tBART: Abstractive summarization based on the joining of Topic modeling and BART

Binh Dang, Dinh-Truong Do, Le-Minh Nguyen
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
binhdang@jaist.ac.jp, truongdo@jaist.ac.jp, nguyenml@jaist.ac.jp

2022 14th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-5281-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/KSE56063.2022.9953613

***Abstract*—Topic information has been helpful to direct semantics in text summarization. In this paper, we present a study on a novel and efficient method to incorporate the topic information with the BART model for abstractive summarization, called the tBART. The proposed model inherits the advantages of the BART, learns latent topics, and transfers the topic vector of tokens to context space by an align function. The experimental results illustrate the effectiveness of our proposed method, which significantly outperforms previous methods on two benchmark datasets: XSUM and CNN/DAILY MAIL.**

***Index Terms*—Topic modeling, knowledge injection, BART, align function.**

I. Introductions

Automatic text summarization is the technique of efficiently extracting and compressing information from input papers while preserving their essential information. This method is crucial to the domains of several natural language processing (NLP) [1]. Currently, extractive and abstractive are two fundamental types of solutions for summarization [2]. The abstractive method creates unique words or phrases with comprehension, whereas the extractive method chooses important words, and sentences or rearranges words and sentences from the original document. The majority of methods in use today are usually designed to encrypt paragraphs and then decode them using a variety of processes. However, during the encoding and decoding procedures, there is a large amount of information loss. As a result, word embedding or contextual contents are the main focus of existing summarization research.

With the popularity of transformer-based models, the challenge of summarization is how to use the pre-trained transformer-based language model to represent and generate. It requires richer semantics information in the representation and training processes. The summary needs coherence and relatedness. And topic information injection is one of the solutions for this problem. The effectiveness by using latent topic information as features for information retrieval, recommendation system, and semantic textual similarity has been pointed out in several studies [3], [4], [5], [6] and [7]. Topic models are more adept at picking up precise document semantics than transformers are; hence, they might be included in transformers to improve their performance even more.

In this work, we suggest a unique approach for adding topic information into BART models to improve their ability to do abstractive summarization. The method is called the tBART. The following are the main contributions of our work:

- The tBART essentially uses the BART architecture. In this method, the latent topics are learned over sub-words instead of documents/words as in previous work. In addition, we transform the representation vector generated by the topic model to convert to context space by an align function. The topic information is added in both encode and decode processes by a general topic distribution.
- We show that the suggested model significantly outperforms a number of earlier studies on the benchmark datasets: XSUM and CNN/DAILY MAIL

II. Related works

Recently, the encoder-decoder (or "Transformers") technique of sequence-to-sequence abstractive summarization has gained widespread recognition.

The BART model' [8] is a generalized pre-training model based on the Transformer model. Token masking, phrase permutation, document rotation, token deletion, and text infilling are five pre-training approaches that are introduced. Each of these methods uses a denoising autoencoder to add noise to the original text and then restore it. In BERT, tokens are randomly masked through token masking. The sentences in a document are randomly rearranged via sentence permutation. Document rotation rotates the text so that it starts with a token chosen at random from within it. Token deletion removes a token from the initial sentence at random. Text infilling puts a mask token into a randomly chosen position or replaces word sequences with a single mask token. The most accurate method is a combination of sentence permutation and text infilling. The decoder is an autoregressive model, whereas the encoder is a bidirectional model. This pre-trained BART model is tailored to a variety of tasks, including the summarizing task, for which the encoder receives a document and the decoder produces a summary of it.

A topic augmented decoder built on an RNN-based pointer-generator network was developed by See et al. [9] in 2017 and delivers a summary dependent on the input document and the latent subjects of the document. They find that latent themes reveal more general semantic information that can be used to influence the decoder's word-generation decisions.

The ability to reflect the background impact and the implicit information passed between texts is one of the main constraints of automatic summarization. As a general extractive and abstractive model for summarization, T-BERTSum [10] was proposed. To direct the acquisition of contextual information, this uses both the BERT architecture and topic data. The model demonstrates that topic embedding is combined to produce high-quality generation in a simple and efficient manner.

Moreover, Wang and his colleagues proposed the topic assistant model (TA) [11] for the transformer-based models. They used a topic model to learn latent semantics. The latent semantics are applied as an assistant model in the training process through three modules including Semantic-informed attention (SIA), Topic embedding with masked attention (TEMA), and Document-related modulation (DRM). Since TA is a plug-and-play model that does not alter the original Transformer network's structure, it is user-friendly and compatible with a variety of Transformer-based models. Transformer+TA can be readily fine-tuned by users using a pre-trained model; TA merely adds a few extra parameters.

