Fill in the BLANC: Human-free quality estimation of document summaries

Zhe HUANG
Linghao ZENG
Wenyi ZHANG
Ngoc Trinh Hung NGUYEN

January 26, 2024

1 Introduction

In our report, we embark on a comprehensive exploration of the paper "Fill in the BLANC: Human-free quality estimation of document summaries. [1]" Our initial step is to carefully analyze the algorithms described in the paper, ensuring we fully understand them. Next, we plan to re-implement the code that the authors of the paper used. The third step involves initially replicating some of the experiments already presented in the paper. Additionally, we will try to come up with new ideas based on what we have learned. Finally, we will wrap up our report with a summary of the algorithm, discussing its effectiveness and potential future uses.

2 Understanding the Algorithm: Fill in the BLANC

BLANC is a novel method for automatically evaluating the quality of document summaries. Unlike other methods such as ROUGE, BLANC does not require human-generated reference summaries.

At its core, BLANC assesses the quality of a summary by measuring how well it aids a pre-trained model in performing tasks on a document. Specifically, we focus on the masked token task, where a model attempts to reconstruct obscured text spans.

There are two versions of BLANC: BLANC-help and BLANC-tune.

BLANC-help: This approach determines the quality of a summary by combining it, or a filler of the same length, with a sentence from the document that has masked tokens. This composite is then processed by a model to assess the accuracy of predicting the masked tokens. The variance in prediction accuracy serves as an indicator of the summary's quality.

BLANC-tune: In this method, a sentence with masked tokens from the document is input into both the original model and a model tuned with the summary. The quality of the summary is inferred by comparing their prediction accuracies for the masked portions.

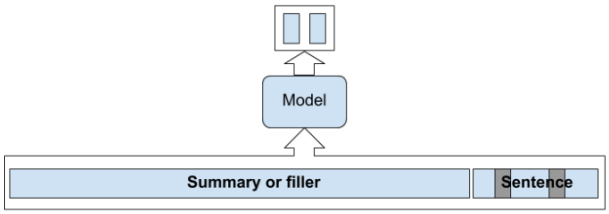

Figure 1: Schematic Diagram of the BLANC-help Method

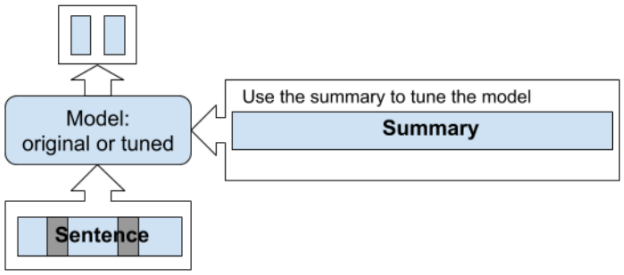

Figure 2: Schematic Diagram of the BLANC-tune Method

3 Code Re-implementation

The implementation of this part is in the '**BLANC_reimplementation.ipynb**' file.

We have reimplemented BLANC-help and BLANC-tune following the pseudo-code provided in the paper. Our goal was to understand the algorithm's behavior and assess its limitations. During our analysis, we made interesting observations and conducted tests using specific examples.

A case study of BLANC-help:

- Document: "Jack drove his minivan to the bazaar to purchase milk and honey for his large family."
- Summary: "Jack bought milk and honey."
- Score: 0.1

We observed the following:

- Original Word: mini
- Base Prediction: white
- Prediction with the help of summary: small

In this case, the term "mini" is related to "small", indicating a synonym relationship that was not considered in scoring. This suggests the potential integration of a method to calculate semantic similarity to enhance scoring accuracy.

- Summary: "Jack drove his car."
- Score: 0.2

- Original Word: jack
- Base Prediction: he
- Prediction with the help of summary: jack

Despite having a higher score, this summary is less informative, indicating a limitation in the BLANC-help's current scoring system. It bypasses the no-copy-pair guard, with the score increase primarily due to the prediction alignment between "jack" and "he" to "jack", which is an unknown aspect to us and may related to the nature of language models.

