174	0.085	0.16	0.113	0.089	0.065	0.12	0.069
SummaQA-P	0.197	0.097	0.179	0.127	0.112	0.082	0.133	0.076
SUPERT	0.297	0.147	0.306	0.222	0.236	0.175	0.175	0.101
BERTScore-F1	0.109	0.053	0.371	0.273	**0.377**	**0.277**	0.142	0.082
BERTScore-P	0.055	0.027	0.268	0.196	0.323	0.238	0.126	0.072
BERTScore-R	0.164	0.081	**0.423**	**0.309**	0.345	0.253	0.12	0.069
BLEU	0.095	0.047	0.213	0.153	0.176	0.128	0.140	0.080
ROUGE-L	0.115	0.057	0.241	0.174	0.170	0.124	0.079	0.045
ROUGE-1	0.137	0.067	0.302	0.220	0.184	0.134	0.080	0.046
ROUGE-2	0.129	0.063	0.245	0.177	0.146	0.105	0.063	0.036
ROUGE-3	0.149	0.073	0.251	0.180	0.160	0.116	0.066	0.038

Table 1: Summary level correlations ρ (Spearman) and τ (Kendall Tau-c) of quality estimators with human experts scores. The top rows evaluation measures are reference-free, separated from the lower rows evaluation measures, which need human references. In each column the highest correlation is bold-typed. The only p-values above 0.01 in this table are p=0.03 for BERTScore-P and p=0.01 for ROUGE-2.

3 Performance on human-annotated machine-generated summaries

3.1 Correlations with expert scores

We used SummEval dataset [3] (Fabbri et al., 2020) for comparing ESTIME with a few well known or promising evaluation measures. The part of SummEval dataset that we use consists of 100 texts, each text is accompanied by 16 summaries generated by 16 different models, making altogether 1600 text-summary pairs. Each text-summary pair is annotated (on scale 1 to 5) by 3 experts for 4 qualities: consistency, relevance, coherence and fluency. We took average of the expert scores for each quality of a text-summary pair. Each text is also accompanied by 11 human-written reference summaries, for the measures that need them. (In latest version of (Fabbri et al., 2020) a 17th model - Pegasus dynamic mix - is added to the annotations.)

We calculated scores of ESTIME and other measures for all the 1600 summaries, and presented their correlations with the average expert scores in Table 1. The measures in Table 1 are split into the group of reference-free measures (top) and the measures requiring human-written references (bottom). All the measures are based on certain principles rather than on finetuning on some human-annotated datasets. Here BLANC-help (Vasilyev et al., 2020a) is calculated in two versions[4], which differ by the underlying models: BLU - bert-large-uncased, and AXXL - albert-xxlarge-v2. ESTIME and Jensen-Shannon (Louis and Nenkova, 2009) values are negated. SummaQA (Scialom et al., 2019) is represented by SummaQA-P (prob) and SummaQA-F1 (F1 score)[5]. SUPERT (Gao et al., 2020) is calculated as single-doc with 20 reference sentences 'top20'[6] (using bert-large-nli-stsb-mean-tokens). BLEU (Papineni et al., 2002) is calculated with NLTK. BERTScore (Zhang et al., 2020) (by default [7] using roberta-large) is represented by F1, precision (P) and recall (R). For ROUGE (Lin, 2004) the ROUGE-L is calculated as rougeLsum[8].

By design ESTIME should perform well for consistency, and indeed it beats other measures in the table. Being a one-sided summary-to-text estimator

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

[4] https://github.com/PrimerAI/blanc#blanc-on-summeval-dataset

[5] https://github.com/recitalAI/summa-qa

[6] https://github.com/yg211/acl20-ref-free-eval

[7] https://github.com/Tiiiger/bert_score

[8] https://github.com/google-research/google-research/tree/master/rouge

of inconsistencies, ESTIME should not and does not perform well for relevance. ESTIME performs better than other measures for fluency, and reasonably well for coherence (ESTIME-21 is better for coherence than the rest of the reference-free metrics). Interestingly, a comparison of ESTIME-12 vs ESTIME-24 shows that the middle of the transformer knows better than the top about all the summary qualities except the fluency. In Appendix A we show and discuss a curious pattern of dependency of correlations on the embeddings layer. Correlations with all qualities peak around the layer 21, then sharply drop by the top layer 24. This is the reason we added ESTIME-21 to the table. While ESTIME-21 is the best choice, each of the three shown ESTIME versions is better than the rest of the measures for consistency and fluency.

