Automatic Paper Assessment

Gyuseok Lee, Sumin Lim, Kyumin Park, Jiwoo Park
Institute of Software Research
Carnegie Mellon University
{gyuseokl, suminlim, kyuminpa, jiwoop}@andrew.cmu.edu

Abstract

With the academic evolution in many fields, a large number of scientific papers are produced over people's ability to read. To help people understand a large number of scientific papers and help authors to receive quick feedback, we develop a paper assessment tool using deep learning. We collect abstracts, short summaries (TL;DR), strengths, weakness, and their acceptance from Openreview API. Then we train the BERT model to generate either a short summary, strength, or weakness from the abstract input. At the same time, we develop an acceptance predictor from the abstract. The acceptance predictor achieves an accuracy score of 0.57. We also show from a demo that our model successfully generates a short summary, strength, or weakness from the abstract. With additional training data and a larger model, we expect our paper assessment generator can provide more useful information to both authors and readers.

1 Introduction

In line with the development of science and technology, we are living in a flood of information nowadays. We can easily feel this circumstance considering the existence of 6 million English articles in 'WIKIPEDIA', which is a popular online encyclopedia. In addition, we encounter numerous texts through diverse media such as news, Twitter, and so on every day. Moreover, as you can see in Figure 1, in 2021, around 335,000 papers were published in the field of AI alone. Because of this reality, automatic text summarization has been regarded as a very important field in real life and many studies have been conducted [3, 7, 4].

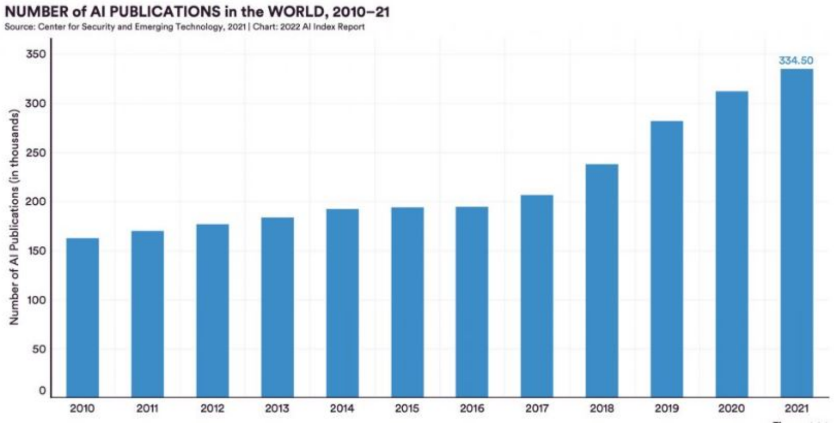


Figure 1: Number of AI Publications in the world, 2010-2021

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

However, only a few pieces of research have been done to summarize scientific research papers [5, 11, 1, 6]. In light of the fact that tremendous papers are pouring into the AI field every day, this is a very unfortunate situation. So, our goal is to make a well-performing automatic paper summarization model. Before we define our problems to solve, we introduce our assumption. We assumed that the abstract of each paper would contain most of the information needed to evaluate the paper. This is because a Language model usually has a limitation on the length of input words and because the abstract is actually a summary of the paper. With this assumption, the problems that we want to solve are 4 tasks using the abstract. First, we want to summarize the paper only in one sentence. Second, we want to generate strengths. Third, we want to generate weakness. Lastly, we want to determine whether it is accepted or not. In fact, there has been no attempt to summarize the paper only in one sentence. Furthermore, there has been no attempt to make a model that produces the strengths and weaknesses of a scientific research paper to the best knowledge. In addition, there has been no attempt to predict whether it would be accepted or not. Therefore, we can say our approach is a novel approach. Also, one can easily realize that it is very helpful if we can get the strengths and weaknesses of a paper for another research. Therefore, we aim to make the model output not only a one-sentence-summarization but also strengths and weaknesses, whether it is accepted or not.

Although we have good motivations and an idea for this project, we run into some difficulties since there is no attempt for these purposes. However, our contribution is that this is the first attempt to output a summary in one sentence, weaknesses, strengths, and whether it would be accepted or not simultaneously, only using the abstract.