Moreover, the BLANC-help score computed using the original author's Python library was 0.22. This variance could be attributed to adjustments in key parameters within the library, particularly the masking frequency (M) and the minimum word length (L_{min}) for masking.

4 Validation Experiments of BLANC measurement

The implementation of this part is in the '**Validation Experiments.ipynb**' file.

To evaluate the effectiveness of the BLANC summary quality assessment method, we conducted four experiments mentioned in the article. We used the CNN_DailyMail_555 dataset, which s offered by the author. This dataset is composed of 555 text-summary pairs, including 100 texts with human-generated summaries randomly selected from the CNN&Daily Mail dataset (Hermann et al., 2015) and 455 machine-generated summaries. Each summary in the dataset is accompanied by a single human score, reflecting the summary's overall quality. These experiments below are glossed over in the third&forth section of the article.

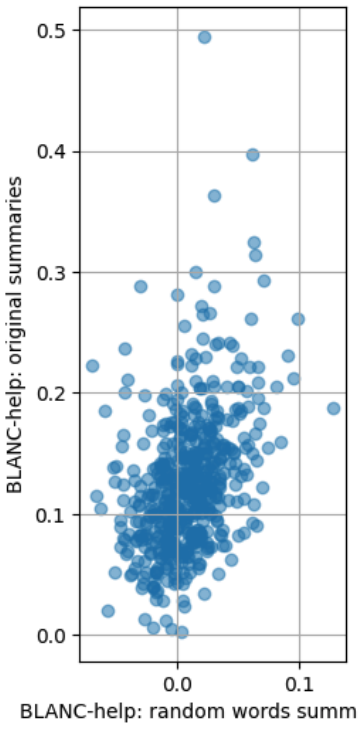

Figure 3: BLANC-help of a generated summary vs. random-words summary (left)

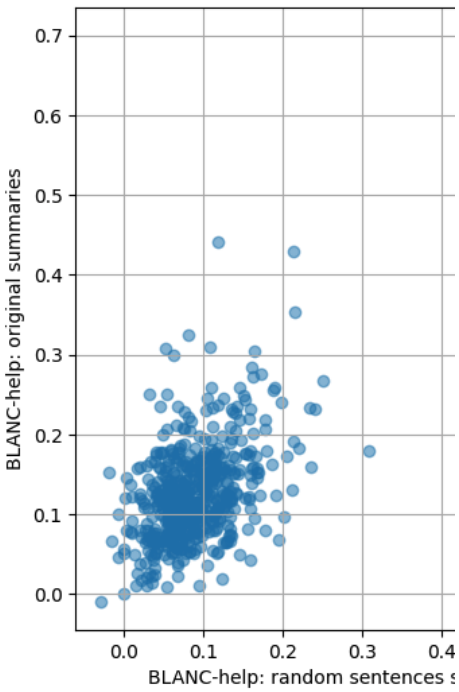

Figure 4: BLANC-help of a generated summary vs. random-sentences "summary" (right)

4.1 Random Words Summary Experiment

The first experiment is Random Words Summary Experiment. Here, our objective was to assess how a poorly constructed summary impacts the BLANC evaluation metric. We first generated summaries by randomly selecting words from the original text, ensuring that their length of words was equal to that of the original summaries. We then employed the BLANC method to evaluate these random words summaries and the original summaries. The BLANC scores obtained for these summaries were compared with the scores of the original human-generated summaries(in Figure 3). In this way, we measured the effectiveness of BLANC in distinguishing between well-constructed and poorly-constructed summaries.

4.2 Random Sentences Summary Experiment

The second Experiment aimed to further test the robustness of the BLANC summary quality assessment method. In this experiment, similarly, summaries were created by randomly selecting complete sentences from the text. And we also compared BLANC scores of the random sentences summaries with that of the original summaries(in Figure 4).