3.2 System level correlations

measure	ρ	τ
BLANC-AXXL	0.812	0.617
BLANC-BLU	0.724	0.567
(-)ESTIME-12	0.756	0.583
(-)ESTIME-21	0.821	0.633
(-)ESTIME-24	0.815	0.633
(-)J-Shannon	0.753	0.533
SummaQA-F1	0.862	0.667
SummaQA-P	**0.912**	**0.750**
SUPERT	0.832	0.633
BERTscore-F1	-0.029	-0.017
BERTscore-P	-0.329	-0.217
BERTscore-R	0.885	0.733
BLEU	-0.150	-0.017
ROUGE-L	0.376	0.283
ROUGE-1	0.694	0.517
ROUGE-2	0.779	0.600
ROUGE-3	0.888	0.717

Table 2: System level correlation ρ (Spearman) and τ (Kendall Tau-c) of quality estimators with human experts scores of consistency. Top rows show reference-free evaluation measures.

In Table 2 we show correlations on system level, meaning that the scores (of automated measures and of human experts) are averaged over the 100 texts, so that each array of scores has length only 16 rather than 1600 (Fabbri et al., 2020). The purpose of this would be a comparison of the summarization models. The results are shown for consistency only; for other qualities some measures have p-value higher than 0.05. The ranking of the measures changes with averaging over the texts (Table 2 vs Table 1). We may speculate that some measures may be more sensitive to the model generation style which can lead to less errors or more errors on average; other measures may be more sensitive to specific factual errors in each summary. If it is true, we would have to be cautious about the measures that do well on the system level and do not do well on the summary level.

3.3 Discussion

While we must be cautious about picking evaluation measure version most fitting human scores (Vasilyev and Bohannon, 2021), using ESTIME-21 is probably justified by simultaneous maximum at level 21 for all four summary qualities, as shown in Figures 1 and 2 in Appendix A. For our definition of ESTIME we preferred N_a of Equation 1 rather than the alternative definition N_w of Equation 3 in Appendix B. N_w is counting all the text tokens that managed to 'win', i.e. to be closer to a summary token than any text occurrence of the summary token. We are concerned that if the summary token is bad (inconsistent with its context), the number of the 'winners' is large and might be fairly arbitrary.

We defined ESTIME in Equations 1 and 2 by using simple scalar product of embeddings. In Appendix C we show that using normalized embeddings only makes the correlations worse, by almost fully erasing the 'Layer-21 maximum'.

In Appendix D we give an example of switching to a simpler underlying model: bert-base-uncased. This slightly weakens the correlations, and makes the dependency on the layer id less sharp.

In Appendix E we give an example of excluding part of speech tokens from consideration by ESTIME. This means that the summation in Equation 1 will use only tokens t_i of some parts of speech, and that the $\max$ will similarly restrict the tokens t'_β. Despite high frequency of determiners in texts, the omission of the determiners from ESTIME makes almost no difference.

As explained in Section 2, in obtaining embeddings we are using a somewhat spoiled context, because we mask many tokens in a single input (albeit requiring the masks to be reasonably separated). In Appendix F we show that our separation requirements are indeed reasonable, and making them twice more strict barely change the correlations.

4 Performance on human summaries with generated subtle errors

Machine-generated summaries, even by abstractive summarization models, generally follow the source text by frequently reproducing large spans from it. Human summaries are more varied in describing the source text, and it is interesting how useful can be ESTIME for evaluating them. Fundamentally, we are asking how flexible are the embeddings in understanding the context. In order to answer this question, we made random selection of 2000 text-summary pairs from CNN/Daily Mail dataset (Hermann et al., 2015). For each human-written summary we then added the same summary modified by generated factual errors. We thus made 4000 text-summary pairs. We assigned the 'golden' scores as 1 to each clean summary, and 0 to each summary with errors.

