Journal of Current Science and Technology, May - August 2024 Vol. 14 No. 2, Article 26

ISSN 2630-0656 (Online)

Cite this article: Sanchan, N. (2024). Comparative study on automated reference summary generation using BERT Models and ROUGE score assessment. *Journal of Current Science and Technology*, *14*(2), Article 26. https://doi.org/10.59796/jcst.V14N2.2024.26

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Comparative Study on Automated Reference Summary Generation using BERT Models and ROUGE Score Assessment

Nattapong Sanchan

School of Information Technology and Innovation, Bangkok University, Pathum Thani 12120, Thailand

E-mail: nattapong.sa@bu.ac.th

Received 6 October, 2023; Revised 31 October, 2023; Accepted 15 December, 2023
Published online 2 May, 2024

Abstract

Automatic text summarization is a sub-area in text mining in which a computer system determines the most informative information in the original text to produce a summary for certain jobs and users. In the development of the systems, one of the most important tasks is to evaluate the quality of summaries produced by the systems. Generally, the evaluation task becomes laborious, time-consuming, and expensive because it requires significant efforts on annotation tasks for humans to manually create reference summaries. Being able to generate automatic reference summaries would promote the development of summarization systems in term of speed and evaluation. In this paper, we proposed an Auto-Ref Summary Generation framework for automatically generating reference summaries used in the generic text summarization evaluation task, the Sliced Summary. Given a set of clusters from a cluster ground-truth label dataset, variants of BERT models were utilized for creating cluster representations. The automatic reference summaries were later generated through a centroid-based summarization approach. Overall, DistilBERT, ROBERTa, and SBERT have played crucial roles in automatic summary generation, achieving the highest ROUGE-1 score of 0.47060. However, this does not meet our expectation on text coherence and readability aspects. Although the summaries generated through our proposed framework could not be used as the replacement of the manual summaries, this study has shed new light on the acquisition of automatic reference summaries from a ground-truth label dataset.

Keywords**:** *automatic summarization; document clustering; k-means; centroid-based summarization; natural language processing; text mining*

1. Introduction

Owing to the exceptional growth of textual information in online media, automatic text summarization has emerged to facilitate users to rapidly digest content in text. Automatic text summarization is a sub-area of text mining in which a system determines the most informative information in the original text to produce a summary for certain tasks and users. To generate a summary, researchers have developed summarization systems for different purposes (i.e., single document summarization (Liu et al., 2019b), multi-document summarization (Chen et al., 2023), aspect-based opinion summarization (Wu et al., 2016), query-focused summarization (Baumel et al., 2016), update summarization (Delort, & Alfonseca, 2012), and cross-language document summarization (Wan, 2011) to summarize different text genres such as product reviews (Yu et al., 2016), news articles (Huang et al., 2011), political text (Sharevski et al., 2021), meeting text (Oya et al., 2014), scientific articles (Altmami, & Menai, 2022), online debates (Sanchan et al., 2017; Sanchan,

Bontcheva, & Aker, 2020), and medical data (Abacha et al., 2021; Tang et al., 2023) with the assistant of Artificial Intelligence i.e., ChatGPT in their summarization task. Later, the generated summaries will be assessed against various criteria such as informativeness, text coherence, readability, and understandability.

To assess the quality of a summary generated by a system (system summary), the evaluation can be extrinsic or intrinsic (Nenkova, & McKeown, 2011). In an extrinsic evaluation, a system summary is assessed on its usefulness for a particular task whereas in an intrinsic evaluation, a summary is compared to a reference summary which is usually created by one or more humans. However, the acquisition of a reference summary is an expensive, laborious, and time-consuming task. For instance, in the summarization of ten thousand tweets, a human must read and understand the entire content expressed in those tweets and then manually create a reference summary that reflects the informative content in the tweets. In addition, more than one set of reference summaries may need to be created by different humans to prevent bias arising by one single reference summary. Therefore, these acquisition difficulties open a research gap for us to explore the automatic generation of reference summaries.

In this paper, we proposed a framework to automatically acquire an automatic generic reference summary using a cluster ground-truth label dataset, DUC 2004 (Task 2). We investigated which model will provide the highest ROUGE-N scores. Those models included ALBERT, BERT-Based, DistilBERT, ROBERTa, SBERT, SqueezeBERT, and a baseline, Bag of Words model using TF*IDF. In the automatic reference summary generation, we generated cluster representations using the aforementioned models and extracted salient sentences to form the summaries through the proposed Sliced Summary methods. Overall, we achieved the highest ROUGE-1 scores and DistilBERT, ROBERTa, and SBERT played significant roles in summary generation. Being able to create automatic referenced summaries would benefit the research communities by expediting the summary evaluation and the development of generic summarization systems.