However, the author points out a problem with this experiment: in the case of purely extractive summaries, the process of calculating BLANC scores may pair a summary with sentences from the text that have been copied into the summary. This exact sentence copying should be unfairly helpful in unmasking words in the original sentence. This effect may be reduced by including a simple guard rule, "no copy-pair", into the measure: We may exclude any pairing of exact copy sentences from the calculation of the measure. In detail, when generating new random sentence summaries, every time we select a sentence randomly from the text, we delete this sentence in the original text.

4.3 Random Sentence Replacement Experiment

In this experiment, we sought to investigate how the BLANC-help metric responds when certain components of a summary are intentionally degraded. Specifically, we only focused on summaries composed of exactly three

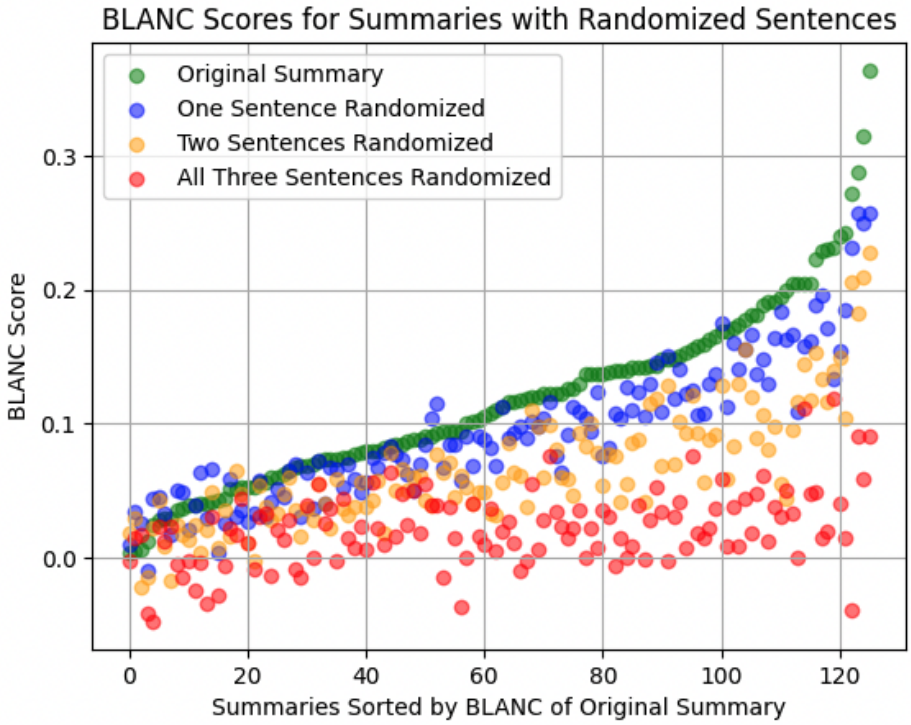

Figure 5: BLANC-help for 3-sentence summaries with one or more sentences replaced by random words from the text. The summaries are sorted by measure of the original summary.

sentences, as this allowed for a controlled alteration of the summary structure. We systematically replaced one, two, or all three sentences with randomly selected words, keeping the same length of the resulting randomized summary as the original summary.

The results, depicted in Figure 5, clearly demonstrate a trend where the BLANC-help scores decrease as more sentences are replaced by random words.

4.4 Comparison with Human Evaluation

The forth experiment focused on comparing the BLANC assessment with human evaluations. The objective was to validate the correlation between the BLANC-help scores and the scores given by human evaluators. For this purpose, we employed BLANC on human-evaluated summaries and calculated Pearson correlation between BLANC scores and human scores to determine the degree of alignment between machine and human assessments of summary quality.

Pearson correlation	**BLANC-help**	**BLANC-tune**
human scores	0.37305	0.35197

5 Innovation, Idea Generation and Implementation

5.1 Innovation and Idea Generation

Cons for Human-free model: Human-free aspect is a strong and a weak point at the same time.

In some cases, the most meaningful summary (in terms of capturing the meaning of the document), is not always the best summary under user's eyes.

The more adapted way to estimate a summary's quality (maybe) is:

- Categorize the given summary and/or take the target users estimations on the summary as a co-estimation.