2 Related Works

Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary [8]
The motivation of this paper is how to measure the quality of treatment well for patient discharge summaries Due to including the various aspects of patient information in the summary, feature extraction technique from corresponding the domain is important. This paper proposes a novel sentiment analysis method for discharge summaries like vector space models, statistical models, association rules, and extreme learning machine autoencoder (ELM-AE). We think that these proposed feature extraction techniques would be very important for our task even if the domain is different from each other. So, we expect to make fine-grained feature representations for our task.

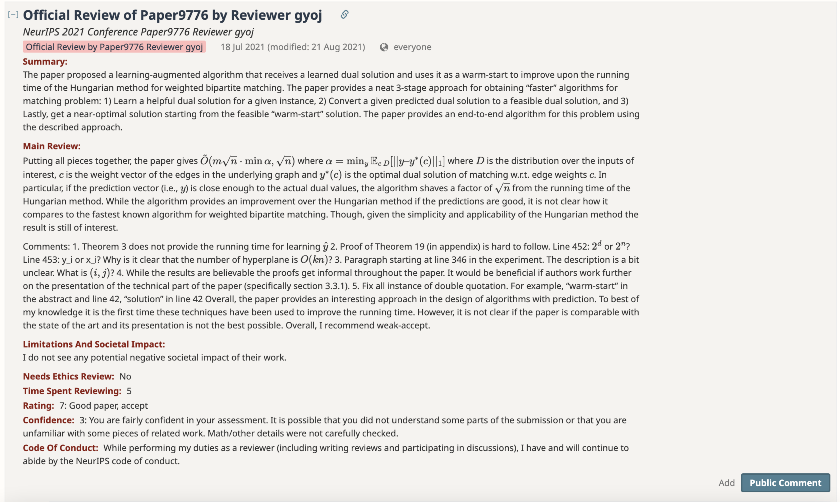
A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning [9]
This paper is designed for text summarization which is one of the core tasks in natural language processing. The main idea of this paper is a hybrid model of extractive-abstractive to combine BERT (Bidirectional Encoder Representations) word embedding with reinforcement learning. Specifically, this paper used BERT word embedding as text representation and pre-train two sub-models (i.e., extractive and abstractive). After that, the extraction network and the abstraction network are bridged by reinforcement learning. Plus, they use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics as the evaluation method. We think that feature extraction and text summarization is related to our task, so this paper would be very useful for the overall structure of our task.

3 Dataset

To achieve our research goal, we mainly focus on `OpenReview.net` data. Multiple assigned reviewers leave their review contents publicly in `OpenReview.net`. All papers submitted to a focal conference has their multiple reviews. Examples of papers' reviews are presented in Figure 2 and Figure 3.

This task has an inherent difficulty in that the true quality of a paper is determined by other researchers and journal editors. Each conference has its own research scope and aims, resulting in the rejection of a high-quality and out-of-scope paper. A similar problem also occurs in the reviewers' aspect. Each reviewer has their own research area, so a focal reviewer may not notice the high-quality paper because of different research interests.

To construct our dataset, we crawled the website, 'openreview.net'. This website keeps track of several conferences in the computer science area. However, there were problems in the crawled



[−] **Official Review of Paper9776 by Reviewer gyoj**

NeurIPS 2021 Conference Paper9776 Reviewer gyoj

Official Review by Paper9776 Reviewer gyoj 18 Jul 2021 (modified: 21 Aug 2021) everyone

Summary:
The paper proposed a learning-augmented algorithm that receives a learned dual solution and uses it as a warm-start to improve upon the running time of the Hungarian method for weighted bipartite matching. The paper provides a neat 3-stage approach for obtaining "faster" algorithms for matching problem: 1) Learn a helpful dual solution for a given instance, 2) Convert a given predicted dual solution to a feasible dual solution, and 3) Lastly, get a near-optimal solution starting from the feasible "warm-start" solution. The paper provides an end-to-end algorithm for this problem using the described approach.