2. Background and related work

Assessment of summaries allows us to measure the quality of the summary created by a system. The assessment generally falls into extrinsic and intrinsic evaluations. Extrinsic evaluation, also called a task-based evaluation, measures the effectiveness of the summary in a certain task. In the intrinsic evaluation, which this paper focuses on, a system summary is assessed based on its coherence and informativeness.

Generally, there are two types of summaries required for intrinsic evaluation: a system summary and a reference summary. A system summary is the summary that is generated by a computer system and a reference summary is that created by one or more human subjects. The following subsections discuss the common intrinsic evaluation metrics.

2.1 Human Judgment

One of the most traditional intrinsic evaluation metrics is the utilization of human subjects to assess the quality of the summary. For example, Minel et al. (1997) asked human subjects to rate a set of system summaries based on specific criteria such as the presence of significant ideas and specific content, repeating of unneeded concepts, the flow sequence of arguments in text, readability, etc. Based on the quality of the text, examples of their rating scales are *clear*, *fairly clear*, *not very clear*, *incomprehensible*, etc. Likewise, Radev et al. (2000) proposed a utility-based evaluation to evaluate both single-document and multi-document summaries. For instance, to evaluate a set of salient sentences in clusters, the researchers defined a scale (called a utility point) in a range of 1-10. A utility of 0 refers to the sentence that does not belong to the cluster whereas a utility of 10 indicates the salient sentence. Furthermore, a recent work by Monsen, & Rennes (2022) studied how users perceived extractive and abstractive summaries. They conducted an online survey to compare legibility, fluency, and simplicity. Due to the need of human assessors, this type of evaluation needs time, labor effort, and costs.

2.2 ROUGE

The next evaluation metric is Recall-Oriented Understudy for Gisting Evaluation (ROUGE). It is one of the most typical evaluation metrics that

measures n-gram overlaps between a system summary and a reference summary (Lin, 2004). Equation 1 illustrates the ROUGE equation where *n* refers to the size of grams. $Count_{match}$ ($gram_n$) is the greatest number of terms appearing in both summaries. As shown in Equation 1 the equation, ROUGE is recall-based since the denominator is the overall addition of the entire number of terms occurring in the reference summary. Adding additional terms in the reference summary, the number in the denominator increases.

$$\text{ROUGE-N}=\frac{\sum_{S\in\{\text{Reference Summaries}\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{\text{Reference Summaries}\}}\sum_{gram_n\in S}Count(gram_n)} \quad (1)$$

There are different variants of ROUGE. For instance, ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-4 measure the overlapping of unigrams, bigrams, trigrams, and fourgrams respectively. ROUGE-L measures the longest common subsequence co-occurrences of terms in system and reference summaries. ROUGE-S measures skip-gram co-occurrences of terms. ROUGE-SU4 measures both unigram and skip-bigram up to four terms. An example of recent related work that reported ROUGE scores in their summarization system is by Sanchan et al. (2018). The authors proposed a system that summarizes online debates by extracting salient sentences from the debates. The sentence extraction was performed using a set of prominent features i.e., sentence position, debate titles, sentence length, adverbs, and term similarities. The results revealed that the best ROUGE scores were derived through the sentence position feature with ROUGE-1, ROUGE-2, and ROUG-SU4 of 0.6124, 0.5375, and 0.487 respectively.

2.3 Evaluation Without Reference Summaries

Aside from the aforementioned metrics, some researchers endeavor to propose a methodology to perform the evaluation without a reference summary. The goals are to be used when 1) no reference summaries are available and 2) there is only one a single reference summary which might yield less accuracy in the evaluation task. The general intuition of using no reference summaries is to evaluate the system summary against the original text. Louis, & Nenkova (2009) proposed a method for evaluating sentence selection in an automatic summarization task. This method does not require the construction of reference summaries. Instead, the researchers assumed that the distribution of terms in the input documents and their generated summary should be similar. This methodology was evaluated by determining the ranking correlation of 1) the summary and the pyramid approach; and 2) the summary and human judgment. By applying different features, JensenShannon divergence yields the highest correlation of 0.88 and 0.73 for the pyramid and human judgment respectively. Another research work was proposed by Saggion et al. (2010). They proposed the correlation of ranking different summarization systems (e.g., generic single-document, generic multi-document, and focused-based summarization systems) with and without reference summaries. Instead of using term distribution, they utilized n-grams and skipped n-grams probability distributions in their system.

The endeavor to evaluate system summaries without using reference summaries is the most relevant related work to our goal. We addressed an alternative for generating an automated generic reference summary when reference summaries are not available. The automated summary can be later used along with ROUGE evaluation metrics for the quality assessment of the system summaries.

3. Materials and Method

We proposed a method for generating automated reference summaries as illustrated in Figure 1, the structure of the Auto-Ref Summary Generation framework.