Cons for BLANC-tune:

Poorly written summaries for tuning the model are not recommended: they are expensive, time-consuming, and may potentially harm the model. (See more in 6.3.2)

Instead, we should explore a simpler approach by using BLANC to select high-quality summaries before fine-tuning the original model.

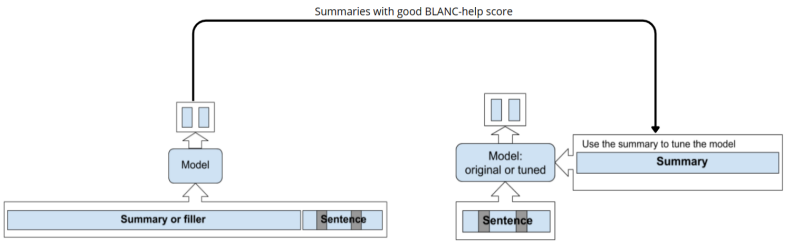

Figure 6: Schematic Diagram of Combined BLANC-help and BLANC-tune

BLANC Score Calculation Problem:

The BLANC score is very low, which is a critical observation when testing BLANC with various documents and their associated summaries.

One of the reasons we have identified is the calculation of the BLANC score:

$k = int(out_{base[i]} == sentence[i])$

$m = int(out_{help[i]} == sentence[i])$

$S_{km} += 1$

A case study in **Code Reimplementation** 3 has shown that, even the prediction is very similar to the original word (e.g. **small** vs **mini**), BLANC still recognize it as an unsuccessful prediction.

$\Rightarrow$ BLANC score index is not reliable in many cases.

Idea and Solution:

Making the BLANC-score calculation more flexible is a commendable approach. Utilizing cosine similarity provides an effective solution for this calculation.

The cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is given by:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

We then set a threshold to decide if that is a successful prediction or an unsuccessful case.

Remark 1: The idea of using similarity between the predicted tokens and original tokens seems promising. However, for models with transformer architecture, the connections between tokens are calculated and optimized based on Keys, Values, and Queries. Therefore, if we use tokens' embeddings or positional embeddings (which are not recommended because the predicted tokens don't have a position in the sequence) to calculate similarity between tokens, it might not yield the results we expect.

Remark 2: BLANC doesn't have a pre-built evaluation metric like ROUGE or BLEU, so there is a real need for in-depth research on how to implement the calculation of the BLANC score. It should not solely rely on calculations proposed in papers or based on cosine similarity.

Remark 3: We might want to use the tokenizer proposed by the pre-trained model we are employing to maintain coherence, instead of employing BERT as described in the papers. However, in the experiment, I am using the BERT model for this section so that we can have a clear vision of the results of the proposed solution in the papers.

Automated Summarization and Evaluation Model:

Auto Generative Model to summarize a document then its output will pass through BLANC for the quality's estimation process.

The index to minimize will be $-BLANC = -\frac{S_{01}-S_{10}}{S_{total}}$

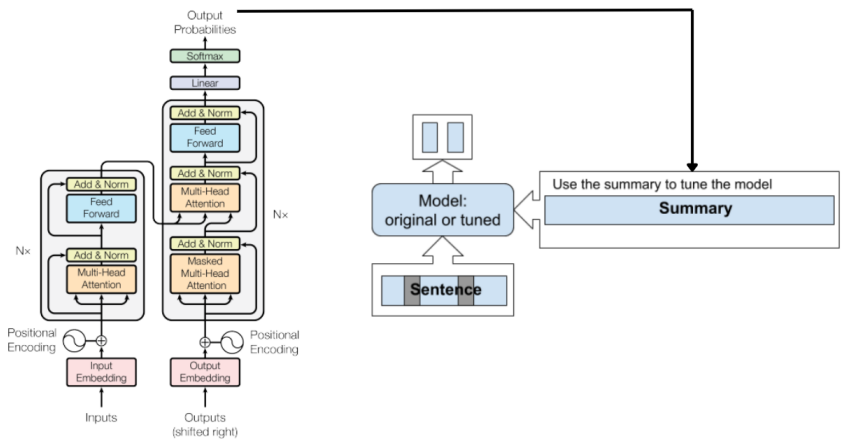

Figure 7: Schematic Diagram of Automated Summarization and Evaluation Model

Idea of Automated Summarization and Evaluation Model Implementation:

Step 1: Use Transformer with a fixed-length Decoder Output L_{summ} in order to maintain the consistency of the context and ensure a consistent length for the summary. We can train a model with documents and their associated summaries so that a model can learn better how to capture the meaning of a given document.