Main Review:
Putting all pieces together, the paper gives $\tilde{O}(m\sqrt{n} \cdot \min \alpha, \sqrt{n})$ where $\alpha = \min_y \mathbb{E}_{c\ D}[||y–y^*(c)||_1]$ where D is the distribution over the inputs of interest, c is the weight vector of the edges in the underlying graph and $y^*(c)$ is the optimal dual solution of matching w.r.t. edge weights c. In particular, if the prediction vector (i.e., y) is close enough to the actual dual values, the algorithm shaves a factor of $\sqrt{n}$ from the running time of the Hungarian method. While the algorithm provides an improvement over the Hungarian method if the predictions are good, it is not clear how it compares to the fastest known algorithm for weighted bipartite matching. Though, given the simplicity and applicability of the Hungarian method the result is still of interest.

Comments: 1. Theorem 3 does not provide the running time for learning $\hat{y}$ 2. Proof of Theorem 19 (in appendix) is hard to follow. Line 452: 2^d or 2^n? Line 453: y_i or x_i? Why is it clear that the number of hyperplane is $O(kn)$? 3. Paragraph starting at line 346 in the experiment. The description is a bit unclear. What is (i, j)? 4. While the results are believable the proofs get informal throughout the paper. It would be beneficial if authors work further on the presentation of the technical part of the paper (specifically section 3.3.1). 5. Fix all instance of double quotation. For example, "warm-start" in the abstract and line 42, "solution" in line 42 Overall, the paper provides an interesting approach in the design of algorithms with prediction. To best of my knowledge it is the first time these techniques have been used to improve the running time. However, it is not clear if the paper is comparable with the state of the art and its presentation is not the best possible. Overall, I recommend weak-accept.

Limitations And Societal Impact:
I do not see any potential negative societal impact of their work.

Needs Ethics Review: No

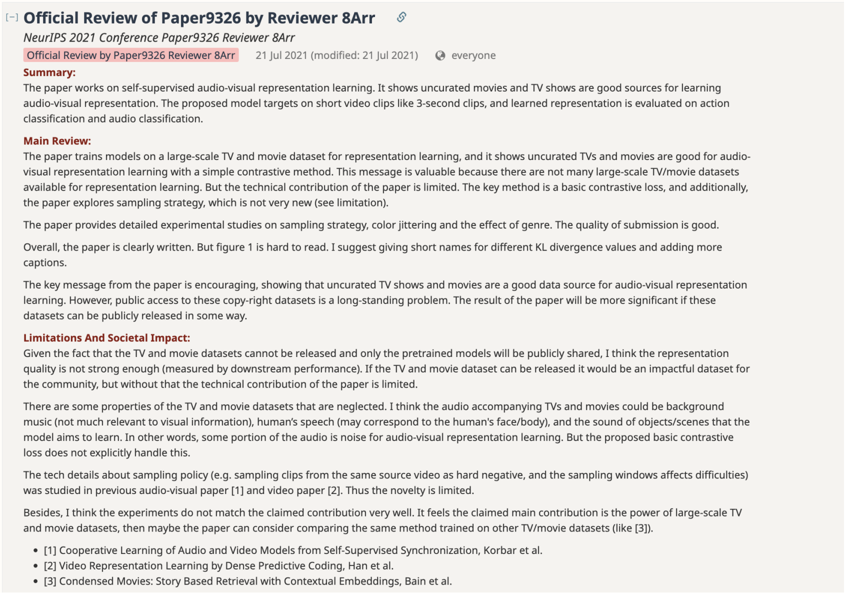
Time Spent Reviewing: 5

Rating: 7: Good paper, accept

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Figure 2: Accepted Paper's Review Example



[−] **Official Review of Paper9326 by Reviewer 8Arr**

NeurIPS 2021 Conference Paper9326 Reviewer 8Arr

Official Review by Paper9326 Reviewer 8Arr 21 Jul 2021 (modified: 21 Jul 2021) everyone

Summary:
The paper works on self-supervised audio-visual representation learning. It shows uncurated movies and TV shows are good sources for learning audio-visual representation. The proposed model targets on short video clips like 3-second clips, and learned representation is evaluated on action classification and audio classification.