3.1 DUC 2004 Dataset

Document Understand Conference (DUC) was organized to study and promote the continuous growth of automatic text summarization research from 2001 to 2007. Each year, the conference provided data and instructions for participants to conduct summarization systems. As well as in DUC 2004, Task 2, a set of 50 news clusters was provided to the participants to develop a system for summarizing a 665-byte summary for each cluster (U.S. Commerce Department, National Institute of

Standards and Technology, 2014). To evaluate the system summaries, DUC 2004 provided 4 model summaries used for the evaluation. We considered that these manual summaries are the summaries of all clusters and therefore could also be used for the evaluation of a single summary generated from all clusters. In this paper, we therefore generated an automatic reference summary from the entire set of clusters and used the 4 manual summaries for the evaluation. In total, the dataset contains 50 clusters, each with 10 news documents, accounting for 336,207 words. We utilized this dataset for creating cluster representations in the next step.

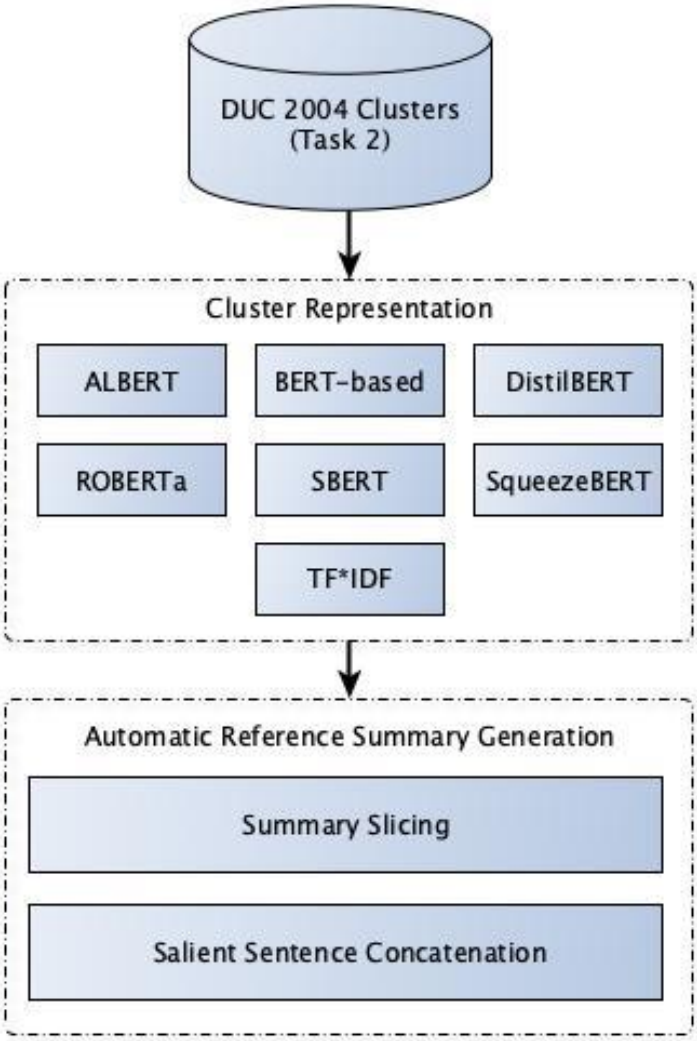

Figure 1 Auto-Ref Summary Generation Framework

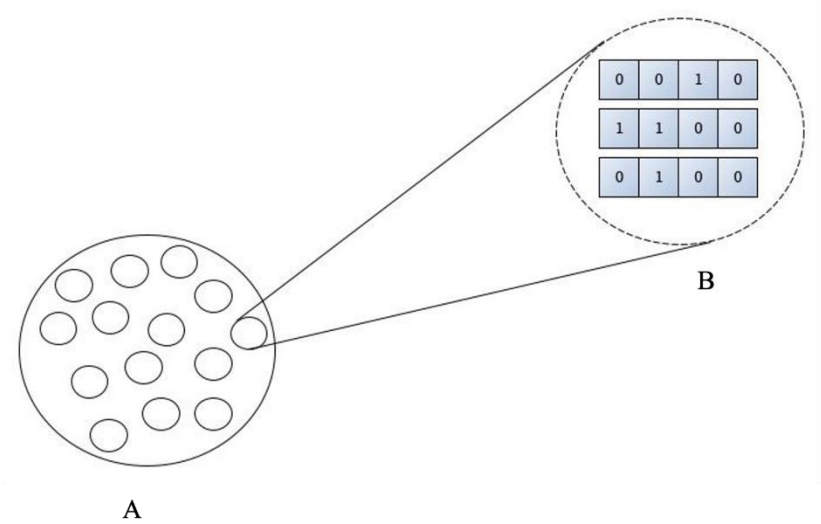

Figure 2 Cluster Representation on the given DUC 2004 news clusters

3.2 Cluster Representation

Text representation refers to the conversion of text into a form that enables systems to understand the context. A decent representation should be able to capture all essential content, semantics, and structures of the original text. To create cluster representation, in this paper, we utilized a traditional Bag of Word (BoW) model (TF*IDF) and several variants of Bidirectional Encoder Representations from Transformers (BERT), including ALBERT, BERT-Based, DistilBERT, ROBERTa, SBERT, and SqueezeBERT. Overall, given 50 DUC 2004 news clusters, we obtained 50 cluster representations through BoW and BERT models. In total, there are 7 models so that we derived 350 cluster representations. For example, circle A, in Figure 2, illustrates cluster representations for the given 50 DUC 2004 news clusters derived through a BERT model. Circle B in the figure shows a text representation for each sentence in the cluster. The following subsections elaborates on concepts and techniques used in the generation of text representation with Bag of Words and BERT models.

3.2.1 Bag of Words

Bag of Words (BoW) is a traditional text representation that embodies words as an unordered set with a frequency of those words. In this paper, we represented each cluster as the BoW model and extracted Term Frequency * Inverse Document Frequency (TF*IDF). Equations 2 – 4 illustrate TF*IDF equations, where *tf* of word *w* in cluster *c* is the frequency of word *w* in cluster *c*. *idf* of word *w* in the entire cluster corpus C is the logarithm of a number of clusters in the corpus C divided by the number of clusters having word *w*. The interpretation of an IDF value of word *w* closer to 0 indicates that the word is more common. The higher the value of TF*IDF, the rarity of the word occurrence. The final output of this BoW model is vector representations of clusters that will be the input for the summary generation.