Although these models have shown the benefits of merging topic models and S2S learning, incorporating topic data into Transformer-based summarization algorithms is still a relatively unexplored field of research.

III. Our approach

Topic knowledge is crucial for understanding texts, as we discussed previously. However, the essential question is how to incorporate the topical information into textual representation. We provide a novel method to gently include the sub-word subject information into Transformer-based language models in order to solve this issue. The figure 1 presents our architecture's specifics.

The tBART have three main components include:

- Topic model: It has the goal that learning the latent topics.
- Representation: It consists the embedding of context, position and topic.
- Summarization: It has an abstractive summarization based on the above two components.

A. Topic model component

The "Topic model" components is the pre-processing for tBART. To learn latent topics, we use a topic model as the core of component. A word-document matrix or a bag-of-words is frequently used as the input to a basic topic modeling method like LDA [12] or NMF [13] to express the relationship between words and documents. The representation is independent of the document's word order. Or, to put it another way, the document's words are interchangeable. Moreover, there is no relationship between the documents in a corpus; they are all independent. Latent topics on a corpus can be found based on statistical methods by looking at the words used in the original texts. Words and documents are represented by topic modeling's outputs in their own latent topic spaces. In this component, the output is the relation of word and latent topics. The representation of output is the matrix $W \in R^{V \times k}$ with V - the size of vocabulary; k - the number of latent topics.

B. Representation component

To increase topic information, the modification of input embedding is necessary. In BART model, the input embedding include token embedding and position embedding. However, we specially add topic embedding in the representation of the input text. The "Representation" component is represent an input text $S = \{w_i\}_{i=1}^{n}$ with m - the internal hidden size of the transformer model n - the length of input text S. We have the token embedding and position embedding with input text S:

$$Token\ embedding = \{Ew_i\}_{i=1}^{n} \in R^{m \times n} \tag{1}$$

$$Position\ embedding = \{Ei\}_{i=1}^{n} \in R^{1 \times n} \tag{2}$$

Each sequence in the topic-based format is encoded into a topic space. The outcomes of the topic model frequently include (i) the relationship between vocabulary tokens and subjects and (ii) the relationship between learned corpus articles and topics. The relationship between vocabulary tokens and themes - W - is exploited in this study to encode input text. This decision was made after taking into account how topic information could improve the meaning of tokens. The topic information of each token is embedded into a vector whose dimension is the number of latent topics - k. Each input text is distinguished by a topic-term matrix of size $k \times n$, denoted by topic embedding where k is the number of latent topics and n is the number of tokens in each input text:

$$Et_i = W(w_i) \in R^k \tag{3}$$

$$Topic\ weight = \{Et_i\}_{i=1}^{n} \in R^{k \times n} \tag{4}$$

To have topic embedding, we apply an align function between token embedding and topic weight. With the align function, the interaction between topic and context becomes stronger. The token embedding is multiplied with topic weight to create the $Align\ weight$ by Equation 5. The matrix shows the effect of the topic and context. It is the alignment from context space to topic space and vice versa. After that, we convert $Topic\ weight$ to context space base on the align matrix by Equation 6.

$$Align\ weight = TRANS(Token\ embedding) \times Topic\ weight \tag{5}$$

$$Topic\ embedding = Topic\ weight \times TRANS(Align\ weight) \tag{6}$$

where $TRANS$ is the transpose function.

So that, the representation of an input is show such as:

$$Input\ embed = Token\ embedding + Position\ embedding + Topic\ embedding \tag{7}$$