Main Review:
The paper trains models on a large-scale TV and movie dataset for representation learning, and it shows uncurated TVs and movies are good for audio-visual representation learning with a simple contrastive method. This message is valuable because there are not many large-scale TV/movie datasets available for representation learning. But the technical contribution of the paper is limited. The key method is a basic contrastive loss, and additionally, the paper explores sampling strategy, which is not very new (see limitation).

The paper provides detailed experimental studies on sampling strategy, color jittering and the effect of genre. The quality of submission is good.

Overall, the paper is clearly written. But figure 1 is hard to read. I suggest giving short names for different KL divergence values and adding more captions.

The key message from the paper is encouraging, showing that uncurated TV shows and movies are a good data source for audio-visual representation learning. However, public access to these copy-right datasets is a long-standing problem. The result of the paper will be more significant if these datasets can be publicly released in some way.

Limitations And Societal Impact:
Given the fact that the TV and movie datasets cannot be released and only the pretrained models will be publicly shared, I think the representation quality is not strong enough (measured by downstream performance). If the TV and movie dataset can be released it would be an impactful dataset for the community, but without that the technical contribution of the paper is limited.

There are some properties of the TV and movie datasets that are neglected. I think the audio accompanying TVs and movies could be background music (not much relevant to visual information), human's speech (may correspond to the human's face/body), and the sound of objects/scenes that the model aims to learn. In other words, some portion of the audio is noise for audio-visual representation learning. But the proposed basic contrastive loss does not explicitly handle this.

The tech details about sampling policy (e.g. sampling clips from the same source video as hard negative, and the sampling windows affects difficulties) was studied in previous audio-visual paper [1] and video paper [2]. Thus the novelty is limited.

Besides, I think the experiments do not match the claimed contribution very well. It feels the claimed main contribution is the power of large-scale TV and movie datasets, then maybe the paper can consider comparing the same method trained on other TV/movie datasets (like [3]).

- [1] Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization, Korbar et al.
- [2] Video Representation Learning by Dense Predictive Coding, Han et al.
- [3] Condensed Movies: Story Based Retrieval with Contextual Embeddings, Bain et al.

Figure 3: Rejected Paper's Review Example

dataset. Before we state the problems, let us briefly introduce the statistics for the dataset. In the 'openreview.net', there are 628 academic events in total, and 31,062 papers were submitted in total. In addition, considering our purpose, we excluded papers with no review, and papers only with the committee's decision. Also, there were only 2010 reviews with strengths and weaknesses as in Table 1. In the other reviews, there were 24,945 reviews without explicit strengths and weaknesses, or with strengths and weaknesses being written in the review text, which makes it hard to extract. To sum up, it was hard to train the models generating strengths and weaknesses due to the lack of data.

4 Model Description

Recent text summarization models use either extractive, abstractive or hybrid (extractive + abstractive) methods. Extractive models can produce well-structured summarizations when using templates and are likely to generate an accurate summary, but they generate a limited variety of formats. Abstractive summarization may produce more natural and various styles of summarization, but they may return

	Number of Submissions	Number of Papers with Pros-and-Cons Reviews	Number of Papers without Pros-and-Cons Reviews
Venues	628	628	628
Mean	49.44	3.20	39.72
Std	293.86	42.35	494.41
Min	0	0	0
25%	0	0	0
50%	0	0	0
75%	20	0	0
Max	4064	888	10026
Total Number of Data Instances		2010	24945

Table 1: Descriptive Statistics of Our Dataset

inaccurate summarization. Since the paper reviews have a large diversity, we use the abstractive summarization technique.

For the backbone model, we use the BERT-based language model, which currently shows the best performance in most NLP tasks. Also, it can be used properly even in an environment with low computing resources comparing other big language models.

After using BERT as the backbone, we proceed with our target tasks like generating summarization, strengths, and weaknesses of the paper, and evaluating whether the paper will be accepted or not by attaching the head (i.e., MLP layer). Figure 4 is the overall structure of our models.

Note that although each task is different, they have the same backbone structure and the only different thing is the heads. At that time, the weight of the model trained by each task is not shared. Even though all tasks lie in the same domain, we think that the aims of the tasks are different and sharing weight would harm the performance.