$$tf(w,c)=\frac{f_c(w)}{\sum_{w\in c} f_c(w)} \quad (2)$$

$$idf_{(w,C)}=\ln\left(\frac{|C|}{|w\in C:w\in c|}\right) \quad (3)$$

$$tf*idf(t,w,C)=tf(w,c)*idf_{(w,C)} \quad (4)$$

3.2.2 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a recent model that generates text representations. It was bidirectionally trained on a massive amount of unannotated text such as Wikipedia text and numerous book corpora. As it is bidirectional, the model recognizes the sequence of

text structures in both directions, from right to left and from left to right (Kenton, & Toutanova, 2019). Figure 3 illustrates the generic structure of the BERT model where a text representation is generated. Initially, a sentence is tokenized and later added with special tokens. *[CLS]* and *[SEP]* added to sentences representing the start and end of each sentence. *[PAD]* will be added to a sentence only when its length is shorter than the maximum one. In the next step, each token will be mapped to a unique identification number representing the token. The final output is a vector representation of the sentence.

3.3 Automatic Reference Summary Generation

Generally, in the evaluation of text summarization paradigm, system summaries are commonly assessed against a set of references summaries created by humans. In case no reference summaries available, the evaluation may not be successful. Therefore, we proposed a method to generate an automatic generic reference summary by employing a centroid-based summarization.

After deriving 350 cluster representation through BoW and BERT models, we calculated the centroid of each cluster using the square-error criterion as shown in Equation 5 (Han, & Kamber, 2006). Within each cluster, we defined the value of *k* to *1* so that we obtained the overall distances among sentences in the cluster. We then extracted salient sentences at the centroid to form the automatic reference summaries, which express the central information of each cluster. Our assumption regarding the generation of automatic reference summaries is that all clusters have equal contributions for summary generation. We used Sliced Summary methods for summary generation.

$$E = \sum_{i=i}^{k} \sum_{p \in C_i} |p\text{-}m_i|^2 \quad (5)$$

In this paper, we extracted salient sentences to form the automatic generic reference summaries which express the central information of each cluster. Our assumption regarding the generation of automatic reference summaries is that all clusters have equal contributions for summary generation.

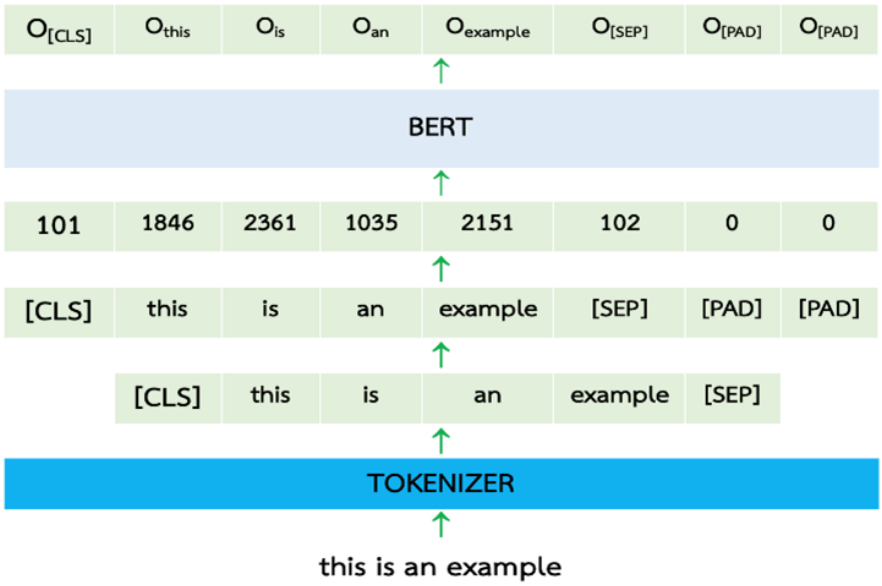

Figure 3 Example of Generic BERT Structure for Generating Text Representation
Adapted from "The performance of BERT as data representation of text clustering" by Subakti et al. (2022)

