

TFCSG: An Unsupervised Approach for Question-retrieval Over Multi-task Learning

Shang Aiguo¹, Michael Danner², Zhu Xinjuan¹, Matthias Rätsch^{2†}

¹ College of Computer Science, Xi'an Polytechnic University, Xi'an, China
(Tel: +86-1580-299-5170; E-mail: zhuxinjuan@xpu.edu.cn, aiguoqing@gmail.com)

² ViSiR, Reutlingen University, Reutlingen, Germany
(Tel: +49-7121-271-7098; E-mail: {michael.danner, matthias.raetsch}@reutlingen-university.de)

Abstract: Most Question-answering (QA) systems rely on training data to reach their optimal performance. However, acquiring training data for supervised systems is both time-consuming and resource-intensive. To address this, in this paper, we propose TFCSG, an unsupervised similar question retrieval approach that leverages pre-trained language models and multi-task learning. Firstly, topic keywords in question sentences are extracted sequentially based on a latent topic-filtering algorithm to construct unsupervised training corpus data. Then, the multi-task learning method is used to build the question retrieval model. There are three tasks designed. The first is a short sentence contrastive learning task. The second is the question sentence and its corresponding topic sequence similarity judgment task. The third is using question sentences to generate their corresponding topic sequence task. The three tasks are used to train the language model in parallel. Finally, similar questions are obtained by calculating the cosine similarity between sentence vectors. The comparison experiment on public question datasets that TFCSG outperforms the comparative unsupervised baseline method. And there is no need for manual marking, which greatly saves human resources.

Keywords: question-retrieval, multi-task learning, topic model, contrastive learning, transfer learning, sentence representation

1. INTRODUCTION

Similar question retrieval has always been a research issue of focus in the field of natural language processing. Generally, it conducts supervised learning through tagged question-and-answer data or similar question pairs, but this method requires manual annotation of a large amount of data. One of the most challenging problems in question-answering (QA) technology is the gap between a large amount of unlabeled existing new data and the limited annotation capability available. In the actual QA system, user questions are often expressed in the form of short text. Compared with long text, the short question text contains fewer characters, the text description is more casual, and the key information of the question is difficult to extract, which brings great challenges to researchers. The previous unsupervised question-retrieval model uses traditional information retrieval techniques[1], such as lexical and semantic text-matching, query extension, etc., which often require a large amount of manual knowledge and cannot effectively deal with the above problems. Subsequently, the vector space model method was proposed[2]. The traditional space vector model carries out text vectorization based on certain features, such as the frequency of occurrence of terms or words, and tends towards high-dimensional sparsity when applied to the representation of short texts, resulting in low retrieval performance.

Recently, large language models (LLMs) based on large-scale corpus training can perform natural language processing tasks well. Compared with previous smaller pre-trained language models (PLM), LLMs

have shown the ability to learn context. The most well-known big language model is ChatGPT proposed by OpenAI/Microsoft. It doesn't need to reason and understand, only talks to people smoothly by generating answers. Nevertheless, LLMs have some drawbacks, such as the tendency to fail dialogues and generate false or fabricated information, which leads to poor performance in specific vertical field question-answering and question-retrieval tasks[3]. With the advent of deep learning, it has become an effective method for pre-trained language models based on large-scale models to construct relevant downstream tasks, such as BERT[4], RoBERTa[5], GPT-2[6], ALBERT[7], etc. These models are designed to build a powerful encoder that is capable of a comprehensive understanding of input text by learning in large corpora[8]. However, the sparsity of short questions and the anisotropy problem of PLM make the question-retrieval inefficient.

