Is Human Scoring the Best Criteria for Summary Evaluation?

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

Abstract

A summary quality measure is judged by how well it correlates with quality scores produced by human annotators. A higher correlation with human scores is considered to be a decisive indicator of a better measure. In this work we present observations that cast doubt on this view. We also show a possibility of an alternative indicator for selecting the best measure from a family of measures, a criterion that does not rely on human scores.

1 Introduction

The goal of summarization is to convey important and only important information of the text in a fluent, comprehensible and concise summary, preserving the factual consistency with the text.

There are several families of automated measures of summary quality. For example, Gabriel et al. (2020) classified the automated measures into four types: question-answering, text reconstruction, semantic similarity and lexical overlap. Each of these types has families of measures, for example SUM-QE (Xenouleas et al., 2019), APES (Eyal et al., 2019), Summa-QA (Scialom et al., 2019), QAEval (Deutsch et al., 2020) and FEQA (Durmus et al., 2020) in question-answering, BLANC (Vasilyev et al., 2020a) in text reconstruction, BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019) and SUPERT (Gao et al., 2020) in semantic similarity, ROUGE (Lin, 2004) and Jensen-Shannon (Louis and Nenkova, 2009) in lexical overlap.

A high correlation with human evaluation scores is currently accepted as the crucial criterion for choosing a good evaluation measure. Arguably, the factual faithfulness can be annotated objectively, with detailed classification of factual errors (Kryscinski et al., 2020; Huang et al., 2020; Vasilyev et al., 2020b; Gabriel et al., 2020). However, other summary qualities are subjective; this forces researchers to be careful in design and usage of human annotations (Bhandari et al., 2020; Fabbri et al., 2020; Iskender et al., 2021). Annotation scores depend on the types of texts and on the qualification of annotators. For example, there is a big difference in expert and crowd-sourced scores in (Fabbri et al., 2020)[1].

Annotators are biased in favor of anything that makes the scoring easier: extractiveness of the summary, and focus of the summary on the top part of the document (Ziegler et al., 2020). The annotation process itself differs from how the summary quality is assessed by a typical human reader. A human reader does not have a goal of scoring a summary, but rather uses the summary to guess the content of the text.

The contribution of this work:

1. We provide an example of a false 'improvement' of an automated evaluation measure: a dubious modification, imitating a human annotator behavior, can increase the correlation with human scores. For a contrast, we also provide an example of a true improvement that increases correlation with human scores for a good reason.
2. We explore an alternative criterion for selecting an optimal evaluation measure from a family of measures, the criterion not relying on human scores. We provide evidence that the criterion is robust across different kinds of texts and summaries.

For our demonstration we will use BLANC family of evaluation measures, because it is easily interpretable as an analogy to a human reader that uses the summary to guess the content of the text. Two families defined in (Vasilyev et al., 2020a) differ

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

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2184–2191
August 1–6, 2021. ©2021 Association for Computational Linguistics

by their setup. The BLANC-help family gets information from the summary by having the model read the summary when reading and reconstructing the text. The BLANC-tune family gets information from the summary by lightly tuning the model on the summary before reading and reconstructing the text. Measures in each of the families, BLANC-help and BLANC-tune, may differ by the parameters defining the setup.

2 Example of False Improvement: Limited Comparison with Text

After reading a summary, an annotator may chose not to review carefully the whole text, but to consider in detail only the parts that look most similar to the summary. We can imitate this by using only the part of the text that is most related to the given summary. In modifying BLANC this way, it is reasonable to expect that correlation with human scores will increase, but this would make a false 'improvement' of the measure.

In Appendix we provide two examples - Figures 6 and 7 - illustrating the bias that we seek to explore in this section. Each example has a summary that truly attempts to cover all important facts, and a summary that we intentionally wrote to cover only a very limited part of text. To create a falsely 'improved' measure, we seek to explore the bias of annotator giving more attention to parts of text most similar to the summary.