For summarization and strength/weakness generation, we utilize BERT masked language model to generate token-by-token. For generation using the light model, we modify [2] for each task. Input is the abstract followed by the mask in the initial step. The model then predicts the most possible token among the vocabulary. Then we modify the input as an abstract and predicted token, followed by a new mask. Until the prediction of the mask reaches the stop token, we recurrently generate by inserting the mask at the end. Figure 5 depicts the generation procedure of our model.

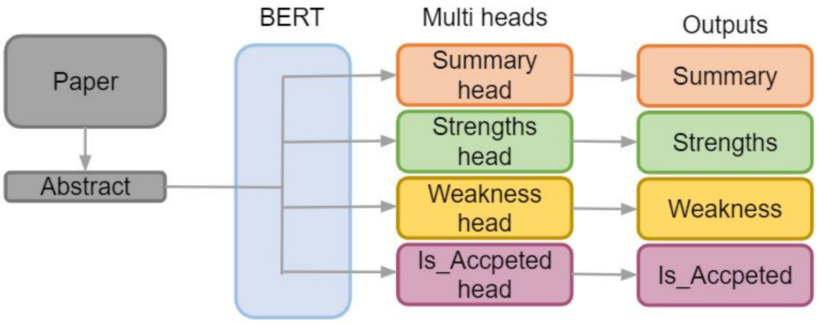


Figure 4: Overall Structure

For the acceptance prediction model, we use a simple BERT sequence prediction model. When the BERT computes representation from the text, we take sentence vector (representation of `[CLS]` token) and predict acceptance probability using multi-layer perceptrons. We use binary cross-entropy as a loss function since acceptance prediction is a binary classification task. Figure 6 shows the architecture of the acceptance prediction model.

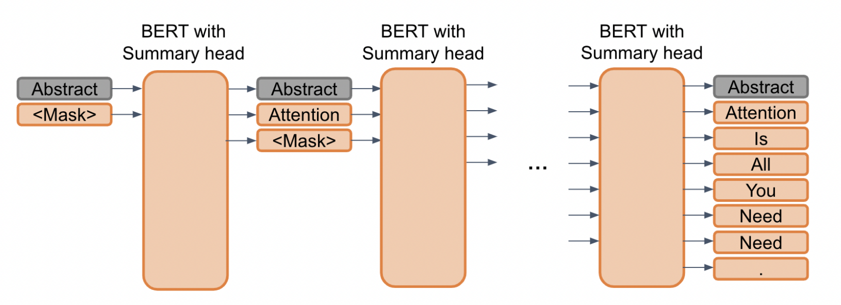


Figure 5: Explanation of Generation Procedure for a summary in one sentence

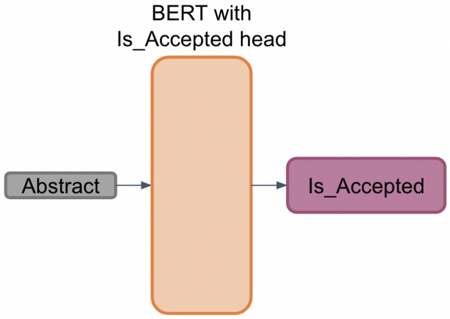


Figure 6: Architecture of acceptance prediction model

4.1 Model Implementation

As a baseline model, we use the BERT-base model to generate a summary (henceforth, TL;DR) with the hyperparameters illustrated in Table 2. We begin to experiment with an abstract part of a focal paper due to the limited input size of the language model. TL;DR texts are gathered from `openreview.net`, which are written by authors of the papers. Usually, TL;DR is comprised of two sentences. To generate this short summary, we fine-tune the pretrained BERT model with [10]. As predefined, masked language modeling (MLM) loss is used during fine-tuning. We get loss in batches - in some NLP tasks, performance optimization may not be achieved due to the incredibly large data size, resulting in an overfitting problem. Validation loss per step is described in Figure 10. Implementation of models generating strengths and weaknesses is almost the same as generating TL;DR. For the loss for classification of whether it would be accepted or not (Is_Accepted), we use Cross Entropy as a loss function.