In this paper, we propose an unsupervised question retrieval approach TFCSG (An Unsupervised Question Retrieval Based on Latent Topic Filtering and Multi-task Learning). TFCSG uses the latent topic model GSDMM-Filter to extract the topic keywords in the question and takes the extracted topic words as the self-supervised labels of the question. Then, it uses the question and its sequential topic keywords to complete the construction and training of the multi-task model. In addition, we designed a multitask model containing three tasks, the first task is a short question contrastive learning task, the second task is a sequence similarity distribution task of questions and their corresponding sequential topic keywords, and the third task is a question generation sequential topic keywords task. The joint training of the three related tasks

† is the presenter of this paper.

aims to obtain greater benefits for PLM and thus improve the efficiency of the question-retrieval task.

2. RELATED WORK

2.1. Rule-based query system

Earlier question retrieval task was done to calculate the similarity between query questions and archived questions using the term-based spatial vector models TF-IDF[9] and BM25[10]. However, the high-dimensional sparse feature of vectors makes their performance poor in short-text retrieval. In order to overcome this shortcoming, Methods based on latent semantic analysis are proposed, such as LSA[11] (Latent semantic analysis), PLAS[12], etc., which translate high-dimensional sparse vectors into latent vector space and improves retrieval efficiency. However, due to the complexity and high computational cost of such algorithms, these methods have not been widely used. One of the most popular methods for implicit topic modeling is LDA (Latent Dirichlet allocation) based on generating probabilistic models[13]. LDA assumes that a document is generated by a mixture of several topics, which is not conducive to its performance in short texts. Due to the sparsity of short text, a short text is likely to contain only one topic. Based on this idea, Yin[14] proposed GSDMM (A collapsed gibbs from algorithm for dirichlet multinomial mixture model). GSDMM is often superior to the LDA model in short text and sparse text analysis[15]. Topic models can ignore noise information, effectively extract subject words, focus on the core intention of sentences, and improve the question-retrieval performance.

2.2. PLM-based query systems

In recent years, transfer learning based on pre-trained language models (PLM) has been widely used in the representation of natural language. Such models are trained on a large-scale unlabeled corpus and then fine-tuned for specific tasks, to extract and understand text feature information more comprehensively. The understanding of natural language is inseparable from the pre-training process of large corpora[17]. Common PLMs such as BERT, RoBERTa, GPT-2, ALBERT, etc. BERT is one of the most popular network pre-trained language models based on a bidirectional transformer[18] encoder. In many NLP tasks, such as question-answering, sentence classification, text representation, named entity recognition, etc., it is superior to many traditional methods. PLM can obtain richer and more effective text representation. In particular, BERT has achieved excellent results in semantic text similarity tasks. It proves that a language model based on a transformer has great potential in extracting and understanding text information. However, previous work has shown that the anisotropy problem constitutes a critical bottleneck for BERT-based sentence representation which hinders the model from fully utilizing the underlying semantic features[19], the dissimilar text representation vectors have a higher similarity score. Therefore, efforts should be made to solve anisotropy problems when

training transformer models. In particular, Bert-Flow[20] attempts to convert BERT's sentence embedding distribution into an isotropic Gaussian distribution by normalizing flows learned under unsupervised targets. In addition to the stream-based method, the whitening operation in BERT-whitening[21] also achieves good results. Specifically, the whitening operation attempts to transform the mean of sentence vectors to zero and the covariance matrix to the identity matrix. In 2021, Gao[22] proposed SimCSE, a learning framework based on sentence representation based on contrastive learning, whose core idea is to narrow the distance between similar samples and increase the distance between dissimilar samples. However, in the short-text questions, the positive examples generated using the dropout strategy are highly similar to the original questions themselves, thus limiting the effect of model training.

2.3. Chatbot-based query system

Large language models are models with more than 100 billion parameters trained in massive data, such as GPT-4[23] proposed by OpenAI-Microsoft, LaMDA[24] proposed by Google, LLaMA[25] proposed by Facebook, ERNIE3.0[26] proposed by Baidu, etc. The emergence of GPT-4 may be considered an early version of artificial general intelligence. LLMs can be applied to many tasks, such as chatbot (ChatGPT, Bard, ERNIE, Dalai, etc.), image retrieval, copywriting, translation, text generation, solving mathematical problems, etc., but there is not enough memory, computing power, and training data to solve all problems. It is currently not suitable for vertical domain information processing, because LLMs can easily generate text with uncontrollable information.