To create a biased BLANC, we can calculate BLANC separately for each sentence of the text, and select n sentences with the highest score. We can consider these selected sentences as the 'text' to deal with, and calculate BLANC on this 'text'. We create such *limited-text* BLANC from BLANC-help[2]. For our illustration we use average expert scores of 1600 text-summary pairs in the dataset SummEval (Fabbri et al., 2020).

Compared to BLANC, the limited text BLANC has indeed higher Spearman correlation with average expert, as shown by thin lines in Figure 1. In this and other figures through this section all p-values are below 0.05.

For Appendix examples, see results in Tables 1 and 2.

We can imagine a human expert paying more attention to several (say three or five) most 'promising' (most similar to the summary) sentences of the text. In evaluating relevance, this might be not very different from working with full text. But for other qualities (coherence, consistency, fluency) the correlation increases.

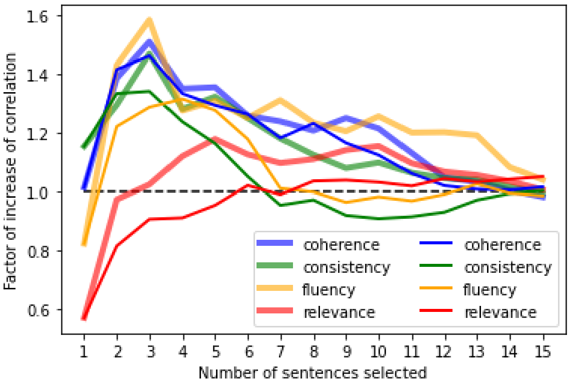


Figure 1: Factor by which Spearman correlation of BLANC with human scores increases when only part of text is used for BLANC. The text part is selected as sentences with top BLANC values (thin lines) or as contiguous sentences with highest BLANC (thick lines).

Naturally, for a human it is easier to review a contiguous piece of text rather than separated pieces, even if this might diminish legitimacy of evaluation of all qualities, including relevance. And, no surprise, BLANC for such contiguous part of text correlates with human scores even better - as shown by thick lines in Figure 1.

Figure 2 illustrates the same trends when the resulting BLANC is calculated for each selected sentence separately, and then averaged over the sentences.

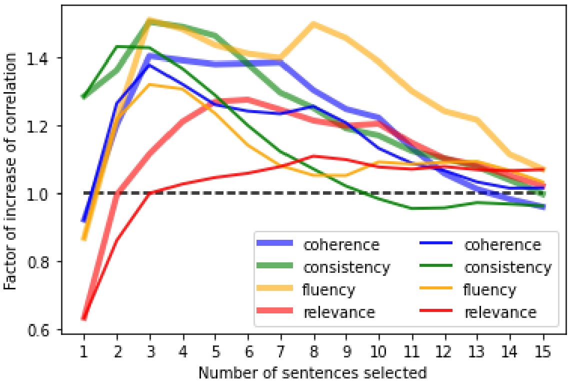


Figure 2: Factor by which Spearman correlation of BLANC with human scores increases when only part of text is used for BLANC. The text part is selected as sentences with top BLANC values (thin lines) or as contiguous sentences having highest average BLANC (thick lines). The resulting BLANC is calculated as average over BLANC of the sentences.

Figure 3 shows the increase of correlations when

[2] https://github.com/PrimerAI/blanc

the text is restricted not by the number of sentences but by a threshold on BLANC of a sentence.

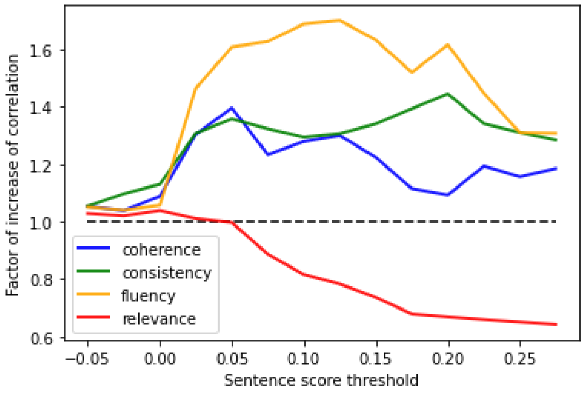


Figure 3: Factor by which Spearman correlation of BLANC with human scores increases when only part of text is used for BLANC. The text part is selected as sentences with BLANC exceeding threshold.