Step 2: Pass the summary to the tuning model process (or pass it to BLANC-help in order to find a good summary before tuning the original model).

Step 3: Use $-BLANC = -\frac{S_{01}-S_{10}}{S_{total}}$ as an index to minimize, the smaller the value, the better the summary (in terms of meaning).

(Optional): Categorize the given summary and/or take the target user's estimations on the summary as a co-estimation.

Step 4: Return good summaries and their documents associated for Decoder and Encoder process (Step 1) and re-do the whole process.

$\Rightarrow$ **Little Cons:** BLANC score is not a continuous function (similar to 0-1 Loss).

5.2 Implementation

5.2.1 Adjusted BLANC-score calculations

The implementation of this part is in the '**Automated Summarization-Evaluation.ipynb**' file.

Since the approach of using similarity scores has proven ineffective until now, I have conducted all experiments and explained their limitations in **Section** 6.3.1.

5.2.2 Automated Summarization and Evaluation Model

Model used: t5 small pre-trained model[2]

Dataset: BillSum

Summarization of US Congressional and California state bills.

Features:

- text: bill text.

- Summary: summary of the bills.

Reason for using pretrained model: I cannot build a model from scratch due to a lack of dataset and limited capacity on my Colab version for training the model. To achieve significant results in terms of BLANC score and during the tuning phase of the generative model, I opted to use and test several pretrained models. The facebook/bart-large-cnn model is not executable (causing my Colab to run out of RAM), so I chose the 't5-small' model instead.

Implementation: In the '**Automated Summarization-Evaluation.ipynb**' file.

Result: Summaries from fine-tuned model is general better than summaries generated by pre-trained model

Case study:

Truncated reference summary:

Existing law authorizes a peace officer to arrest a person without a warrant if the officer has probable cause to believe that the person has committed a public offense in the officer's presence or if the officer has probable cause to believe that the person has committed a felony. This

Pre-trained model's summary:

a peace officer may arrest a person in obedience to a warrant, warrant or, pursuant to the authority granted to him or her by Chapter 4.5. the officer has probable cause to believe that the person to be arrested has committed a felony, although not in the officer's presence.

Fine-tuned model's summary - 2 epochs:

Existing law authorizes a peace officer to arrest a person in obedience to a warrant, warrant, or, pursuant to the authority granted to him or her by Chapter 4.5, without a warrant, to arrest a person warrant whenever any of the following circumstances occur: (1) The officer has probable cause to believe that the person to be arrested has committed a public offense

Even without seeing the original document, we can observe that the fine-tuned model's summary is more detailed and meaningful compared to the pre-trained model alone. This has been confirmed by ChatGPT and me.

Note: All details about truncated original text, results after certains epochs will be store in the "**Summarization-Evaluation-CaseStudy.pdf**" file.

About BLANC-score:

We **applied no-copy-pair guard** to eliminate sentences from the text that have been copied into the summary.

We **don't apply Adjusted BLANC-score calculations** due to their lack of reliability.

The result with filler consisting of a number of "." equal to the length of the provided summaries :

Truncated reference summary score = 0.16

Pre-trained summary score = 0.15

2 Epochs trained summary score = 0.22

6 Epochs trained summary score = 0.19

The result with filler consisting of a number of "." equal to the number of tokens of the provided summaries. :

Truncated reference summary score = 0.14

Pre-trained summary score = 0.17

2 Epochs trained summary score = 0.22

6 Epochs trained summary score = 0.14

We can clearly observe the instability of the BLANC score across various given summaries, with different lengths of filler.