Our 'subtle errors' generation method is simple, heuristic-free and easily reproducible. In order to generate an error, we randomly select a whole-word token in the summary, mask and predict it by an LM model (we used bert-base-cased). We then select the top predicted candidate that is not equal to the real token, and substitute it for the real token. The resulting *subtle errors* are similar to real machine-generated mishaps and hallucinations, with the fluency preserved.

The evaluation task is now more difficult: the summaries are human-written, and the errors are subtle. Without labeling of the generated errors, we cannot be confident of always having real factual errors: large part of the generated errors are indeed truly factual errors, but the rest disturb coherence or fluency, or make synonyms. For purposes of a preliminary simple evaluation here, and to ensure high probability of having true errors, we generated 3 random errors in each 'score=0' summary. Table 3 shows that ESTIME is more sensitive to the generated errors than other measures. Only reference-free measures could be applied in this situation. All p-values in the table are less than 10^{-3}, except 0.023 for BLANC-AXXL, 0.002 for Jensen-Shannon and 0.001 for SummaQA-P.

In Table 3 the correlation of ESTIME-21 with generated errors turns out to be lower than the correlation of ESTIME-24. If we guessed correctly in Apendix A about the reasons for the drop of the correlations between the layers 21 and 24 in Figures 1 and 2, then the relatively high value of ESTIME-24 indicates that it may have additionally benefited from an information relevant to predicting tokens, even when the generated token replacements are not factual errors. In the near future we plan to follow up these evaluations on a large fully labeled dataset of 'subtle errors'.

measure	ρ	τ
BLANC-AXXL	0.036	0.042
BLANC-BLU	0.076	0.087
(-)ESTIME-12	0.138	0.159
(-)ESTIME-21	0.163	0.188
(-)ESTIME-24	**0.169**	**0.195**
(-)J-Shannon	0.048	0.055
SummaQA-F1	0.055	0.064
SummaQA-P	0.054	0.062
SUPERT	0.107	0.123

Table 3: Correlation ρ (Spearman) and τ (Kendall Tau-c) of quality estimators with the presence of generated subtle errors in human summary. The dataset of 4000 text-summary pairs was created by random pick of 2000 test-summary pairs from CNN / Daily Mail dataset, duplicating these 2000 pairs, and by generating subtle errors in the 2000 duplicated summaries.

5 Conclusion

We introduced ESTIME: estimator of summary-to-text inconsistency by mismatched embeddings, - a measure of summary quality with emphasis on measuring factual inconsistency between the summary and the text. The fact that this simple measure correlates with human-labeled consistency and fluency much better than more complex measures tells us about the current state of summary evaluation, and about the power of contextual embeddings.

We also introduced a method for generating *subtle errors*; the method has a potential for creating consistent and realistic benchmark datasets for factual consistency. In the near future we intend to release such fully labeled flexible dataset.

Acknowledgments

We thank Nidhi Vyas and anonymous reviewers for review of the paper and valuable feedback.

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 Meth-*

ods in Natural Language Processing, pages 9347–9359. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6251–6258. Association for Computational Linguistics (2020).

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *arXiv*, arXiv:2010.00490.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

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. Association for Computational Linguistics.

Nicholas Egan, Oleg Vasilyev, and John Bohannon. 2021. Play the Shannon game with language models: A human-free approach to summary evaluation. *arXiv*, arXiv:2103.10918.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65. Association for Computational Linguistics (Hong Kong, China, 2019).

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*, pages 3938–3948. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. SummEval: Re-evaluating summarization evaluation. *arXiv*, arXiv:2007.12626v4.

Lisa Fan, Dong Yu, and Lu Wang. 2018. Robust neural abstractive summarization systems and evaluation against adversarial information. *arXiv*, arXiv:1810.06065.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure! A meta evaluation of factuality in summarization. *arXiv*, arXiv:2010.12834.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 446–469. 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*, pages 9332–9346. Association for Computational Linguistics.

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

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306–314. Association for Computational Linguistics.

Klaus-Michael Lux, Maya Sappelli, and Martha Larson. 2020. Truth or error? Towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, pages 1–10. Association for Computational Linguistics (2020).