Selection of a part of the text is used in SUPERT multi-document evaluation measure (Gao et al., 2020) as a tool for creating 'reference' from each document and then evaluating a summary on the created references. In the context of BLANC here, the selection of a part of the text is done differently and has a clear interpretation: instead of estimating usefulness of the summary in guessing the whole text, we estimate how much the summary would help to guess only the most 'relevant' part of the text. Here 'relevant' means the part of the text for which the summary turned out to be most helpful. This is equivalent to using only the most easy (for annotator, after reading the summary) part of the text. The summary may as well relate only to a small piece of text of no importance. This means that the evaluation measure became worse, even though the correlation with human scores is stronger.

The human bias exploited by the limited-text BLANC does not necessarily manifest itself through a low inter-annotator agreement. The reported in SummEval (Fabbri et al., 2020) inter-annotator agreement of experts is 0.71, which is an acceptable value (Artstein and Poesio, 2008). While achieving a reasonable inter-annotator agreement is an important problem in human annotations, our example shows that another problem may be in the nature of the human evaluation of summary qualities, where a summary is presented to human for scoring (rather than for guessing about the text content), and the text is presented to facilitate the scoring.

3 Example of True Improvement: Learning More from Summary

In this section we provide an example of a legitimate increase of correlations with human scores, as opposed to the described in previous section false 'improvement'. We can combine BLANC-help and BLANC-tune by tuning the model on the summary (BLANC-tune), and then using the tuned model to read the summary while doing Cloze task on the text (BLANC-help).

Such *full* BLANC version is equivalent to a human that first learns the summary, and then, while guessing missed words in the text, is still looking at the summary again and again. Using both opportunities to learn from the summary makes sense, it should legitimately extract more help from the summary. The worst that may happen is that a model used by BLANC-help is already so perfect in reading the summary that its additional tuning on the summary will not improve the measure (but will not hurt either).

As expected, the full BLANC has substantially higher correlations with annotations of experts on the 1600 text-summary pairs of (Fabbri et al., 2020). Compared to BLANC-help, the Spearman correlation of full BLANC with human scores can increase by 13%-18% for coherence, 2%-3% for consistency,13%-15% for fluency, and 7%-8% for relevance, - all this depending on the number of epochs (10-20) and learning rate (1.0e-4 to 2.0e-4) of the BLANC-tune used.

4 Max-Help Criterion and its robustness

As we have seen in section 2, the correlation with human scores is not always a reliable method to select the best evaluation measure. The fact that we were able to recognize the falsity of 'improvement' in section 2 and the legitimacy of improvement in section 3 suggests that we may find a no-human criterion, at least for some setups.

In previous sections we used BLANC-help as an initial version for our modifications. As stated in Vasilyev et al. (2020b), based on the dataset introduced in there[3], BLANC-help with interval $gap = 2$ between masking locations in the text provided the highest correlations with human scores. It was noted that BLANC's average score across

[3] https://github.com/PrimerAI/blanc

the dataset was also the highest at this setup, implying that BLANC extracted maximal help from the summaries. Such coincidence is not a rule: the ”max-help” measure, selected for having maximal average score, is not always the same as the ”max-human” measure, selected for having maximal correlation with human scores.

The max-help criterion - selection of a measure that has highest average score - makes sense under two conditions:

1. The measure is being selected from a family of measures that have the same definition of the output score - as assessment of a usefulness of a summary. The score may be derived, for example, from how many text tokens were successfully reconstructed with help of the summary (BLANC), or from how many questions about the text were successfully answered with help of the summary (QA-based measures). The condition is that the definition is fixed for the family.
2. The average score is being measured with a large enough dataset representing the domain on which we are interested to use the measure.

The meaning of the criterion is simple: the better is the measure in extracting useful information from summaries, the better it should be in judging summaries by their usefulness. The criterion does not require human scoring. All we need is a measurement of an average score.

