Mongolian-Chinese Cross-lingual Topic Detection Based on Knowledge Distillation

Yanli Wang, Yatu Ji, Baolei Sun, Qing-Dao-Er-Ji Ren, Nier Wu,
Na Liu, Min Lu, Chen Zhao, Yepai Jia
Inner Mongolia University of Technology
Hohhot, China
{MLjyt, wunier04, csnaliu, cslumin}@imut.edu.cn,
{2834267714, bl.sun, 854766886, 981627837}@qq.com,
15847146757@163.com

***Abstract*—Most topic detection research focuses on cross-language information processing in languages with abundant resources, such as English-Chinese and English-German, while there are also some studies that focus on low-resource languages, for instance, Vietnamese-Chinese and Tibetan-Chinese cross-lingual topic detection. However, there is a scarcity of research on Mongolian-Chinese cross-lingual topic detection, mainly due to the scarcity of Mongolian language data and the limitations of traditional topic detection in Mongolian text representation, clustering, and topic representation. Therefore, this paper proposes a Mongolian-Chinese cross-lingual topic detection method based on knowledge distillation. First, the knowledge distillation method is utilized to integrate Mongolian-Chinese cross-language semantics, enabling Mongolian news to be mapped into the Chinese semantic space with rich semantics through a pre-trained Chinese language model. Then, during the fine-tuning process of the original model, a Parameter Efficient Fine Tuning (PEFT) module is added to prevent catastrophic forgetting of the pre-trained language model. Finally, a density-based hierarchical clustering algorithm is used to cluster the news texts. The experimental results show that the proposed method using the RoBERTa-WWM pre-trained model combined with the density-based hierarchical clustering algorithm has improved at least 6 percentage points in the F1 score and 2 percentage points in topic coherence evaluation metrics compared to other baseline models. Furthermore, after adding the PEFT module, the model has improved at least 4 percentage points in F1 score, 3 percentage points in topic coherence, and 3 percentage points in topic diversity, respectively. The experiments prove that this method can effectively improve the accuracy of Mongolian-Chinese cross-lingual topic detection.**

***Index Terms*—Mongolian-Chinese topic detection, pre-trained language model, knowledge distillation**

I. Introduction

Topic detection technology provides the discovery of new information and focuses on specific hot topics through the detection and tracking of targeted topics [1]. Through steps such as data collection, preprocessing, relevance analysis, and hot topic tracking, relevant content is automatically clustered, and the development of news events is tracked, providing users with the trajectory and trends of event evolution. Current research on topic detection primarily focuses on the monolingual domain and has resulted in three mainstream research methods: methods based on topic models, text feature clustering, and pre-trained language models. Cross Language Topic Detect and Track (CLTDT) involves clustering analysis of news reports in different languages to discover and summarize related topics. Currently, CLTDT methods are mainly divided into two categories: translation-based methods and utilizing bilingual dictionaries or parallel corpora to train bilingual word embeddings. Parameter Efficient Fine Tuning (PEFT) is a fine-tuning strategy for large pre-trained models, aiming to adapt the model to the needs of specific tasks or domains through a small number of training steps while preserving most of the parameters. Knowledge distillation techniques can transfer the knowledge from a complex neural network model to a smaller, more lightweight neural network model, enhancing the model's generalization and robustness.

The research on Mongolian-Chinese cross-language topic detection helps public opinion regulatory departments promptly obtain special and sensitive hot topics from vast amounts of Internet data, enabling effective supervision, dynamic tracking, and result analysis of these topics. However, Mongolian-Chinese cross-language topic detection faces some challenges. First, the scarcity of Mongolian data means that there is a limited amount of Mongolian language data for training and evaluation, which has a certain impact on the model performance and generalization ability. Second, Mongolian-Chinese cross-language topic detection needs to address language differences, as Mongolian and Chinese have significant differences in grammar, vocabulary, and semantics, increasing the complexity of text representation and similarity calculation. Additionally, imbalanced data distribution and translation alignment are also challenges that need to be overcome. To address the aforementioned issues, this paper proposes a Mongolian-Chinese cross-language topic detection method based on knowledge distillation. First, a pre-trained semantic encoder is fine-tuned using Mongolian-Chinese parallel corpora through knowledge distillation, enabling the encoder to integrate cross-language semantic information and map the semantic encodings of Mongolian and Chinese into the same semantic space, thus improving the semantic encoding capabilities for both languages. Then, to better preserve the original model's semantic representation ability for Chinese and enable the model to have cross-language semantic fusion capabilities, a PEFT module is introduced into the student

979-8-3315-4085-2/24/$31.00

model. Finally, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is utilized to perform clustering operations on news texts.