According to ChatGPT, the Truncated reference summary is more comprehensive compared to pre-trained summary, yet the BLANC score suggests the opposite.

In contrast, ChatGPT indicates that the summary trained for 2 Epochs is more comprehensive compared to the one trained for 6 Epochs, as indicated by the BLANC score.

Indeed, the way we generate the **filler** and its length will affect the BLANC score, through the unmasking process.

Conclusion: Combine with the result from **Code Re-implementation** 3, there is a need for deeper research into all aspects of BLANC. Thus far, despite decent scores across various summaries, the results of BLANC score calculations, BLANC-tune structure, and others have not been entirely reliable.

6 Conclusion on Limitations, Future Prospects, Remaining Problems, Solutions, and Contributions Section

6.1 Limitations of Fill in the BLANC

As we've mentioned in many places in this report, **Fill in the BLANC** is an interesting approach for human-free quality estimation of document summaries. However, under our observations, BLANC still has some limitations.

- The effect of **filler** on a model's prediction capability.
- BLANC score calculation
- BLANC-tune limitations
- Human-free approach

Note: Since we have already discussed certain limitations and solutions in **Section** 5, either through observations or logical reasoning, we are just revisiting these limitations here.

6.2 Future Prospects

We can hardly make any definitive statements about the future prospects of BLANC.

While BLANC is a promising concept with an interesting approach, it lacks many essential components in various aspects, which require considerable work to address.

In comparison to ROUGE or BLEU, which are widely used and already have their own prebuilt matrices, we believe BLANC still has a long way to go. This journey includes adjusting BLANC score calculations, selecting pre-trained models for BLANC score calculation, and establishing a reliable and stable matrix to evaluate the training process.

6.3 Remaining Problems and Solutions

6.3.1 Cosine Similarity for BLANC score calculations

We have built a cos similarity calculation in order to calculate the similarity between a predicted token and masked token.

Model: bert-base

There are some case studies:

token1 = "man" token2 = "man" Cosine Similarity: 0.9999997019767761

token1 = "man" token2 = "boy" Cosine Similarity: 0.9007138609886169

token1 = "cat" token2 = "dog" Cosine Similarity: 0.9107443690299988

⇒ As mentioned in the Remarks of Section 5.1, adjusting the BLANC score is necessary to make a fair judgment on the predicted tokens, but it requires further research. We observed that setting a threshold to determine the quality of predictions for masked tokens is challenging: the cosine similarity between 'man' and 'boy' is lower than that between 'cat' and 'dog'.

This challenge is particularly pronounced when using base models like BERT, making the results not reliable at all.

Further research is indeed necessary in order to make BLANC a reliable evaluation matrix, like ROUGE or BLEU.

6.3.2 Automated Summarization and Evaluation Model

Limited max-length for pre-trained model:

The max-length is typically fixed for pre-trained models; for example, it's 1024 for 'facebook/bart-large-cnn' and T5 small. This limitation is imposed to prevent information loss, as large max-length values may lead to issues such as vanishing/exploding gradients. Additionally, this constraint is influenced by computational limitations.

Solution: Chunking long documents into smaller ones is a useful and widely-used solution. [3]

Cons: Tokens at the barrier (of every 1024 tokens for the models mentioned above) might lose their information.

Note 1: Our first idea was to cut a very long document into paragraphs (which is much smaller compared to max-length), but the result didn't meet my expectations. With this approach, we very likely lose the connection between paragraphs, thereby losing all the power of the Transformer model. The Transformer model is designed to address the issue of losing connection between two tokens when the distance between them is very far in the document, which is a problem with the LSTM model

Note 2: We also posted this question to M. Tristan Cazenave [4]. He also proposed this method and mentioned that losing information at the barrier is a situation that must be accepted.

We have already tested this technique with some lengthy documents, and the generated summaries are decent. Unfortunately, Colab has a hard time fine-tuning a dataset with extra-long documents (always crash). Therefore, We have limited the max-length to 512 and reduced the max-length of the summary accordingly.