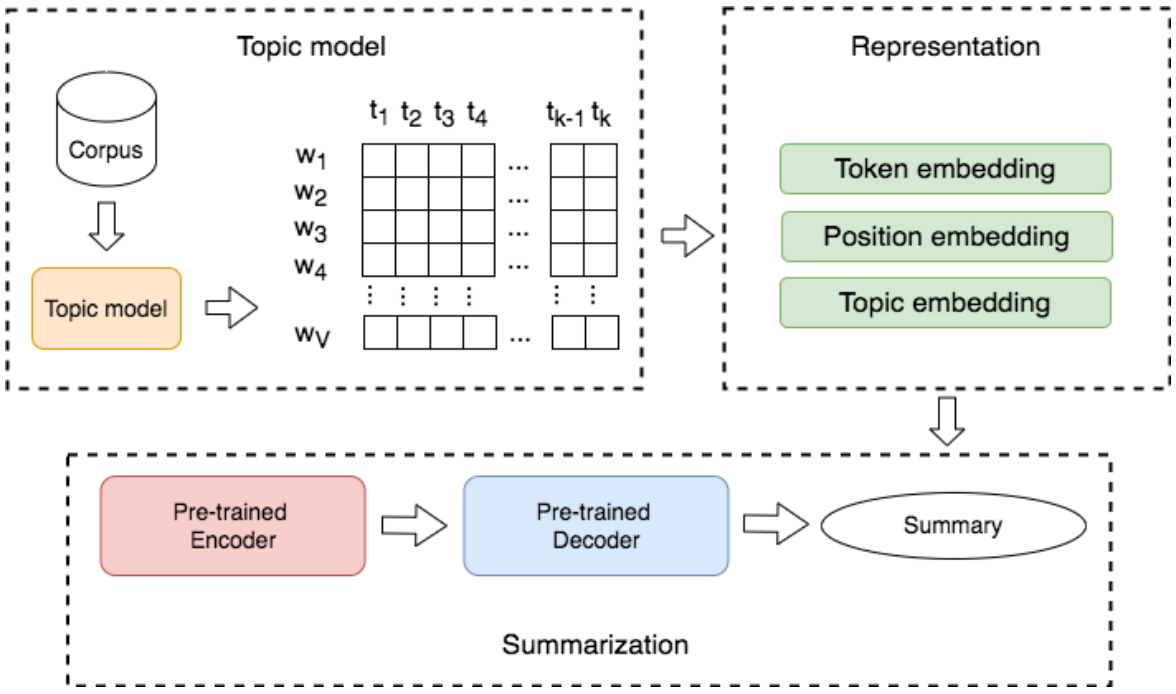

Fig. 1. The architecture of tBART

C. Summarization component

In this component, the BART is apply as the core of the component. In BART, the encoder is Bidirectional Encoder of BERT model' [14] and the decoder is Autoregressive Decoder of GPT model [15].

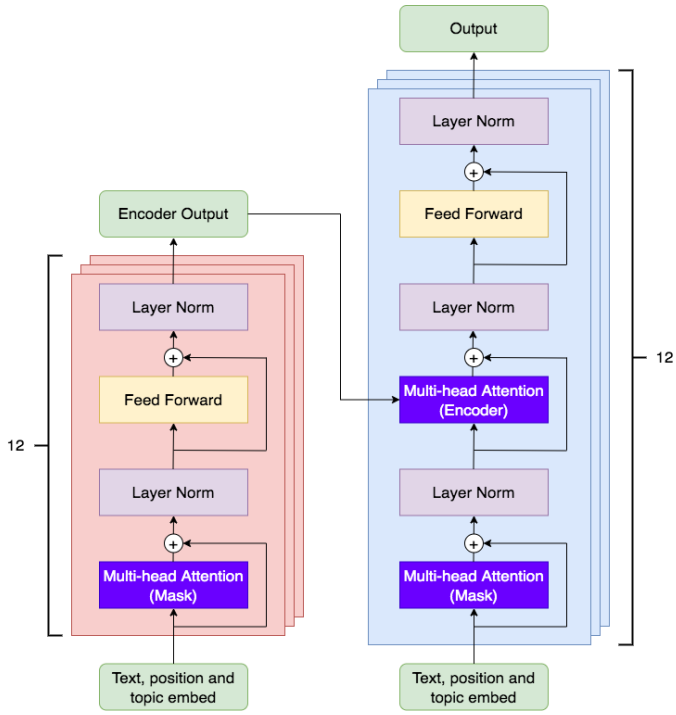

Fig. 2. Encoder - Decoder architecture

The BERT encoder outputs a vector comprising sentence-level information in addition to an embedding vector for each token in each text sequence in its input. By learning for both token- and sentence-level tasks in this way, the decoder becomes a solid starting point for any upcoming fine-tuning tasks. The previously stated and illustrated masked sequences are used for the pre-training. BART empowers the BERT encoder by using more difficult types of masking mechanisms in its pre-training while BERT was taught using a straightforward token masking technique. Each encoder layer has mask multi-head self attention layer and feed forward layer. After each step, layer norm was applied to normalize.