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. Association for Computational Linguistics (2020).

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 (ACL), pages 311–318, Philadelphia. 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*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Oleg Vasilyev and John Bohannon. 2021. Is human scoring the best criteria for summary evaluation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2184–2191. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020a. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, Nicholas Egan, Charlene Chambliss, and John Bohannon. 2020b. Sensitivity of BLANC to human-scored qualities of text summaries. *arXiv*, arXiv:2010.06716.

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 s 5008–5020. Association for Computational Linguistics (2020).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics (2020).

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a BERT-based summary quality estimation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6005–6011, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. *arXiv*, arXiv:1904.09675v3.

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*, pages 563–578. Association for Computational Linguistics.

A Dependency on layers

It is natural to expect that embeddings from top layer would be good in characterizing context for a token. In Figures 1 and 2 we show correlations of SummEval expert scores with ESTIME versions that are defined by a model layer from which the embeddings are taken. (The model is bert-large-uncased-whole-word-masking.) Immediate observation about the dependency of the correlation value on the model layer is that after reaching maximum around layer 21, the correlation value quickly drops at higher layers. At low levels the correlation value increases fast by layer 5 (for coherence and relevance or 7 (for consistency and fluency) and then grows much slower, sometimes going flat or even dropping down.

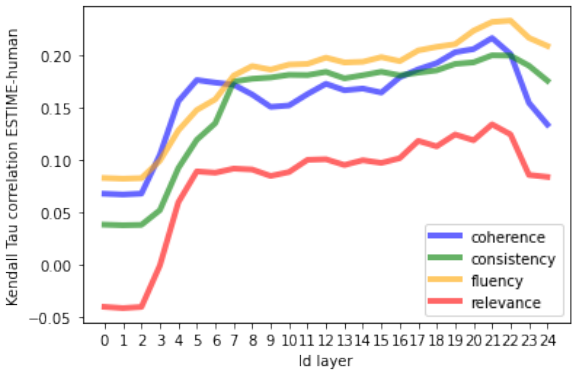

Figure 1: Kandall Tau-c correlation between SummEval experts scores and ESTIME using embeddings taken from different layers of the model.

We have no guess why the dependency of the correlations on the layer Id is so strong immediately after the layer #2 and why it is weak further in the wide range of the middle layers. However, we can speculate about the sharp drop after the 'layer #21 peak'. It may be that below the layer #21 peak, the BERT model keeps a lot of generic contextual information for two reasons: it is trained for two tasks (next sentence prediction and masked token prediction), and each node has to be useful for all or for the most of the nodes above. But after

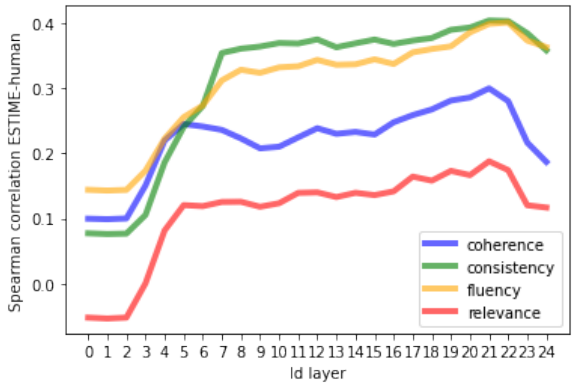

Figure 2: Spearman correlation between SummEval experts scores and ESTIME using embeddings taken from different layers of the model.

the peak, the last few layers at positions close to each text token are strongly influenced by the token prediction task.

In the plots shown in this Appendix, as well as in all other plots through the paper, the correlations p-values are below 0.05 (mostly far below). Unlike the summary level correlations, the system level correlations have not much data. This is why, keeping only the correlations with p-values below 0.05, we can show in Figure 3 only the consistency quality, and even for the consistency we have less range of layers.