3. METHODOLOGY

Define the scenario of similar question sentence retrieval in this paper as follows: there is a question dataset $Q = \{q_1, q_2, \dots, q_n\}$, with n question sentences. And for the user input q_0 , it is necessary to find K questions similar to q_i from the question dataset Q , where $1 \leq i \leq n$. The whole retrieval process can be denoted as $q_{sim} = retrieval(q_0, Q, K)$.

In this section, we will introduce the TFCSG method, which is structured as shown in Figure 1. TFCSG consists of four modules: (1) Topic keywords extraction module(GSDMM-Filter). (2) Short question contrastive learning module. (3) Question similarity module with its sequential topic keywords. (4) Question sentence generation sequential topic keyword module. In particular, q_{stk} is the sequential topic keywords extracted in question q .

3.1. GSDMM-Filter

The topic keywords in the question sentence play a significant role in understanding the question sentence. We design the topic keyword filtering algorithm (GSDMM-Filter) based on GSDMM which is different from the previous GSDMM [14] algorithm and the GSDMM-Filter algorithm can extract the topic keywords. The GSDMM

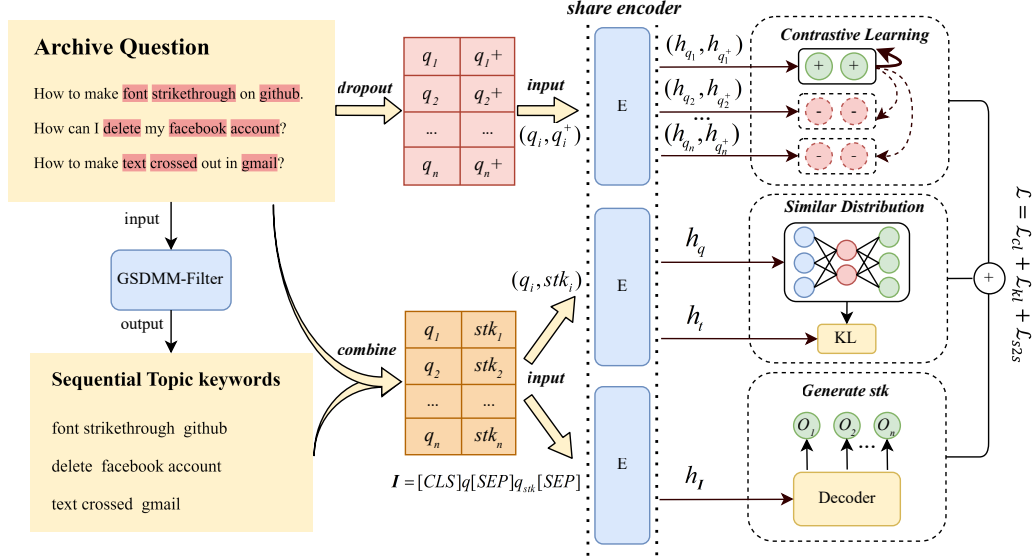


Fig. 1 . Description of the structure of the TFCSG method

The original archived questions are topic filtered to obtain the sequential topic keywords, and the training corpus is input to the shared encoder through dropout and combines operations. Finally, fine-tuning the pre-trained language model is completed by multi-task learning. In particular, in the contrastive learning module, thick arrows indicate closing the distance between positive instances, dashed lines indicate increasing the distance between negative instances, and other arrows indicate data flow direction.

automatically counts the keyword distribution of each topic and the number of times that the keywords appear in a topic when completing the interrogative information processing and clustering. Based on such features, we sequentially keep the keywords in the question that occur more frequently than t in the topic. For example, “How to make font strikethrough on GitHub?” is filtered by GSDMM-Filter as “font strikethrough GitHub”.