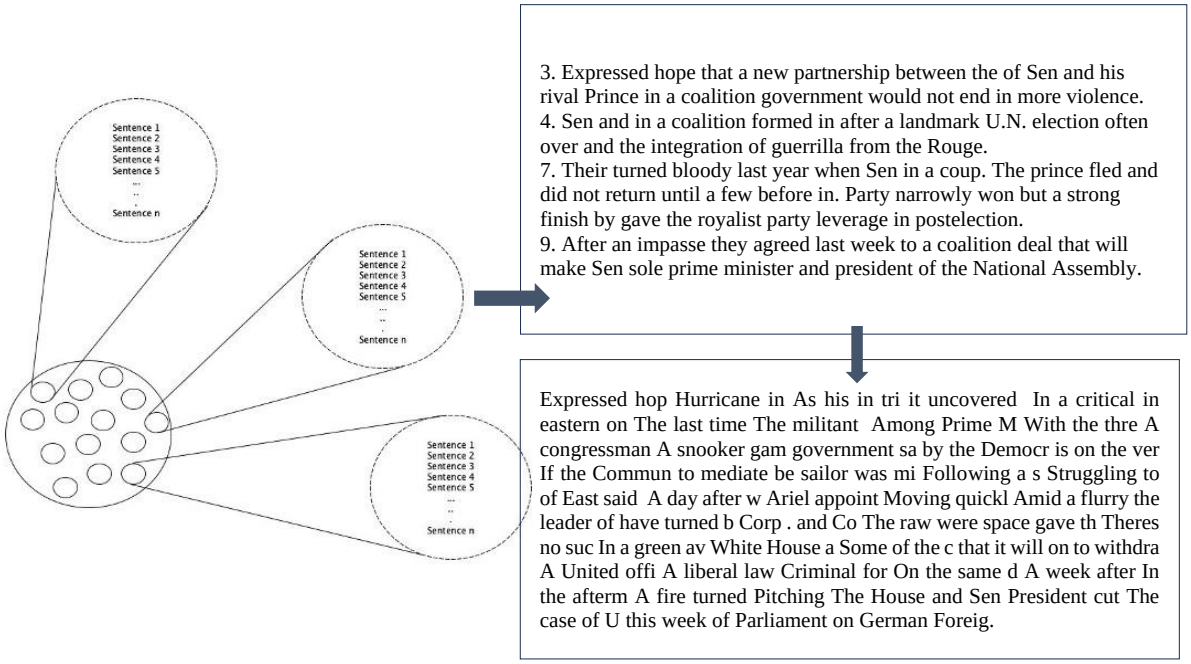

Figure 4 Sequential Steps for Generating Automatic Generic Reference Summary using the Sliced Summary Method

3.3.1 Sliced Summary

In this method, we considered that every cluster contains significant pieces of information and thus should be included in the automatic reference summary. As shown in Figure 4, within each cluster, we extracted one salient sentence closeted to the cluster's centroid. Then we sliced the sentences according to the proposed formular shown in Equation 6. From the equation, N_w is the number of maximum characters that can be extracted from one cluster which is calculated by the floor of N_c (the number of clusters in the dataset) divided by C (the compression rate as the number of characters). For instance, with a compression rate of 655 bytes (C), when the equation returns the value of 10 for a sentence in a cluster, only first 10 characters can be extracted from the sentence in that cluster.

$$N_w = \lfloor \frac{N_c}{C} \rfloor \quad (6)$$

3.3.2 Salient Sentence Concatenation

In this method, we follow the common text summarization approach for the automatic generic reference summary generation as presented in Rossiello et al. (2017) and Saggion, & Gaizauskas (2004). We named this methodology Salient Sentence Concatenation, as from each cluster, only one sentence, in which its distance is nearest to the center of the cluster was extracted until the compression rate of 655 byes was satisfied. We also prevented the selection of similar sentences in the summary by measuring semantic similarities among the selected sentences. If the semantic similarity is empirically greater than 0.5, the sentence is ignored.

4. Results

To evaluate the quality of the automatic generic reference summaries, we compared those to four manual summaries provided by DUC2004 and reported ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, and ROUGE-L scores as shown in Table 1. The bold values in the table indicated the highest scores in each ROUGE metric.