The max-help criterion can be credible if it does not depend too strongly on the types of texts and summaries. In order to verify this assumption, we considered four types of summaries (and the corresponding texts): (1) CNN summaries from the CNN / Daily Mail dataset (Hermann et al., 2015); (2) Daily Mail summaries from the CNN / Daily Mail dataset; (3) First two sentences from random daily news; (4) Random two sentences from random daily news.

The random daily news were selected as three random news documents per day over one year, with the summaries of the document being two first and two random sentences. We used 1000 samples for each of the four types of summaries.

We intentionally selected summaries of so different styles and quality: if the criterion selects the same best measure for so different types of summaries, then it is indeed a very robust criterion.

For BLANC-help family, we found that for all four datasets the optimal setup (accordingly to the max-help criterion) happens to be the same: $gap = 2$; minimal length of whole-word token allowed to be masked is 6 characters $L_{normal} = 6$; the word-split tokens are always masked ($L_{lead} = 1$ for first token, and $L_{follow} = 1$ for follow-up tokens).

This setup is almost the same as the parameters found in (Vasilyev et al., 2020b) to maximise correlation with human scores, except L_{normal} and L_{follow} which have low influence. The question we asked: does the ’optimal’ max-help evaluation measure remain optimal (or near-optimal) for different kinds of texts and summaries? Figure 4 provides convincing evidence for a positive answer.

Figure 4 shows the average BLANC-help value obtained with sub-optimal (different from max-help) setup. We consider a change of gap and gap_mask (number of tokens allowed to be masked at each masking location) to explore a less frequent and a more frequent masking, and a change in the token length thresholds for masking tokens. Remarkably, the average BLANC-help value drops in each case for all four datasets in a similar manner. The token length thresholds have almost no influence, making a drop just a few percents. Change in frequency of masking has a larger effect, leading to a drop 10%-20%.

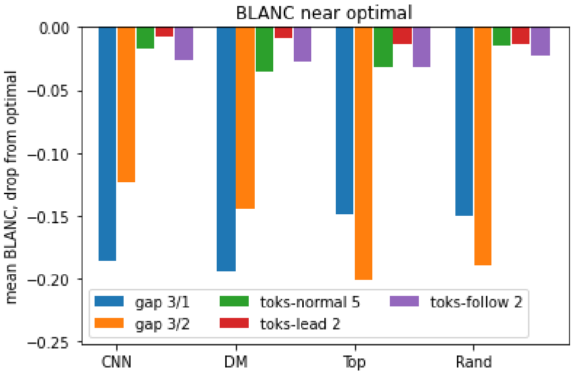


Figure 4: Drop of mean BLANC-help value when parameters differ from optimal. The drop is shown as a fraction of the optimal mean BLANC value. The summaries probed are: CNN and DM (from the CNN/Daily Mail dataset), Top and Rand (first two sentences and random two sentences from random news articles). The parameters probed are: ’gap 3/1’ is $gap = 3$ and $gap\_mask = 1$; ’gap 3/2’ is $gap = 3$ and $gap\_mask = 2$; ’toks-normal 5’ is $L_{normal} = 5$; ’toks-lead 2’ is $L_{lead} = 2$; ’toks-follow 2’ is $L_{follow} = 2$.

For BLANC-tune family, similar to BLANC-help, the max-help optimal setup is the same for all four datasets: $gap = 3$; number of tokens allowed to be masked at each masking location for inference $gap\_mask = 2$; for tuning $gap_{tune} = 4$ and $gap\_mask_{tune} = 3$; $L_{normal} = 6$; $L_{lead} = 1$; $L_{follow} = 1$; probability of replacement of a masked token by another random token at tuning $p_{replace} = 0$.

Probability $p_{replace} = 0$ differs from the value 0.1 used in the standard BERT training, but $p_{replace}$ has only weak influence on the BLANC-tune. Figure 5 shows a few examples of changes of the setup, which illustrate that the optimal measure remains optimal across all four datasets.