3.2. Contrastive Learning for questions

In recent years, contrastive learning has achieved good performance in textual representation with the concept of pulling relevant instances closer together and pushing away irrelevant instances [27]. It assumes a set of question pairs $p = \{(q_i, q_i^+)\}_{i=1}^b$, where b denotes the size of a batch of training data, and q_i and q_i^+ are semantically related. Cross-entropy is used to construct the objective function in a batch of data with negative instances, but not positive ones. Let h_i and h_i^+ be the vector representation of q_i and q_i^+ , respectively, and the objective function of b sentences in a batch of data is shown in Eq. (1):

$$\ell_{cl} = -\log \frac{\exp \text{sim}(h_i, h_i^+)/\tau}{\sum_{j=1}^b \exp \text{sim}(h_i, h_j^+)/\tau} \quad (1)$$

where τ is the temperature hyperparameter; and $\text{sim}(h_1, h_2)$ is the cosine similarity $h_1^T h_2 / (\|h_1\| \|h_2\|)$. In this work, we use the PLM to represent the input sentences, such as BERT or RoBERTa. $h = \text{PLM}_\theta(q)$, where θ is a parameter of PLM.

3.3. Similar distribution learning

The distribution of similar question vectors and the distribution of sequential topic keywords improve the training of the PLM. Specifically, $h_q = \text{PLM}_\theta(q)$,

$h_t = \text{PLM}_\theta(q_{stk})$, where q is the current input question, and q_{stk} is the sequential topic keywords generated by the GSDMM-Filter algorithm. We adopt the structure of auto-encoder[28] for similar distribution tasks, such that $z = f_{\theta_1}(h_q) = \sigma(\mathbf{W}_1 h_q + \mathbf{b}_1)$, $\hat{x} = f_{\theta_2}(z) = \sigma(\mathbf{W}_2 z + \mathbf{b}_2)$, normalize the implicit output vectors h_t and \hat{x} , i.e., $h_t = \text{softmax}(h_t)$, $\hat{x} = \text{softmax}(\hat{x})$, and define the objective function of the similarity between the question and sequential topic keywords as Eq. (2).

$$\ell_{kl} = KL(h_t || \hat{x}) = \sum_{h_t, \hat{x} \in \chi} h_t \log(h_t / \hat{x}) \quad (2)$$

where χ represents the vector space of PLM output.

3.4. Question generation sequential topic keywords

When people read documents, their brains can extract important feature information, such as core words, pictures, proper nouns, etc. As a result, we want the model to have a similar capability, so we use Seq-to-Seq[29] structure to complete the generation from question to sequential topic keywords. Splice q and q_{stk} using [CLS] and [SEP], with $\mathbf{I} = [\text{CLS}]q[\text{SEP}]q_{stk}[\text{SEP}]$ as the model’s input and $\mathbf{O} = q[\text{SEP}]q_{stk}[\text{SEP}]$ as the target sequence. Specifically, $h_{\mathbf{I}} = \text{PLM}_\theta(\mathbf{I})$, where $h_{\mathbf{I}}$ is the vector output for each word. $h'_{\mathbf{I}} = \text{decoder}(h_{\mathbf{I}})$, where $\text{decoder}(\cdot)$ is projecting $h_{\mathbf{I}}$ into the word vector space. Define the objective function of this task as Eq. (3).

$$\ell_{s2s} = -\sum_{j=1}^{|\mathbf{O}|} \mathbf{O}_j \log h'_{\mathbf{I}} \quad (3)$$

3.5. Multi-task learning

The preceding approach employs a contrastive learning framework to optimize the representation of question vectors, extracts keywords sequentially with the assistance of a topic model so that the distributions of the

original question and sequential topic keywords are similar, and finally introduces a generative task that allows the model to generate topic keyword sequences from question sentences. The efficiency of the question sentence retrieval task is improved by parallel fine-tuning the PLM through multi-task learning so that the PLM can output a high-quality representation of the question sentences. The final model training objective function is $\ell = \ell_{cl} + \ell_{kl} + \ell_{s2s}$.