Table 1 ROUGE Scores of the Automatic Generic Reference Summaries Generated through Each Model

Models	ROUGE-1	ROUGE -2	ROUGE -3	ROUGE -4	ROUGE -L
Sliced Summary					
- ALBERT	0.39800	0.14300	0.10480	0.08480	0.26720
- BERT-Based	0.37330	0.15050	0.10460	0.08670	0.26400
- DistilBERT	0.43300	**0.17780**	0.11840	0.09630	**0.29830**
- ROBERTa	**0.43710**	0.14680	0.10440	0.08650	0.27630
- SqueezeBERT	0.38460	0.13720	0.09320	0.07780	0.25100
- SBERT	0.38450	0.16740	**0.12110**	**0.10190**	0.28170
- TF*IDF	0.42220	0.16050	0.11390	0.09580	0.29010
Salient Sentence Concatenation					
- ALBERT	0.36950	0.11160	0.07550	0.07440	0.22610
- BERT-Based	0.34930	0.10470	0.07390	0.07140	0.23520
- DistilBERT	**0.47060**	**0.24930**	**0.21930**	**0.21560**	**0.35490**
- ROBERTa	**0.47060**	**0.24930**	**0.21930**	**0.21560**	**0.35490**
- SqueezeBERT	0.39620	0.12210	0.07910	0.07440	0.22710
- SBERT	**0.47060**	**0.24930**	**0.21930**	**0.21560**	**0.35490**
- TF*IDF	0.38280	0.12080	0.08480	0.07530	0.23610

Table 2 Example of two summaries generated by the Sliced Summary and Salient Sentence Concatenation methods through ROBERTa

Sliced Summary	Salient Sentence Concatenation
Expressed hop Hurricane in As his in tri it uncovered In a critical in eastern on The last time The militant Among Prime M With the thre A congressman A snooker gam government sa by the Democr is on the ver If the Commun to mediate be sailor was mi Following a s Struggling to of East said A day after w Ariel appoint Moving quickl Amid a flurry the leader of have turned b Corp . and Co The raw were space gave th Theres no suc In a green av White House a Some of the c that it will on to withdra A United offi A liberal law Criminal for On the same d A week after In the afterm A fire turned Pitching The House and Sen President cut The case of U this week of Parliament on German Foreig.	Expressed hope that a new partnership between the of Sen and his rival Prince in a coalition government would not end in more violence. Sen and in a coalition formed in after a landmark U.N. election often over and the integration of guerrilla from the Rouge. Their turned bloody last year when Sen in a coup. The prince fled and did not return until a few before in. Party narrowly won but a strong finish by gave the royalist party leverage in postelection. After an impasse they agreed last week to a coalition deal that will make Sen sole prime minister and president of the National Assembly.

From the table above, automatic reference summaries were generated through two methods: the Sliced Summary and the Salient Sentence Concatenation. In the first method, ROBERTa achieved the highest ROUGE-1 score of 0.43710 while the highest ROUGE-3 and ROUGE-4 scores were achieved by SBERT. For ROUGE-2 and the longest matching of terms, ROUGE-L, DisilBERT achieved the highest scores of 0.17780 and 0.29830 respectively. For the second method, DisilBERT, ROBERTa, and SBERT achieved the highest ROUGE scores in every aspect. These outperformed the scores derived from the baseline, TF*IDF.

Overall, the primary reason that ROBERTa yielded the highest ROUGE scores for both Sliced Summary and Salient Sentence Concatenation methods was presumably, the quantity and variety of data used in the pretraining phase. ROBERTa was pretrained with 160GB of data gathered from various sources, compared to the original BERT with that of 16GB (Liu et al., 2019a), and ALBERT (Lan et al., 2020) and DistilBERT with 16GB of book and Wikipedia text. Moreover, in general, the results in each ROUGE-N were not significantly greater than the others. Other possible factors leading to the different results included the number of layers,

training time, parameter reduction, and methods used for training BERT models. For instance, while other models were trained with and without masked language modeling (MLM) and next-sentence prediction (NSP), DistilBERT was integrated with knowledge distillation in which the model was trained to reproduce the behavior of a larger model, BERT. By containing the distillation token, the model enhances the quality of textual sequence representation.

In addition, another factor leading to the similar ROUGE scores derived from the Salient Sentence Concatenation method is the adjustment of semantic similarity. When a salient sentence was extracted from each cluster, it will be included in the automated reference summary. We eliminated the information overlap in the summary by defining a semantic similarity threshold of 0.5, indicating that when a salient sentence is extracted, it is compared to the prior sentences in the summary. If the semantic similarity is greater than the defined threshold, that sentence is contemplated as having overlapped information and not be included in the summary.