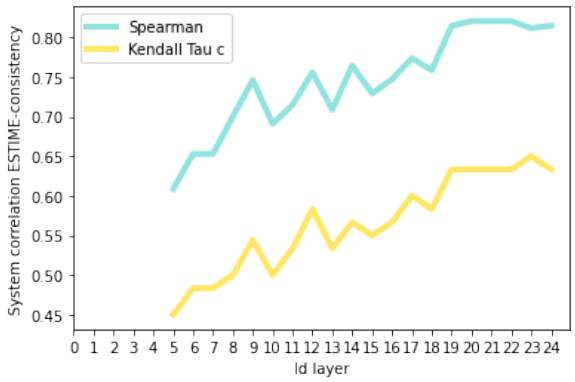

Figure 3: Spearman and Kendall Tau-c correlations - system level - between SummEval experts scores of consistency and ESTIME using embeddings taken from different layers of the model.

B Count of winning tokens in text

ESTIME is defined in Equation 1 and is considered through the paper as a count of 'alarming' summary tokens. It could be alternatively defined as a count of all winner-tokens from the text, as defined in Equation 3.

$$N_w = \sum_{i:t_i \in T} \sum_{\beta:t'_\beta \neq t_i} H((e_i e'_\beta) - s(i)) \quad (3)$$

In Figures 4 and 5 we see how N_w differs from ESTIME (N_a).

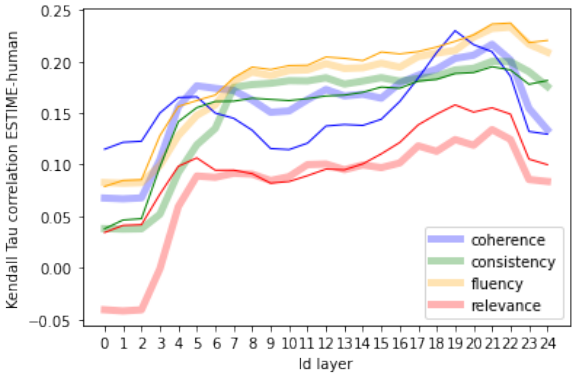

Figure 4: Kandall Tau-c correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: ESTIME (as defined by Equation 1 and considered through the paper). Thin lines: N_w as defined by Equation 3.

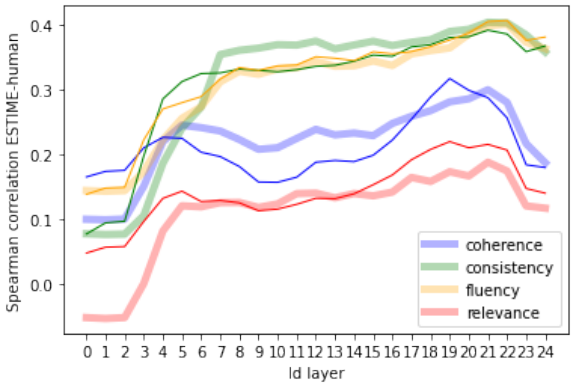

Figure 5: Spearman correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: ESTIME (as defined by Equation 1 and considered through the paper). Thin lines: N_w as defined by Equation 3.

C Normalization of embeddings

We used unnormalized embeddings for ESTIME. From Figures 6 and 7 it is clear that normalizing embeddings does not improve ESTIME. Curiously, the effect of the normalization on the correlations is in destroying the peak around the layer 21.

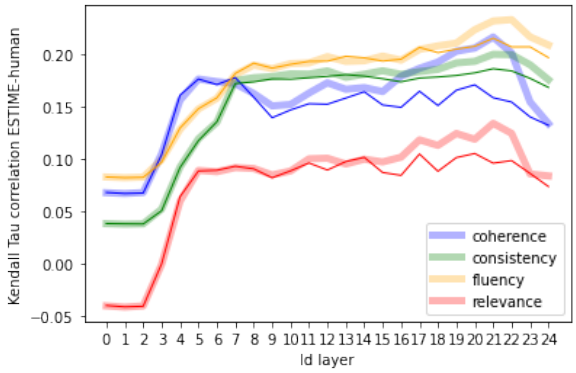

Figure 6: Kandall Tau-c correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: unnormalized embeddings (as used through the paper). Thin lines: normalized embeddings.