We also want to test the BLANC-score between truncated reference summaries and summaries generated by a pre-trained model, but it seems impossible due to a lack of computational capability.

Reference Summary Problem:

The trick above introduced a new problem: the order in the document vs. the order in the summary.

Because the summaries in the dataset, used as references for fine-tuning the model, are written by humans, so that the order in the document is sometimes not preserved in the corresponding summary. In other words, in the summary, they present a general situation before starting the actual summarization. To address this, we attempted to provide ChatGPT with some truncated summaries from the dataset and summaries generated by a pre-trained model. Pre-trained model summaries are highly recommended for their generality.

I have personally checked them, and they are truly comprehensive compared to the truncated summaries (in some cases).

Note 1: In general, the quality of truncated summaries is better than those generated by pretrained models. This is because the majority of summaries and writing styles (not just in the dataset) tend to be in a listing format i.e., the initial information in the documents will be summarized first rather than a general-to-detail style. Moreover, the general statement (if present) in the summaries typically comprises about 10-20 tokens, functioning similarly to a headline. While it has a minor impact on the initial part of the summarization process, the overall quality of the truncated summaries is acceptable.

Note 2: In general, the maximum length for input documents is 1024, and for summaries, it is 128, which is more than enough to handle almost every truncated document situation. However, in my case, we have limited the maximum length to 512 and the summary to 64, which might cause some trouble.

Fine-tuning model with bad summaries:

The problem mentioned earlier leads us to another issue: a lack of generality in the truncated summaries.

Since we will be fine-tuning the model with truncated summaries from the dataset, which sometimes do not follow the same order of information as the corresponding documents, the model might have difficulty learning how to align with the summaries in the dataset.

Without a proper fine-tuning process for the generative model, we might compromise the integrity of the pretrained model.

Case study:

Pre-trained model's summary: plastic microbeads nuisance prevention law 42360 is added to Part 3 of division 30 of the Public Resources Code. the Legislature finds and declares all of the following: a) plastic does not biodegrade into elements or compounds commonly found in nature like other organic materials. a) plastic pollution is the dominant type of anthropogenic debris found throughout the

Fine-tuned model with 2 epochs: Existing law provides for the licensure and regulation of personal care products by the State Department of Public Resources by the State Department of Public Resources. Existing law provides for the regulation and regulation of personal care products by the State Department of Public Resources. Existing law requires the State Department of Public Resources to

The fine-tuned model's result has some attributes aligned with the reference summary but lacks generality and is hard to understand due to insufficient training and the quality of truncated summary.

Solution 1: We are considering aligning the reference summaries with the same order of information as the corresponding documents. This way, when we chunk a very long document into smaller segments, we can maintain corresponding truncated summaries that align with each chunk. This alignment facilitates the fine-tuning process for the model.

Solution 2: Consider heavily condensing and slightly summarizing the document before fine-tuning it with the corresponding reference summary (we can then apply solution 1 for this phase).
While this approach may be time-consuming, it has the potential to yield a more natural summary.

Question about evalution model: Do we really need a very well-trained evaluation model?

Fine-tuning a model is required to help an evaluation model understand the text's concept in a specific domain. For example, we may want to fine-tune a model for a finance-related task to help it evaluate finance texts. But if the model is very well-trained and can predict (somehow) the correct (or very similar) masked tokens in a document/summary and an average person may struggle to understand the document or summary due to its poorly written, lacking information, etc, is that a good choice? Or should an average model (just like an average human) be used to judge the text?

We can consider Bard or ChatGPT as examples: when faced with a poorly structured question and numerous linguistic errors, the given answer is somehow correct. If we use the evaluation model as a filter before passing the summary to a human for further tasks, an extraordinary model that can unmask every token correctly might not be the best choice. Instead, an average model with knowledge comparable to that of an average human might be a better solution.

⇒ This brings us back to the problem of categorizing the type of summary.

About BLANC-tune: Tuning a model is really expensive and time-consuming, (as we can see in the fine-tuning process of the pretrained model), and its necessity and output result is still in question (remarks above). Therefore, our first approach will be with BLANC-help.