Furthermore, Table 2 shows an example of the automatic summaries generated by the Sliced Summary (left) and the Salient Sentence Concatenation (right) methods via ROBERTa. By measuring the coherency and readability aspects, on the left, it can be seen that the summary generated by the Sliced Summary method was incoherent and hardly legible as a certain number of characters were cut out from all clusters. This led to the inferior sentence structure and thus the lower ROUGE scores. For the automated reference summary generated thought the Salient Sentence Concatenation method, the sentences in the summary were more coherent and readable. We achieved the highest ROUGE-1 score of 0.47060. Although the model achieved the highest ROUGE-1 score, it confirms that ROUGE does not directly measure text readability and text coherency. In other words, the measurement of legibility and coherency should be performed separately.

To summarize, the ideal ROUGE scores of automatic reference summaries should be equal to 1 so that the summaries could be used as the replacement for the manual summaries. However, we achieved the highest ROUGE-1 score of 0.43710 for the proposed work. We concluded that the summaries derived from the Sliced Summary method are not practical and therefore cannot be replaced the manual summaries. Future investigation to acquire higher scores, such as the determination of only essential clusters used for generating the summaries, the adaption of the proposed framework to other text genres, and the application of different methodologies to generate the automatic references summaries should be considered.

5. Conclusion

Generally, in the evaluation of text summarization paradigm, system summaries are commonly assessed against a set of references summaries created by humans. In case no reference summaries are available, the evaluation may not be successful. In this paper, we addressed an alternative for generating automatic reference summaries for a generic text summarization evaluation task. Given a set of 50 news document clusters, several variants of BERT models were used for creating cluster representations. We viewed that all clusters have equal contributions to generating the summary. Therefore, in each cluster, we determined cluster centroids and extracted salient sentences to form the summaries using two methods: the Sliced Summary and the Salient Sentence Concatenation. Overall, DistilBERT, ROBERTa, and SBERT have played crucial roles in the generation of summaries.

When generating a summary using the Sliced Summary method, we found that the summary had high ROUGE scores but lacked legibility and coherency. For that generated through the Sliced Summary, the summary lacks coherency and legibility. When compared to that of the Salient Sentence Concatenation, the summary is more readable but not completely coherent. While the summaries generated through our proposed framework could not be entirely used as the replacement of the manual summaries, this study has shed new light on acquiring automatic reference summaries from a cluster's ground truth label dataset. In addition, this study confirmed that the assessment of summaries should not only be assessed through ROUGE but also evaluated on readability and coherency.

Our proposed framework illuminated a methodology to acquire an automatic generic

reference summary without human involvement. We have considered that all clusters made equal contributions to the generation of automatic reference summaries. In future work, we aim to investigate the opposite aspect, the inequality of the clusters' contributions to the generation of the summaries. A method to select only necessary clusters, together with the adaption of our proposed framework in different genres of datasets, is also worth investigating. Moreover, an issue of requiring domain experts for reference summary creation should be also addressed. Furthermore, future explorations on algorithms and text representations remain until we meet the success of creating automatic reference summaries. Being able to generate a set of ideal automatic reference summaries could benefit the NLP research communities by expediting the summary evaluation and the implementation of automatic text summarization systems.

6. References

Abacha, A. B., M'rabet, Y., Zhang, Y., Shivade, C., Langlotz, C., & Demner-Fushman, D. (2021, June). *Overview of the MEDIQA 2021 Shared Task on Summarization in The Medical Domain* [Conference Presentation]. In Proceedings of the 20th Workshop on Biomedical Language Processing. https://doi.org/10.18653/v1/2021.bionlp-1.8

Altmami, N. I., & Menai, M. E. B. (2022). Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences, 34*(4), 1011-1028. https://doi.org/10.1016/j.jksuci.2020.04.020

Baumel, T., Cohen, R., & Elhadad, M. (2016, February 12-17). *Topic Concentration in Query Focused Summarization Datasets* [Conference Presentation]. The Thirtieth Conference on Artificial Intelligence (AAAI-16), Phoenix, Arizona, US. https://doi.org/10.1609/aaai.v30i1.10323

Chen, H., Zhang, H., Guo, H., Yi, S., Chen, B., & Zhou, X. (2023, September 18). *SALAS: Supervised Aspect Learning Improves Abstractive Multi-document Summarization Through Aspect Information Loss* [Conference Presentation]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin, Italy. https://doi.org/10.1007/978-3-031-43421-1_4

Delort, J. Y., & Alfonseca, E. (2012, April 23-27). *DualSum: A Topic-Model Based Approach For Update Summarization* [Conference Presentation]. 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France. https://aclanthology.org/E12-1022.pdf

Han, J., & Kamber, M. (2006). Data mining concepts and techniques (2ed.). USA: Morgan Kaufmann.