4. EXPERIMENTAL RESULTS

4.1. Datasets

The experiments were carried out on two publicly available datasets, and the relevant statistics for the four datasets are shown in Table 1, where Avg/max denotes the average and maximum number of characters in the dataset, and Num denotes the number of question sentences categories.

Table 1 Dataset information statistics.

Dataset	Train	Test	Avg/max	Num
StackOverFlow	16000	4000	8.31/34	20
FAQIR	4313	1233	164.9/1083	50

4.2. Analysis of the number of topics

To explore the optimal number of topics for the unsupervised approach proposed in this paper. We calculated coherence scores[16] for a range of topics, varying from 10 to 50. The results of the coherence scores are presented in Table 2. A higher topic coherence score indicates that the generated topics are more logically reasonable. Notably, the ideal number of topics for StackOverflow and FAQIR datasets were determined to be 30 and 50, respectively. The TFCSG method will adopt the number of topics accordingly.

Table 2 Coherence scores for different number of topics.

Datasets	number of topics				
	10	20	30	40	50
StackOverFlow	0.28	0.38	0.41	0.35	0.35
FAQIR	0.39	0.39	0.38	0.40	0.42

4.3. Experimental settings

The frequency parameter t of the GSDMM-filter algorithm is set to 100. The pre-trained language models BERT-Base and RoBERTa-Base were used as shared encoders between multiple tasks with a learning rate of $3e-5$ and a training batch size of 32. The Adam optimizer was used to optimize the training, and the pre-trained models were output with their default dimension of 768. A single NVIDIA Quadro P5000 16GB was used to train the multi-tasking model with maximum lengths of 32, and 256 on StackOverflow and FAQIR datasets, respectively, and the epoch was set to 10 on the datasets.

4.4. Baselines

We select baseline models related to advanced unsupervised methods, such as SimCSE. The baseline models used in this paper are summarized below:

BERT-Base[4]: BERT is a very important model in the domain of natural language processing, and its training consists of two main tasks, MLM and NSP. In the retrieval task, we use its output CLS vectors to represent the sentence vectors.

RoBERTa-Base[5]: RoBERTa is a BERT model optimization that removes the NSP task of BERT and replaces it with a larger amount of intra-batch data and a dynamic mask mechanism.

BERT-Flow[20]: BERT-Flow is a method to reversibly map the output space of BERT from a cone to a standard Gaussian distribution space.

BERT-Whitening[21]: BERT-Whitening converts all BERT sentence vectors into vectors with mean 0 and covariance matrix as the unit matrix, i.e., it performs whitening on the output vectors to enable efficient computation of similarity between sentence vectors.

SimCSE[22]: SimCSE is a training method that uses contrast learning and simple data augmentation.

4.5. Results and Analysis

The experimental results of StackOverflow and FAQIR are shown in 3. Among the two transformer-based models, BERT-Base and RoBERTa-Base, the retrieval efficiency of the Roberta-base model is significantly higher than that of the BERT-base model, the reason for which is that the RoBERTa-base model is more compatible and more closely related to the one sentence task. On the StackOverflow dataset, TFCSG outperforms our compared baseline method in five metrics, P@1, P@5, P@10, MAP, and MRR, and surpasses unsupervised SimCSE-BERT-Base by 8.9%, 15.8%, 19.7%, 8.8%, and 7.9% in the five metrics, respectively. It can be observed that the scores of P@1, P@5, and P@10 are closer because TFCSG constitutes a topic keyword-based method, and similar question sentences of different lengths are mapped onto more similar representations under the guidance of topic words. Essentially, the model can learn the crucial components of the sentences through the topic keywords sequence, reducing the effect of noise, and thus similar question sentences are more likely to be retrieved and ranked first. The performance of TFCSG on the FAQIR dataset is similar to the previous one, which shows the effectiveness of our proposed method.