6.4 Contributions

Ngoc Trinh Hung NGUYEN: My main works are Innovation, Idea Generation and Implementation 5 and Conclusion on Limitations, Future Prospects, Remaining Problems, Solutions, and Contributions Section 6.

I also implemented the code for the *Automated Summarization-Evaluation.ipynb*, with the helpful hand of W.ZHANG in the training process, and the assistance of resources such as the Hugging Face [2] and Long Document Discussion on Hugging Face [3].

I had a close discussion with W. ZHANG on the adjustment of the BLANC score calculation. I also implemented and tested it on the BLANC-help model (no-copy-pair guard function has been implemented), using examples provided in the summarization process (stocked in Summarization-Evaluation-CaseStudy.pdf file).

Up to this point, we have gathered a substantial number of elements to construct a comprehensive Automated Summarization-Evaluation model. However, due to BLANC's instability in certain aspects, I have not integrated these elements to create an evaluation matrix based on BLANC.

Regarding BLANC-tune, due to a lack of time, resources, and certain constraints on tuning the model and BLANC-tune itself, as discussed in Section 6.3.2, I did not perform all the tests with this model.

ChatGPT and Bard also assisted me in evaluating summaries and correcting linguistic errors, provided guidance on loading data from HuggingFace, using HF-Token, and helped fix bugs in my code, my Latex's file.

Linghao ZENG: My main contribution is the reimplementation of BLANC-help and BLANC-tune in *BLANC reimplementation.ipynb*, and conducting several case studies in Section 3. I followed the pseudo-code from the original paper closely to ensure our version of BLANC-help was accurate. This reimplementation allowed us to explore how BLANC works in detail.

In our case studies, we looked at how different summaries for the same document affected the BLANC scores. This process helped us see how changes in summaries could lead to different scores. From these observations, we got an idea: it might be good to make the way BLANC calculates scores more flexible. This means the BLANC system could maybe consider the meanings of words more when it gives scores, which could make it better at evaluating summaries (Section 5).

I also explored how changing hyperparameters, such as the number of masked words and their spacing, affects BLANC scores. Although this experiment could help fine-tune BLANC's accuracy, I haven't fully analyzed the results due to time and report length constraints.

ChatGPT assisted in correcting grammatical errors in my slides and reports, and also helped in debugging my code.

Zhe HUANG: My main contribution is conducting the Validation Experiments of BLANC measurement 4. Diligently reviewing the source article, I meticulously coordinated the experimental design, implementation, and data analysis processes.

Utilizing the dataset provided by the authors on their GitHub repository, I replicated as much as possible the authors' intention in the validating experiments of the BLANC method.In response to the no-copy-pair issue raised during our last presentation discussion, I implemented a straightforward solution(which involved deleting the selected sentences from the original text) that effectively resolved the matter. Consequently, all four experiments were conducted with success, yielding results consistent with the original paper.

In addition to the empirical work, I engaged in extensive discussions with **W. ZHANG** and **LH. ZENG** regarding our findings related to the BLANC methodology and potential enhancements.

ChatGPT was used in refining the language used in my slides and report, and providing assistance in debugging the code.

Wenyi ZHANG:My contribution is centered around drafting the introduction of our report, aiding team members in understanding the paper, checking their code, and collectively resolving any issues. Additionally, I was involved in the organization and coordination of our group efforts, as well as the final review and proofreading of our report.

References

[1] O. Vasilyev, V. Dharnidharka, and J. Bohannon, *Fill in the blanc: Human-free quality estimation of document summaries*, 2020. arXiv: 2002.09836 [cs.CL].

[2] Huggingface, *Summarization*. [Online]. Available: https://huggingface.co/docs/transformers/tasks/summarization.

[3] Huggingface-Discussion, *Summarization on long documents*. [Online]. Available: https://discuss.huggingface.co/t/summarization-on-long-documents/920.

[4] T. Cazenave, *Professeur at lamsade dauphine psl*. [Online]. Available: https://www.lamsade.dauphine.fr/~cazenave/index.php.