Huang, X., Wan, X., & Xiao, J. (2011, June 19-24). *Comparative news summarization using linear programming* [Conference Presentation]. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA. https://link.springer.com/chapter/10.1007/978-981-16-9012-9_21

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June 2-7). *Bert: Pre-training of deep bidirectional transformers for language understanding* [Conference Presentation]. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, Minnesota, USA. https://aclanthology.org/N19-1423.pdf

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020, February 9). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* [Conference Presentation]. The International Conference on Learning Representations, Addis Ababa, Ethiopia. https://doi.org/10.48550/arXiv.1909.11942

Lin, C. Y. (2004). *Rouge: A Package for Automatic Evaluation of Summaries* [Conference Presentation]. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain. https://aclanthology.org/W04-1013/

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019a). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* [Conference Presentation]. The International

Conference on Learning Representations, Online.

Liu, Y., Titov, I., & Lapata, M. (2019b). *Single Document Summarization As Tree Induction* [Conference Presentation]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/n19-1173

Louis, A., & Nenkova, A. (2009, August). *Automatically evaluating content selection in summarization without human models* [Conference Presentation]. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore. https://doi.org/10.3115/1699510.1699550

Minel, J. L., Nugier, S., & Piat, G. (1997). *How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and Their Results on SERAPHIN* [Conference Presentation]. In Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain.

Monsen, J., & Rennes, E. (2022, June 20-25). *Perceived text quality and readability in extractive and abstractive summaries* [Conference Presentation]. 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

Nenkova, A., & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval, 5*(2–3), 103-233. https://doi.org/10.1561/9781601984715

Oya, T., Mehdad, Y., Carenini, G., & Ng, R. (2014, June 19-21). *A Template-Based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships* [Conference Presentation]. Proceedings of the 8th International Natural Language Generation Conference (INLG), Pennsylvania, U.S.A. https://doi.org/10.3115/v1/w14-4407

Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management, 40*(6), 919-938. https://doi.org/10.3115/1117575.1117578

Rossiello, G., Basile, P., & Semeraro, G. (2017, April 3). *Centroid-based text summarization through compositionality of word embeddings* [Conference Presentation]. Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres, Valencia, Spain. https://doi.org/10.18653/v1/w17-1003

Sanchan, N., Aker, A., & Bontcheva, K. (2017). *Automatic summarization of online debates* [Conference Presentation]. In Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017, Varna, Bulgaria. https://doi.org/10.26615/978-954-452-038-0_003

Saggion, H., & Gaizauskas, R. (2004). *Multi-document summarization by cluster/profile relevance and redundancy removal* [Conference Presentation]. In Proceedings of the Document Understanding Conference.

Saggion, H., Torres-Moreno, J. M., da Cunha, I., SanJuan, E., & Velázquez-Morales, P. (2010). *Multilingual summarization evaluation without human models* [Conference Presentation]. In Coling 2010: Posters, Beijing, China.

Sanchan, N., Aker, A., & Bontcheva, K. (2018, April 17-23). *Gold Standard Online Debates Summaries and First Experiments Towards Automatic Summarization of Online Debate Data* [Conference Presentation]. Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary. https://doi.org/10.1007/978-3-319-77116-8_37

Sanchan, N. Bontcheva, K. & Aker, A. (2020, October 21-22). *An Adoption of a Contradiction Detection Task to Assist the Summarization of Online Debates* [Conference Presentation]. The 5th International Conference on Information Technology (InCIT), Chonburi, Thailand. https://doi.org/10.1109/incit50588.2020.9310941

Sharevski, F., Jachim, P., & Pieroni, E. (2021, November 15). *Regulation TL; DR:*

Adversarial Text Summarization of Federal Register Articles [Conference Presentation]. In Proceedings of the 3rd Workshop on Cyber-Security Arms Race, New York, USA. https://dl.acm.org/doi/abs/10.1145/3474374.3486917

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of big Data*, *9*(1), 1-21. https://doi.org/10.1186/s40537-022-00564-9

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., ... & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, *6*(1), 158. https://doi.org/10.1101/2023.04.22.23288967

U.S. Commerce Department, National Institute of Standards and Technology (2014, September 9). *DUC 2004 Past Data*. https://www-nlpir.nist.gov/projects/duc/data/2004_data.html

Wan, X. (2011, June 19-24). *Using bilingual information for cross-language document summarization* [Conference Presentation]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA. https://dl.acm.org/doi/10.5555/2002472.2002659

Wu, H., Gu, Y., Sun, S., & Gu, X. (2016, July 24-29). *Aspect-Based Opinion Summarization with Convolutional Neural Networks* [Conference Presentation]. In 2016 International Joint Conference on Neural Networks, Vancouver, BC, Canada. https://doi.org/10.1109/ijcnn.2016.7727602

Yu, N., Huang, M., Shi, Y., & Zhu, X. (2016, December 11-17). *Product Review Summarization By Exploiting Phrase Properties* [Conference Presentation]. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan. https://aclanthology.org/C16-1106/