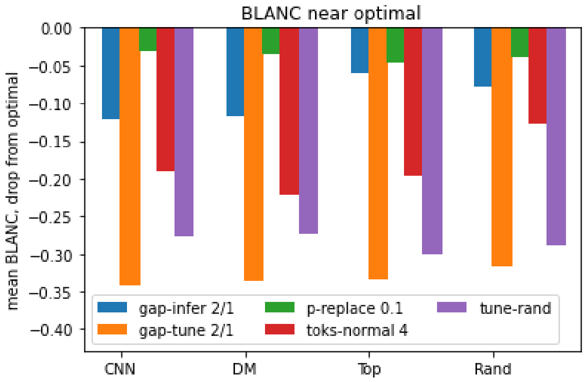


Figure 5: Drop of mean BLANC-tune value when parameters differ from optimal. The drop is shown as a fraction of the optimal mean BLANC value. The summaries probed are: CNN and DM (from the CNN/Daily Mail dataset), Top and Rand (first two sentences and random two sentences from random news articles). The parameters probed are: 'gap-infer 2/1' is $gap = 2$ and $gap\_mask = 1$; 'gap-tune 2/1' is $gap_{tune} = 2$ and $gap\_mask_{tune} = 1$; 'p-replace 0.1' is $p_{replace} = 0.1$; 'toks-normal 4' is $L_{normal} = 4$; 'tune-rand' is making tokens masking random rather than even at tuning.

The demonstrated evidence for robustness suggests that in finding an optimal measure we do not need even human summaries: we can apply the max-help criterion utilizing random sentences from the texts.

5 Conclusion

In this paper, we critically reviewed the assumption that maximal correlation with human scores defines the best evaluation measure for summarization; we provided observations supporting our scepticism. Using good interpretability of BLANC evaluation measure, we provided examples of both illegitimate 'improvement' and legitimate improvement of the correlation of BLANC scores with human scores.

We stated the motivation for an alternative criterion for choosing an optimal summary evaluation measure: the maximal average extracted usefulness of summary. We provided evidence that the criterion is robust across very different kinds of summaries, including such 'summaries' as first sentences or random sentences of the text. This means that the criterion can be applied without the need of human summaries.

While in this work we used BLANC, we think that similar observations and the same conclusions could be made using a question-answering based evaluation measure.

Acknowledgments

We thank Nidhi Vyas and Delenn Chin for review of the paper and valuable feedback.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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*, pages 9347–9359. Association for Computational Linguistics.

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.

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.

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.

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*, volume 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.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96. Association for Computational Linguistics (2021).

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.

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. 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.

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. 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.

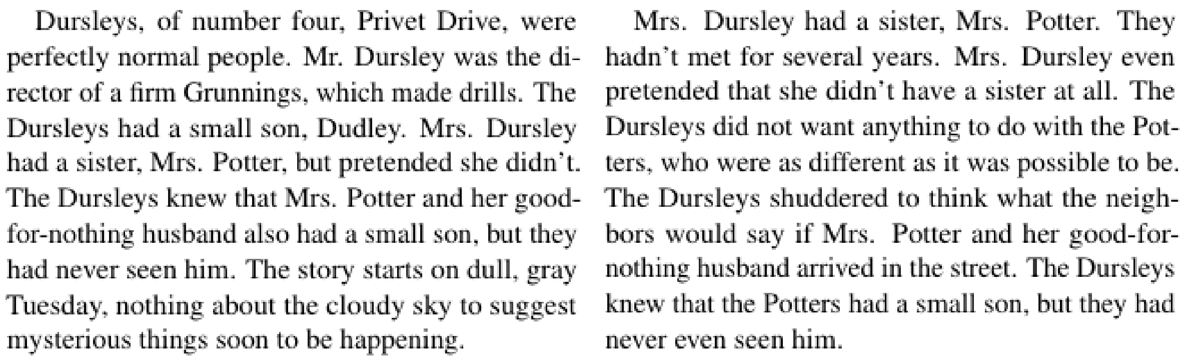
Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *arXiv*, arXiv:1909.08593v2.