Parameter	Value
Pretrained Model	bert-base-uncased
Learning rate	$5e-5$
Batch size	8
Patience step	2000

Table 2: Hyperparameter Settings

5 Experiment

In this section, we explain what experiments we conducted for each task. The backbone is BERT pretrained model and the only different thing is the header part for each task. For extracting TL;DR, strengths, and weaknesses, each task is trained in an MLM - based manner. In the case of acceptance, we use binary classification for predicting acceptance.

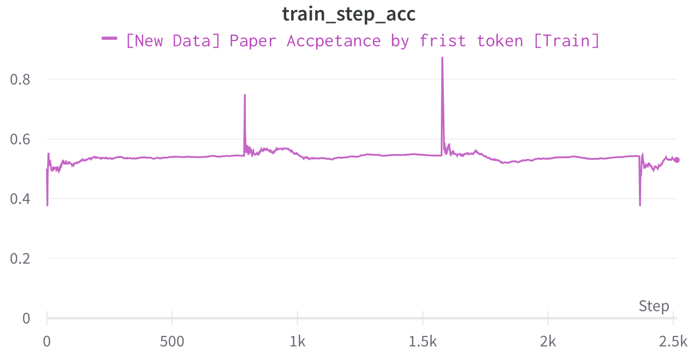


Figure 7: Train Step Accuracy

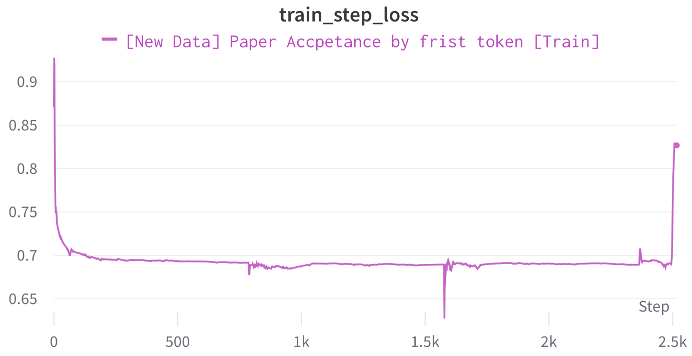


Figure 8: Train Step Loss

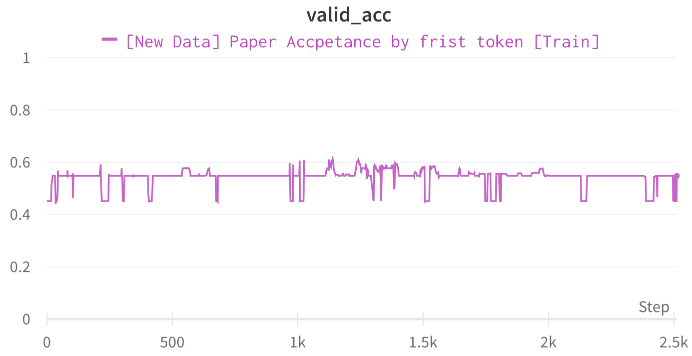


Figure 9: Valid Accuracy

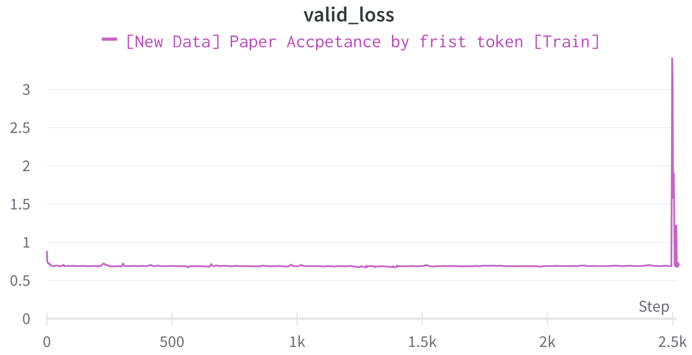
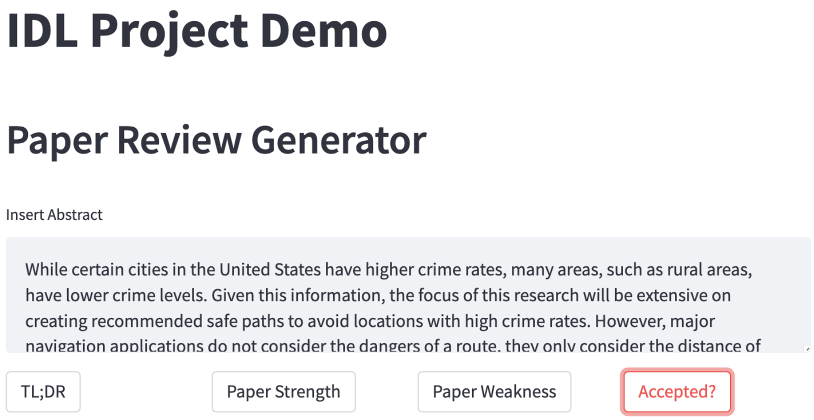