Additionally, we compared the proposed method in the paper with different-sized training datasets based on the P@1 score. The results in Table 4 demonstrate that our method achieves higher accuracy than the baseline method SimCSE with a smaller amount of data.

5. CONCLUSION

Aiming to achieve a marked improvement over the traditional question-retrieval model which requires a lot of manual tagging, an unsupervised question retrieval ap-

Table 3 Experimental results on the StackOverflow and FAQIR dataset.

Model	StackOverFlow					FAQIR				
	P@1	P@5	P@10	MAP	MRR	P@1	P@5	P@10	MAP	MRR
BERT-Base	0.510	0.409	0.355	0.542	0.625	0.684	0.614	0.554	0.614	0.774
BERT-Base-flow	0.642	0.605	0.512	0.682	0.701	0.701	0.651	0.591	0.757	0.814
BERT-Base-whitening	0.664	0.598	0.552	0.694	0.742	0.717	0.652	0.574	0.762	0.804
SimCSE-BERT-Base	0.727	0.656	0.612	0.744	0.767	0.741	0.701	0.624	0.795	0.801
TFCSG-BERT-Base	0.816	0.814	0.809	0.832	0.846	0.837	0.824	0.821	0.851	0.861
RoBERTa-Base	0.623	0.541	0.463	0.607	0.666	0.711	0.654	0.604	0.761	0.779
RoBERTa-Base-flow	0.702	0.664	0.634	0.742	0.805	0.754	0.684	0.641	0.781	0.824
RoBERTa-Base-whitening	0.712	0.662	0.641	0.752	0.773	0.745	0.671	0.654	0.794	0.851
SimCSE-RoBERTa-Base	0.776	0.714	0.684	0.809	0.838	0.791	0.722	0.681	0.834	0.856
TFCSG-RoBERTa-Base	0.832	0.822	0.812	0.854	0.882	0.851	0.849	0.845	0.872	0.901

Table 4 P@1 on training data of different sizes.

Model (BERT-base)	partition of training data		
	10%	30%	50%
SimCSE	0.231	0.358	0.510
TFCSG	0.244	0.533	0.694

proach TFCSG is proposed in this paper. We designed a GSDMM-filtering model to extract the topic key information in questions, constructed a similar task of question distribution and topic keyword sequences distribution, and alleviated the problem of dissimilarity in the representation of synonymous questions of different lengths. In order to alleviate the anisotropy problem of the pre-trained text representation model, we use a contrastive learning framework to get representation vectors. In order to overcome the problem that the short-text questions, we incorporated the task of generating topic keyword sequences using questions to complicate the training process of the model, so that the model can understand the short-text questions to a greater extent. Finally, we trained the three tasks in parallel, and the experimental results demonstrated that TFCSG was superior to the compared unsupervised baseline models. In particular, TFCSG retrieves similar questions much further ahead than the other baseline methods.

Acknowledgment This work is partially supported by the National Key Research and Development Program of China (2019YFC1521405), Graduate Scientific Innovation Fund for Xi'an Polytechnic University (chx2022023)