A Examples of bad judgement by falsely ’improved’ measure

In Figure 6 we provide two toy summaries, each summary is supposed to summarize the first four paragraphs of ’Harry Potter’ book [4]. In order to illustrate how the false improvement described in Section 2 affects ranking of summaries, we intentionally wrote one summary with ’wide coverage’ (attempting to cover all the most important facts), and another summary with ’narrow coverage’ (focusing on a limited part of the text).

The original BLANC, judging the summaries by how useful they are in predicting tokens of the text, gives higher score to the wide-coverage summary, - see Table 1. The falsely ’improved’ BLANC - limited-text BLANC of section 2 - gives higher score to the narrow-coverage summary. It does not matter whether the small part of text covered by the second summary is important or not, the limited-text BLANC makes judgement by how well the summary covered that part of the text. In this case,

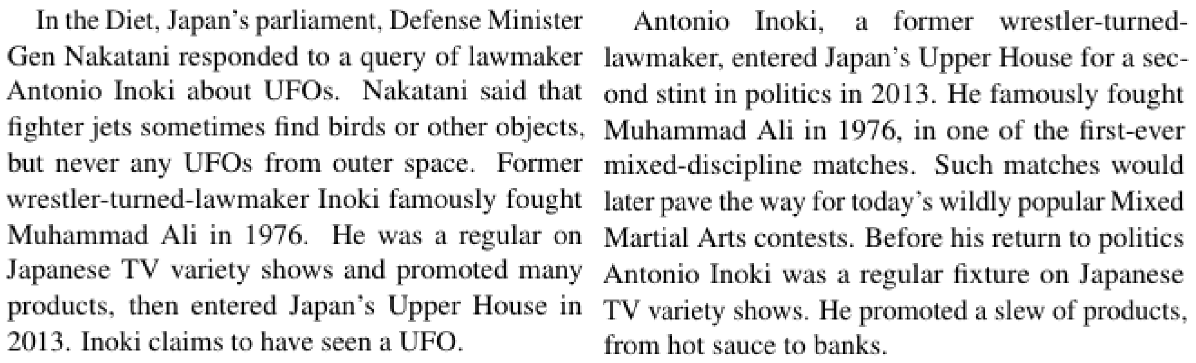
[4] J.K.Rowling (1997). *Harry Potter And the Sorcerer’s Stone*. Bloomsbury



Dursleys, of number four, Privet Drive, were perfectly normal people. Mr. Dursley was the director of a firm Grunnings, which made drills. The Dursleys had a small son, Dudley. Mrs. Dursley had a sister, Mrs. Potter, but pretended she didn't. The Dursleys knew that Mrs. Potter and her good-for-nothing husband also had a small son, but they had never seen him. The story starts on dull, gray Tuesday, nothing about the cloudy sky to suggest mysterious things soon to be happening.

Mrs. Dursley had a sister, Mrs. Potter. They hadn't met for several years. Mrs. Dursley even pretended that she didn't have a sister at all. The Dursleys did not want anything to do with the Potters, who were as different as it was possible to be. The Dursleys shuddered to think what the neighbors would say if Mrs. Potter and her good-for-nothing husband arrived in the street. The Dursleys knew that the Potters had a small son, but they had never even seen him.

Figure 6: Example of a summary with a wide coverage (left) and a narrow coverage (right). Both summaries are supposed to cover first four paragraphs of 'Harry Potter And the Sorcerer's Stone' by J.K.Rowling.



In the Diet, Japan's parliament, Defense Minister Gen Nakatani responded to a query of lawmaker Antonio Inoki about UFOs. Nakatani said that fighter jets sometimes find birds or other objects, but never any UFOs from outer space. Former wrestler-turned-lawmaker Inoki famously fought Muhammad Ali in 1976. He was a regular on Japanese TV variety shows and promoted many products, then entered Japan's Upper House in 2013. Inoki claims to have seen a UFO.