The main contributions of this paper are as follows :

1.To address the issue of traditional topic detection methods struggling to extract semantic information from Mongolian-Chinese cross-lingual news texts due to the scarcity of Mongolian language data, this paper proposes a knowledge distillation approach that integrates Mongolian-Chinese cross-lingual semantics. This allows Mongolian news to be mapped into the rich semantic space of Chinese through Chinese pre-trained language models, enhancing the performance of Mongolian-Chinese cross-lingual topic detection.

2.To address the issue of catastrophic forgetting during the fine-tuning process, we introduce the PEFT module. By freezing the parameters outside the PEFT module, we ensure that the pre-trained language model retains its original semantic representation capabilities while enhancing training efficiency.

3.To address the issue of traditional clustering algorithms like k-means requiring a predefined number of categories, this paper introduces the HDBSCAN algorithm for topic clustering. The algorithm not only does not require specifying the number of clusters beforehand but also has the advantages of being insensitive to noise, being suitable for data with multi-density distributions, and having low computational complexity. It is more suitable for the current distribution of Mongolian news data.

4.Experiments have demonstrated that the model achieved improvements over other baseline models on various evaluation metrics.

II. RELATED WORK

A. Topic detection

The Topic Model is a special kind of Probabilistic Graphical Model. It plays a pivotal role in topic detection, whose primary objective is to identify and extract the core issues or themes within a text. By mining the distribution of topics in documents, the Topic Model provides an effective tool for topic detection.

The most widely used topic detection methods currently are mostly based on the Latent Dirichlet Allocation (LDA) topic model [2] or its improved models [3]. LDA is an unsupervised machine learning technique that can be used to identify potential hidden topic information in large-scale document collections or corpora. This method assumes that each word is drawn from an underlying, latent, hidden topic. LDA relies on the Bag of Words (BOW) assumption, representing text as an unordered collection of words while ignoring word order and contextual information. Due to the inherent limitations of the BOW model, these methods tend to perform better when processing long texts but relatively worse when dealing with short texts. Yan et al. [4] proposed the Biterm Topic Model (BTM), which is an improvement on the LDA model. BTM leverages word co-occurrence relations to enrich phrases, thus alleviating the sparsity issue of short texts. This model has achieved a certain improvement in the topic detection effect for short texts. However, the detection performance of the model is poor when text features are too sparse or there is noise and polysemy.

The topic detection method based on text feature clustering divides multiple news texts into different clusters through clustering, assuming that the texts in each cluster express the same topic. Sayyadi et al. [5] proposed the Key Graph topic detection model based on keyword co-occurrence. This model argues that keywords in a document have a greater ability to represent the topic to which the document belongs and assumes that frequently co-occurring keywords in a document tend to describe the same topic. Therefore, the model first extracts keywords from each document and constructs a graph based on the co-occurrence frequency of keywords in a window within the corpus. The nodes in this graph are keywords, and the weights on the edges represent the co-occurrence strength, measured by Pointwise Mutual Information (PMI). Subsequently, the GN community detection algorithm [6] is applied to this graph, and the resulting communities are considered topics. Finally, the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is used to represent the topic and document vectors, and documents are assigned to the closest topic category based on the cosine distance between the vectors.

In recent years, pre-trained language models have gradually been applied to topic detection technology. Asgari-Chenaghlu et al. [7] proposed the TopicBERT model, which combines BERT (Bidirectional Encoder Representations from Transformers), graph strategies, and a multi-modal entity recognizer, for topic detection in social media. Asgari-Chenaghlu et al. [8] also experimented with using the Transformer structure to encode tweets, calculating text similarity based on the text encoding, and inputting the results into the K-means algorithm for clustering to extract topics related to the COVID-19 pandemic from Twitter tweets.

B. Cross language topic detect and track

Current research on CLTDT mainly relies on methods such as machine translation, bilingual dictionary translation, and bilingual parallel corpora. The core challenge of CLTDT lies in accurately determining the textual relevance of two news reports in a multilingual environment. This challenge involves the representation of news report texts and the computation of their similarity.

Leek et al. [9] adopted the method of machine translation to translate texts in different languages into the same language for computation. Yang et al. [10] utilized a probabilistic topic model to extract topic words from texts and calculated the similarity of topic distributions across different language texts by translating the topic words for clustering. While the machine translation method has achieved certain results, its performance in CLTDT models declines significantly in low-resource language scenarios due to limitations in translation accuracy.