Figure 10: Valid Loss



IDL Project Demo

Paper Review Generator

Insert Abstract

While certain cities in the United States have higher crime rates, many areas, such as rural areas, have lower crime levels. Given this information, the focus of this research will be extensive on creating recommended safe paths to avoid locations with high crime rates. However, major navigation applications do not consider the dangers of a route, they only consider the distance of

Figure 11: Demo Capture 1

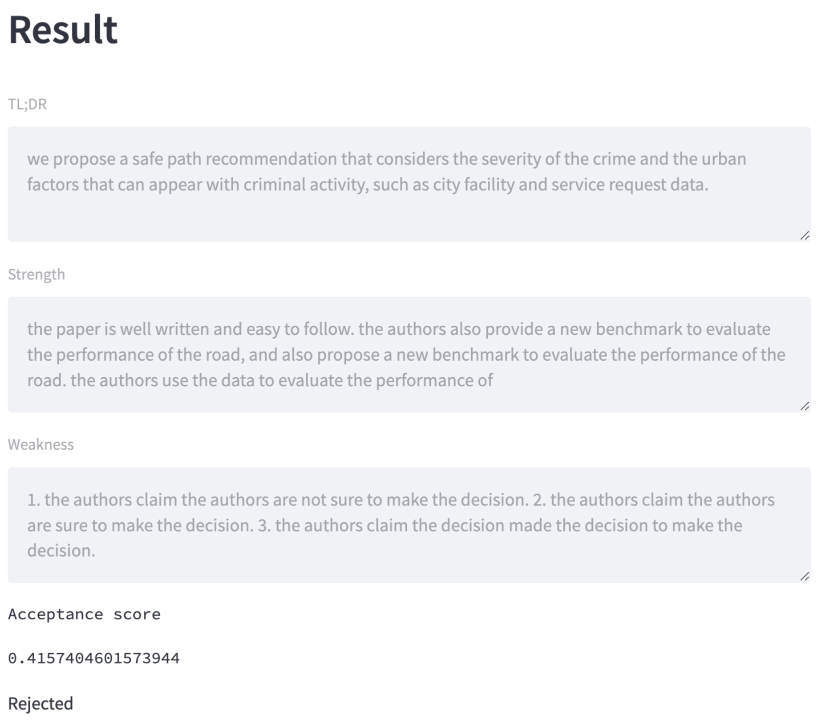
The experimental details are as follows. The learning rate is set to 0.0005, the batch size is set to 8, and epochs to 100. At this time, training is stopped if the validation loss did not decrease for 200 steps by applying early stopping. Note that we implement our method using PyTorch and all experiments are conducted using four NVIDIA Titan RTX GPUs.

The specific experimental results are as follows. Figure 7 to 10 show the performance and convergence of our model when performing the paper acceptance task. As you can see, training and validation accuracy are higher than 0.5, and loss continues to decrease. This shows our model with some degree of performance and convergence. Note that each value was measured at each step, not the epoch.

Figure 11 and 12 show our demonstration. Since our approach is easily implemented as an application, demonstrations like this show the usefulness of our model. You can test any abstract for extracting TL;DR, strengths, weaknesses, and acceptance.