Antonio Inoki, a former wrestler-turned-lawmaker, entered Japan's Upper House for a second stint in politics in 2013. He famously fought Muhammad Ali in 1976, in one of the first-ever mixed-discipline matches. Such matches would later pave the way for today's wildly popular Mixed Martial Arts contests. Before his return to politics Antonio Inoki was a regular fixture on Japanese TV variety shows. He promoted a slew of products, from hot sauce to banks.

Figure 7: Example of a summary with a wide coverage (left) and a narrow coverage (right). Both summaries are supposed to cover the same text taken from CNN/Daily Mail dataset. The text is shown in Figure 8.

BLANC	wide	narrow
Original	0.244	0.218
Limited n=4	0.539	0.584
Limited n=3	0.563	0.652
Limited n=2	0.594	0.778

Table 1: Scores of BLANC versions for wide and narrow coverage summaries of Figure 6. Top row is the original BLANC. Lower rows are for falsely 'improved' BLANC with selection of n top text sentences, as described in Section 2.

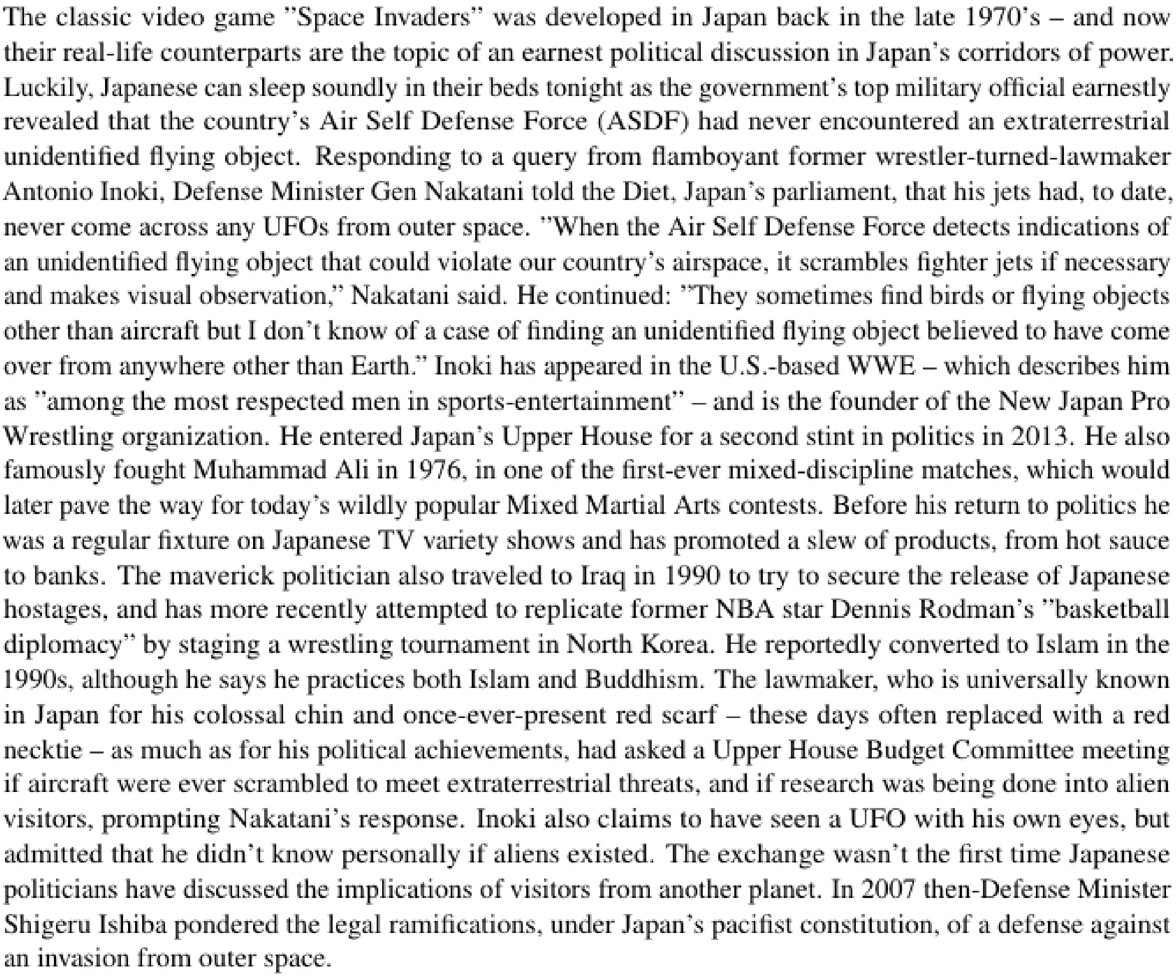
BLANC	wide	narrow
Original	0.159	0.134
Limited n=4	0.327	0.433
Limited n=3	0.365	0.516
Limited n=2	0.428	0.536