The GPT model's decoder utilized an architecture resembling that of the original Transformers' decoder section. GPT sequentially stacks 12 of these decoders such that changing the current token computation only affects prior tokens. Above is a picture of the architecture. The GPT decoder also employs the masked multi-headed self-attention block and a feed-forward layer, as seen in the original Transformer decoder. The multi-head attention of the transformers is chosen to help the decoder learn the soft alignment between the summary document and the original document in order to successfully decode the sequence and more accurately capture the encoded information.

In some related research as Topic Assistant [11], the supporting topic information is learned from the original document in the encoder. After that, this information is presented by vector embedding and added to the decoder. However, we used a general topic space to represent topic information in our approach. Moreover, the topic information is added to both encoder and decoder. So that, the topic information of encoder

and decoder are uniformity when the output of encoder was used in decoder. We discovered that the model can benefit from the knowledge communicated between these two jobs without significantly altering its architecture to give a more comprehensive sequence.

IV. Experimental

A. Experimental setup

We evaluate the performance of tBART on two datasets XSUM ans CNN/Daily Mail with statistics information shown in Table I.

- XSUM: It is a dataset for evaluating abstractive single-document summarizing methods. 226,711 news articles and a one-sentence summary make up the dataset. The articles span a wide range of topics and were compiled from BBC pieces published between 2010 and 2017.
- CNN/Daily Mail: The English-language CNN/DailyMail Dataset is made up of just over 300,000 unique news stories that were authored by reporters for CNN and the Daily Mail. Although the initial version was developed for automated reading, comprehension, and abstractive question answering, the current version supports both extractive and abstractive summarization.

TABLE I
THE INFORMATION OF BENCHMARK DATASETS

Dataset	train/dev/test	#avg length of doc	#avg length of summary
XSUM	204045 11332 11334	431.07	23.26
CNN/Daily Mail	287113 13368 11490	781.6	55.6

We quantitatively compare the tBART with several previous methods based on the ROUGE score(ROUGE-1, ROUGE-2, ROUGE-L). The baselines include: Transformer‘ [16]; BART [8]; BERTSum [16]; PTGEN and PTGEN+Cov [9]; T-BERTSum [10]; BERTSum+TA and BART+TA [11].

We chose the pre-trained of BART includes (i) *facebook/bart-large-cnn* ;(ii) *facebook/bart-large-xsum* to apply for “Summarization” component. The LDA is used for learning latent topics, which is better than other topic models such as GSDMM [17] as mentioned in the study on tBERT [18], SubTST [5]. We set k = 1 for the number of latent topics. The greatest summary generation suggestion is when k=1. When K is greater than 1, the model will become erratic; we have seen that word’s capacity to express many themes is insufficient. Overall, though, many themes won’t veer off-topic much, which is more like the summary document than the outcome of setting k to 0. Each word’s probability distribution across topics was determined by us, and any new document can deduce its topic distribution.

B. Experimental results

We make a comparison between the proposed method and baseline systems that is shown the results in Table II and III.

TABLE II
OUR APPROACH - TBART ON XSUM DATASET WITH ROUGE SCORE RESULTS

Methods	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	29.41	9.77	23.01
BART	45.14	22.27	37.25
BERTSum	38.81	16.50	31.27
PTGEN	29.70	9.21	23.24
PTGEN + Cov	28.10	8.02	21.72
T-BERTSum	39.90	17.48	32.18
BertSum + TA	39.77	17.39	32.39
BART + TA	45.76	22.68	38.03
tBART	**45.84**	**22.73**	**38.90**

TABLE III
OUR APPROACH - TBART ON CNN/DAILY MAIL DATASET WITH ROUGE SCORE RESULTS

Methods	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	40.21	17.76	37.09
BART	44.16	21.28	40.90
BERTSum	42.13	19.60	39.18
PTGEN	36.44	15.66	33.42
PTGEN + Cov	39.53	17.28	36.38
T-BERTSum	42.12	20.45	39.74
BertSum + TA	43.06	20.58	39.67
BART + TA	44.47	21.39	41.32
tBART	**44.55**	**21.40**	**41.61**

As shown in Table II and III, the first part of table is the baslines without topic information support. The second part is the baselines with topic information support. The last part is our approach - tBART. Overall, the tBART significantly outperforms baseline models in XSUM and CNN/Daily Mail benchmark datasets. The experimental results prominently show the effectiveness of tBART.

The tBART model outperforms conventional transformer-based models in a variety of evaluation criteria, showing that the topic may effectively collect more important details and summarize reliable material without resorting to conventional methods. No matter how our model is compared to the baselines, the score demonstrates its superiority, which suggests the need for the theme to be introduced to direct the generation.