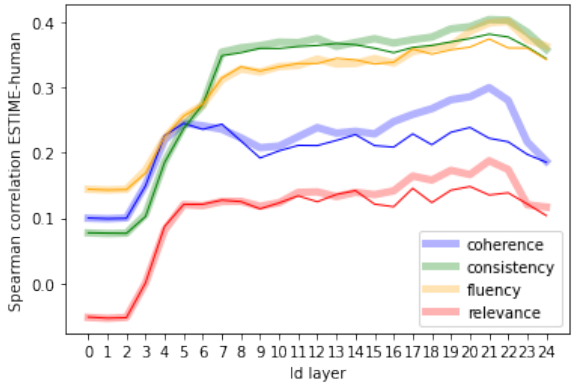

Figure 7: Spearman correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: unnormalized embeddings (as used through the paper). Thin lines: normalized embeddings.

This could mean that the lengths of the embeddings carry all the information necessary for creating the peak at the layer 21.

D Comparison with base BERT

As default and through the paper ESTIME uses the model bert-large-uncased-whole-word-masking. In Figures 8 and 9 we show an example of a comparison with another model - bert-base-uncased. Unlike the large model, the bert-base-uncased has 0-12 range of its layers, and in the plots here we rescaled them by x2, interpolating in between for odd layer Ids. This allows to compare the trends of the correlations along the relative depth of the transformer. We observe the familiar quick rise at low depth, a drop at high levels close to the output, and a slow growth or plateau in between, - but all these features are less sharp for the bert-base-uncased. It is puzzling that a larger transformer, with twice longer 'distance' for backpropagation to travel from the top to the bottom, has more distinct features of quick rise, plateau, peak and drop, with exact locations of the end of the quick rise and of the peak.

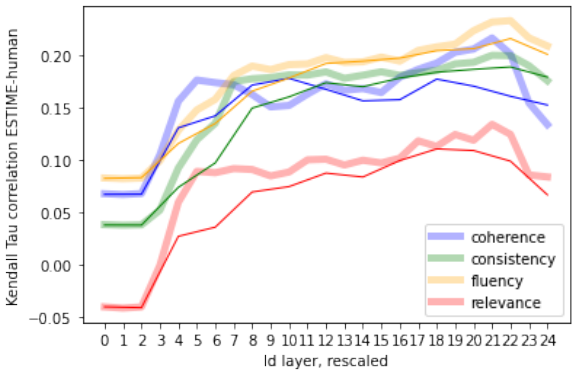

Figure 8: Kandall Tau-c correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: the model is bert-large-uncased-whole-word-masking. Thin lines: the model is bert-base-uncased.

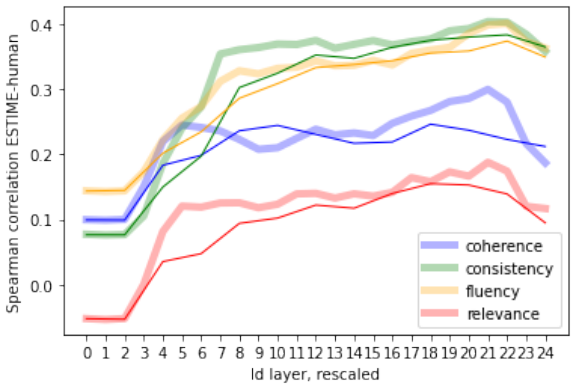

Figure 9: Spearman correlation between SummEval experts scores and EESTIME by embeddings from different layers of the model. Thick lines: the model is bert-large-uncased-whole-word-masking. Thin lines: the model is bert-base-uncased.

E Example of excluding a part of speech

Throughout the paper we used all text tokens for ESTIME. In Figures 10 and 11 we show an example of excluding from consideration a part of speech: determiners. Determiners occur very frequently in the text, but exclusion of them does not make much difference in the resulting correlations with human scores, especially for the quality we are most interested in: consistency.