6 Discussion

In this section, we evaluate the performances of our models and discuss why the performance is not good. We conducted a qualitative evaluation of the models generating a one-sentence summary, of strengths, and weaknesses. Also, we conducted a quantitative evaluation of the model that predicts whether it would be accepted or not. As you can see in Figure 12, the output of the model generating 'TL;DR' is quite plausible. However, the outputs of the models generating strengths and weaknesses are inappropriate since there are too many repeated words in the outputs. We suspect this is because of the lack of training data as illustrated in Section 3. In addition, one can easily notice that the performance of the model predicting whether it would be accepted or not is not good since the



Result

TL;DR

we propose a safe path recommendation that considers the severity of the crime and the urban factors that can appear with criminal activity, such as city facility and service request data.

Strength

the paper is well written and easy to follow. the authors also provide a new benchmark to evaluate the performance of the road, and also propose a new benchmark to evaluate the performance of the road. the authors use the data to evaluate the performance of

Weakness

1. the authors claim the authors are not sure to make the decision. 2. the authors claim the authors are sure to make the decision. 3. the authors claim the decision made the decision to make the decision.

```
Acceptance score

0.4157404601573944
```

Rejected

Figure 12: Demo Capture 2

acceptance score for the given paper is around 0.4157 although it was an accepted paper. There could be several reasons for this situation. First, it would be not sufficient to determine whether a paper would be accepted or not only using the abstract of it. Second, whether a paper would be accepted or not might depend on the conference to which the paper was submitted and the sub-field on which the paper focus. If sufficiently numerous data are given to us, we could answer these questions.

7 Conclusion

Throughout this project, we propose a deep-learning-based tool to analyze the unknown paper. We first divide the task into four subtasks: generating a one-sentence summary, strengths, and weaknesses of the given paper, and determining whether it would be accepted or not. Our models use only the abstract of the given paper for those subtasks. Using the simple BERT model and language modeling, we successfully generate TL;DR, strengths, and weaknesses from the model. For the acceptance prediction task, the accuracy record 0.57. Although the performances of the 3 tasks (generating strengths and weaknesses of the given paper, and predicting whether it would be accepted or not) are not satisfying, our attempt is meaningful there was no attempt for those 3 tasks. In future work, we may consider using a broader field of papers. Since our data have numerous papers dealing with some datasets, the diversity of strength/weakness generations is limited to some words like dataset. If more diverse types of papers are included, we expect the model can generate more accurate strengths and weaknesses. Another future work should be designing models that use the entire paper as model input. One of the problems of strength and weakness generation is that reviewers write strengths and weaknesses based on the entire paper, not the abstract. Due to the limited resource amount and time, we could only use abstract as an input. With a full paper input in the future, the model would provide better analyses considering our result that only abstract problems provide useful analyses of the paper.

References

[1] Amjad Abu-Jbara and Dragomir Radev. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 500–509, 2011.

[2] Ying-Hong Chan and Yao-Chung Fan. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, 2019.

[3] Mohamed Abdel Fattah and Fuji Ren. Automatic text summarization. *World Academy of Science, Engineering and Technology*, 37(2):192, 2008.

[4] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[5] Vahed Qazvinian and Dragomir R Radev. Scientific paper summarization using citation summary networks. *arXiv preprint arXiv:0807.1560*, 2008.

[6] Xiaoping Sun and Hai Zhuge. Summarization of scientific paper through reinforcement ranking on semantic link network. *IEEE Access*, 6:40611–40625, 2018.

[7] Oguzhan Tas and Farzad Kiyani. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213, 2007.

[8] Samer Abdulateef Waheeb, Naseer Ahmed Khan, Bolin Chen, and Xuequn Shang. Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. *Information*, 11(5):281, 2020.

[9] Qicai Wang, Peiyu Liu, Zhenfang Zhu, Hongxia Yin, Qiuyue Zhang, and Lindong Zhang. A text abstraction summary model based on bert word embedding and reinforcement learning. *Applied Sciences*, 9(21), 2019.

[10] 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. 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, Online, October 2020. Association for Computational Linguistics.

[11] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393, 2019.