When compared with other models with topic information support, tBART also outperforms them. The additional topic information in representation was directed to semantics in sentences. The topic is raised for all encode and decode processes. It achieves much more efficiency than just using for decoder such as Topic assistant (+ TA). With T-BERTSum, our model is higher than 3 - 6 points on the ROUGE-1 score.

The Table IV provides a few generated summaries by BART and tBART. As can be seen, topic information is used to generate some commonly overlooked words, such as “outside”

TABLE IV
COMPARISON OF ORIGINAL DOCUMENT, GOLD SUMMARY AND GENERATED SUMMARIES OF BASELINES AND OUR APPROACH

Gold summary	BART	tBART
youtube user serpentor filmed his feline friend in action footage shows the tabby producing bizarre noises as she petted the video has been seen many times.	a user filmed his feline friend in action footage shows the tabby pet producing a range of gurgling noises the show has been seen for more than 1600 times	the **youtube** user **serpentor** filmed his feline friend in action footage shows the tabby pet producing a range of gurgling noise the show has been seen for many times
A shot was reportedly fired at a car outside a primary school in Liverpool as parents were taking their children inside, police have said.	A man has been arrested on suspicion of attempted murder after a shot was fired at a car at a primary school in Liverpool.	A man has been arrested on suspicion of attempted murder after a shot was fired at a car **outside** a primary school in Liverpool

and "youtube". It demonstrated the value of subject knowledge during the generation process.

V. CONCLUSION

This chapter presents a new method for incorporating latent topic information with BART model, called the tBART. This method aims to add information and guide semantic meaning in the generation process. The experimental results show that our model outperforms all baseline models in summarization. Hence, this indicates the effectiveness of our proposed method. In addition, the tBART is built based on BART architecture, so it has more advantages in practical applications. Our work also reveals the effectiveness of latent topics in semantic tasks. In the future, we towards develop the latent topic online in the learning process and increase the quality of topic information for knowledge injection.

REFERENCES

[1] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156 043–156 070, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3129786

[2] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 9815–9822. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33019815

[3] Z. Qin, M. Thint, and Z. Huang, "Ranking answers by hierarchical topic models," in *Next-Generation Applied Intelligence*. Springer Berlin Heidelberg, 2009, pp. 103–112.

[4] M. Ovsjanikov and Y. Chen, "Topic modeling for personalized recommendation of volatile items," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2010, pp. 483–498.

[5] B. Dang, T. Dang, and L. Nguyen, "Subtst: A combination of sub-word latent topics and sentence transformer for semantic similarity detection," in *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, ICAART 2022, Volume 3, Online Streaming, February 3-5, 2022*. SCITEPRESS, 2022, pp. 91–97. [Online]. Available: https://doi.org/10.5220/0010775100003116

[6] T.-B. Dang, H.-T. Nguyen, and L.-M. Nguyen, "Latent topic refinement based on distance metric learning and semantics-assisted non-negative matrix factorization," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. Hanoi, Vietnam: Association for Computational Linguistics, Oct. 2020, pp. 70–75. [Online]. Available: https://aclanthology.org/2020.paclic-1.8

[7] G. Wu, Y. Sheng, M. Lan, and Y. Wu, "ECNU at SemEval-2017 task 3: Using traditional and deep learning methods to address community question answering task," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 365–369.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[9] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://aclanthology.org/P17-1099

[10] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "T-bertsum: Topic-aware text summarization based on BERT," *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 3, pp. 879–890, 2022. [Online]. Available: https://doi.org/10.1109/TCSS.2021.3088506

[11] Z. Wang, Z. Duan, H. Zhang, C. Wang, L. Tian, B. Chen, and M. Zhou, "Friendly topic assistant for transformer based abstractive summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 2020, pp. 485–497. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.35

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[13] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context

correlations," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1105–1114.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[15] R. Alec, N. Karthik, and S. Tim, "Improving language understanding by generative pre-training." 2018. [Online]. Available: https://s3-us-west-2. amazon-aws.com/openaiassets/researchcovers/languageunsupervised/language understanding paper

[16] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3730–3740. [Online]. Available: https://aclanthology.org/D19-1387

[17] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 233–242.

[18] N. Peinelt, D. Nguyen, and M. Liakata, "tBERT: Topic models and BERT joining forces for semantic similarity detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 7047–7055.