Table 2: Scores of BLANC versions for wide and narrow coverage summaries of Figure 7. Top row is the original BLANC. Lower rows are for falsely 'improved' BLANC with selection of n top text sentences, as described in Section 2.

the second summary concentrates on the attitude of the Dursleys to the Potters, and does not provide any other information.

In Figure 7 we again provide two summaries, this time we wrote them for the text shown in Figure 8, taken from SummEval dataset (Fabbri et al., 2020). Again, as shown in Table 2, the original BLANC gives higher score to the overall more useful wide-coverage summary. The falsely 'improved' limited-text BLANC gives higher score to the narrow-coverage summary which focuses only on information about Antonio Inoki and ignores his query in the parliament.

The scores by the limited-text BLANC are higher than the scores by the original BLANC. The reason is that only text sentences with highest scores are selected, and the less the number of the sentences, the higher is the average score. Naturally, for the wide-coverage summary the score increase is not as great as for the narrow-coverage summary.



The classic video game ”Space Invaders” was developed in Japan back in the late 1970’s – and now their real-life counterparts are the topic of an earnest political discussion in Japan’s corridors of power. Luckily, Japanese can sleep soundly in their beds tonight as the government’s top military official earnestly revealed that the country’s Air Self Defense Force (ASDF) had never encountered an extraterrestrial unidentified flying object. Responding to a query from flamboyant former wrestler-turned-lawmaker Antonio Inoki, Defense Minister Gen Nakatani told the Diet, Japan’s parliament, that his jets had, to date, never come across any UFOs from outer space. ”When the Air Self Defense Force detects indications of an unidentified flying object that could violate our country’s airspace, it scrambles fighter jets if necessary and makes visual observation,” Nakatani said. He continued: ”They sometimes find birds or flying objects other than aircraft but I don’t know of a case of finding an unidentified flying object believed to have come over from anywhere other than Earth.” Inoki has appeared in the U.S.-based WWE – which describes him as ”among the most respected men in sports-entertainment” – and is the founder of the New Japan Pro Wrestling organization. He entered Japan’s Upper House for a second stint in politics in 2013. He also famously fought Muhammad Ali in 1976, in one of the first-ever mixed-discipline matches, which would later pave the way for today’s wildly popular Mixed Martial Arts contests. Before his return to politics he was a regular fixture on Japanese TV variety shows and has promoted a slew of products, from hot sauce to banks. The maverick politician also traveled to Iraq in 1990 to try to secure the release of Japanese hostages, and has more recently attempted to replicate former NBA star Dennis Rodman’s ”basketball diplomacy” by staging a wrestling tournament in North Korea. He reportedly converted to Islam in the 1990s, although he says he practices both Islam and Buddhism. The lawmaker, who is universally known in Japan for his colossal chin and once-ever-present red scarf – these days often replaced with a red necktie – as much as for his political achievements, had asked a Upper House Budget Committee meeting if aircraft were ever scrambled to meet extraterrestrial threats, and if research was being done into alien visitors, prompting Nakatani’s response. Inoki also claims to have seen a UFO with his own eyes, but admitted that he didn’t know personally if aliens existed. The exchange wasn’t the first time Japanese politicians have discussed the implications of visitors from another planet. In 2007 then-Defense Minister Shigeru Ishiba pondered the legal ramifications, under Japan’s pacifist constitution, of a defense against an invasion from outer space.

Figure 8: Example of text from SummEval dataset.