REFERENCES

- [1] M. Karan, J. Šnajder, “Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval”, *Expert Systems with Applications*, Vol. 91, pp. 418–433, 2018.
- [2] F. Günther, L. Rinaldi, M. Marelli, “Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions”, *Perspectives on Psychological Science*, Vol. 14, No. 6, pp. 1006–1033, 2019.
- [3] Floridi, Luciano, “AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models”, *Philosophy and Technology*, pp. 15–36, 2023.
- [4] J. Devlin, M. Chang, K. Lee, et al, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint*, 2018.
- [5] Y. Liu, M. Ott, N. Goyal, et al, “Roberta: A robustly optimized bert pretraining approach”, *arXiv preprint*, 2019.
- [6] Y. Yang, Y. Li, X. Quan, “Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2”, *International Conference of AAAI*, pp. 14230–14238, 2021.
- [7] Z. Lan, M. Chen, S. Goodman, et al, “Albert: A lite bert for self-supervised learning of language representations”, *arXiv preprint*, 2019.
- [8] Y. Zou, H. Liu, T. Gui, et al, “Divide and Conquer: Text Semantic Matching with Disentangled Keywords and Intents”, *International Conference of ACL*, pp. 3622–3632, 2022.
- [9] Aizawa, Akiko, “An information-theoretic perspective of tf-idf measures”, *Information Processing & Management*, Vol. 39, No. 1, pp. 45–65, 2003.
- [10] S. Robertson, H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond”, *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.
- [11] S. Deerwester, S. Dumais, G. Furnas, et al, “Indexing by latent semantic analysis”, *International Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.
- [12] Hofmann, Thomas, “Probabilistic latent semantic analysis”, *arXiv preprint*, 2013.
- [13] D. Blei, A. Ng, M. Jordan, “Latent dirichlet allocation”, *International Journal of Journal of Machine Learning Research*, pp. 233–242, 2014.
- [14] J. Yin, J. Wang, “A dirichlet multinomial mixture model-based approach for short text clustering”, *International Conference of SIGKDD*, Vol. 3, No. 1, pp. 993–1022, 2003.
- [15] C. Weisser, C. Gerloff, A. Thielmann, et al, “Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using Twitter data”, *International Journal of Computational Statistics*, pp. 1–28, 2022.
- [16] S. Blair, Y. Bi, M. Mulvenna, “Aggregated topic

- models for increasing social media topic coherence”, *International Journal of Applied Intelligence*, Vol. 50, pp. 138–156, 2020.
- [17] H. Zhu, P. Tiwari, A. Ghoneim, et al, “A collaborative ai-enabled pretrained language model for aiot domain question answering”, *International Journal of IEEE Transactions on Industrial Informatics*, Vol. 18, No. 5, pp. 3387–3396, 2021.
 - [18] A. Vaswani, N. Shazeer, N. Parmar, et al, Łukasz and Polosukhin, Illia, “Attention is All you Need”, *International Conference of Neurips*, Vol. 30, pp. 5998–6008, 2017.
 - [19] Ethayarajh, Kavin, “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”, *International Conference of EMNLP*, pp. 55–65, 2019.
 - [20] B. Li, H. Zhou, J. He, et al, “On the sentence embeddings from pre-trained language models”, *arXiv preprint*, 2022.
 - [21] J. Su, J. Cao, W. Liu, et al, “Whitening sentence representations for better semantics and faster retrieval”, *arXiv preprint*, 2021.
 - [22] T. Gao, X. Yao, D. Chen, “SimCSE: Simple Contrastive Learning of Sentence Embeddings”, *International Conference of EMNLP*, pp. 6894–6910, 2021.
 - [23] OpenAI, “GPT-4 Technical Report”, *arXiv preprint*, 2023.
 - [24] R. Thoppilan, D. Freitas, J. Hall, “Lamda: Language models for dialog applications”, *arXiv preprint*, 2022.
 - [25] H. Touvron, T. Lavril, G. Izacard, et al, “Llama: Open and efficient foundation language models”, *arXiv preprint*, 2023.
 - [26] Y. Sun, S. Wang, S. Feng, et al, “Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation”, *arXiv preprint*, 2021.
 - [27] R. Hadsell, S. Chopra, Y. Lecun, “Dimensionality reduction by learning an invariant mapping”, *International Conference of CVPR*, Vol. 2, pp. 1735–1742, 2006.
 - [28] F. Feng, X. Wang, R. Li, “Cross-modal retrieval with correspondence autoencoder”, *International Conference of ACM Multimedia*, pp. 7–16, 2014.
 - [29] L. Dong, N. Yang, W. Wang, et al, “Unified Language Model Pre-training for Natural Language Understanding and Generation”, *International Conference of Neurips*, Vol. 32, pp. 13063–13075, 2019.