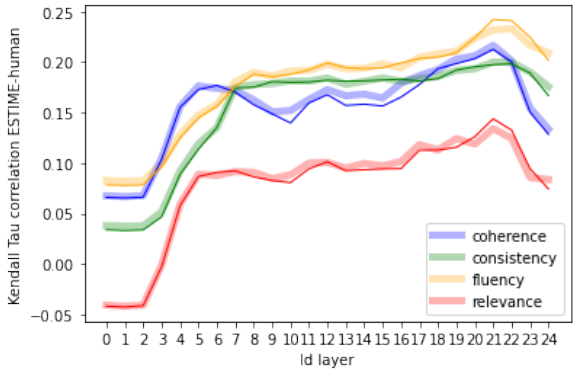

Figure 10: Kandall Tau-c correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: all tokens are used, as is done throughout the paper. Thin lines: tokens of determiners (part of speech) are not used.

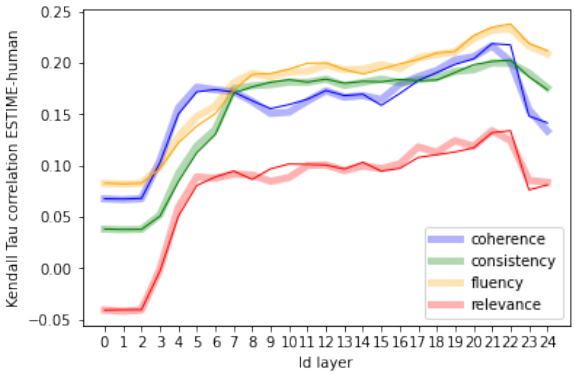

Figure 12: Kandall Tau-c correlation between SummEval experts scores and ESTIME by embeddings from different layers of the model. Thick lines: sparsity of the masking is defined by the distance 8 and the margin 50 (see Section 2), as used through the paper. Thin lines: Distance 8, margin 100.

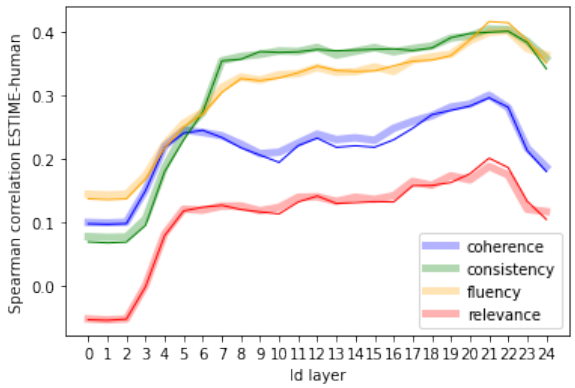

Figure 11: Spearman correlation between SummEval experts scores and EESTIME by embeddings from different layers of the model. Thick lines: all tokens are used, as is done throughout the paper. Thin lines: tokens of determiners (part of speech) are not used.

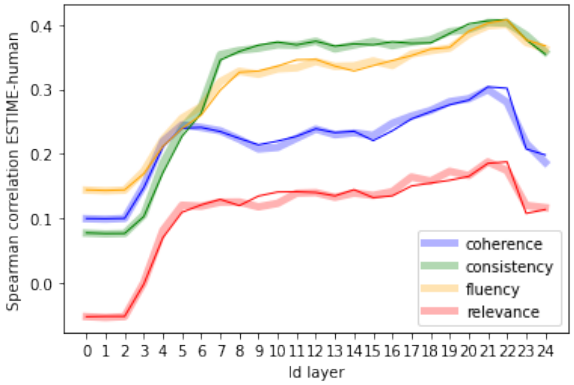

Figure 13: Spearman correlation between SummEval experts scores and EESTIME by embeddings from different layers of the model. Thick lines: sparsity of the masking is defined by the distance 8 and the margin 50 (see Section 2), as used through the paper. Thin lines: Distance 8, margin 100.

F Parameters for sparse masking

In Section 2 we explained that for faster processing we take embeddings not one at a time, but as much as fit into an input window, as long as the masking is done with 8 tokens separation, and within the margin 50 tokens from the input edges (unless input edge touches the edge of the text). In Figures 12 and 13 we compare our default parameters with a twice more sparse version: 16 tokens separation, and 100 tokens margin. The sparser version should be better, but slower to run. From the figures it is clear that the sparser version has almost the same correlations; our default sparsity is good enough.