Mathieu et al. [11] and Pouliquen et al. [12] proposed using bilingual dictionary translation instead of machine translation,

translating the words in news texts through Chinese-English dictionary translation, and comparing the similarity of Chinese and English news texts based on the correspondence of named entities (such as people, places, and organizations). This approach significantly improved the system's detection accuracy. Chang et al. [13] utilized Wikipedia to construct a bilingual dictionary and mined co-occurring topics under Chinese and English news events. Mimno et al. [14] proposed to establish connections between cross-language texts by leveraging the hypothesis that the topic distributions of comparable multilingual corpora are similar, thus achieving research on cross-language topic tasks. Hao et al. [15] introduced the concepts of hard links and soft links into traditional probabilistic topic models and established bilingual connections through parallel aligned documents and bilingual dictionaries to achieve topic clustering tasks. Hong et al. [16] established alignment relationships between news elements based on the unique features of news, using a bilingual dictionary, and clustered news texts through graph clustering, achieving good clustering results. All these methods effectively address the issue of bilingual language differences. While bilingual dictionaries excel at improving cross-language word semantic alignment, they are limited by the size of the dictionary. In low-resource language scenarios, it is difficult for bilingual dictionaries to match all the feature words in news, and they cannot solve issues such as polysemy.

Inspired by this, an increasing number of researchers have begun to explore the use of deep neural networks for CLTDT methods. Bianchi et al. [17] employed a multilingual word embedding approach, training a multilingual BERT, and utilizing a variational auto-encoder to fuse the multilingual BERT to predict the topics of multilingual news articles. This effectively addressed the cross-language language difference issue. The above methods can achieve good results when there are sufficient data resources, such as those between Chinese and English. However, in the case of scarce Mongolian-Chinese data resources, due to the sparsity of parallel corpora, it is difficult to construct bilingual dictionaries and train bilingual word embeddings, making it challenging to integrate Mongolian and Chinese news texts into the same semantic space.

III. Mongolian-Chinese Cross-Language Topic Detection Based on Knowledge Distillation

This paper first utilizes the knowledge distillation method to integrate Mongolian-Chinese cross-lingual semantics, enabling Mongolian news to be mapped into a rich Chinese semantic space through a pre-trained Chinese language model. Then, the PEFT module is introduced into the student model to preserve the original model's semantic representation capabilities for Chinese, thereby enabling the model to possess cross-lingual semantic fusion capabilities. Finally, HDBSCAN is utilized to achieve topic detection for Mongolian and Chinese news. The overall structure is illustrated in Figure 1.

Due to the scarcity of Mongolian corpus resources, which cannot directly support the training of large-scale language models, and the ineffectiveness of topic detection based on

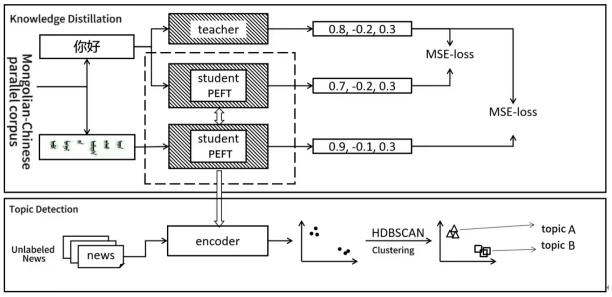

Fig. 1. Overall structure of Mongolian-Chinese cross-language topic detection.

topic models and text features in a multilingual environment, this paper first improves the training structure of knowledge distillation by fine-tuning a Chinese pre-trained language model to achieve better Mongolian semantic representation capabilities. As show in Figure 2, unlike the traditional knowledge distillation training process, this paper adopts the same Chinese pre-trained language model as the foundation for both the student model and the teacher model. While the teacher model encodes Chinese, the student model encodes Mongolian, ensuring that the student model can integrate Mongolian semantic expression capabilities into the vector space of Chinese semantic representation.

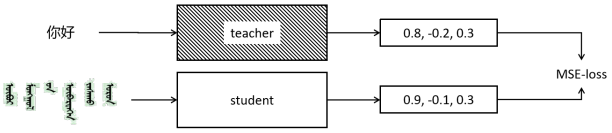

Fig. 2. Improved knowledge distillation.

To preserve the Chinese semantic representation capabilities of the pre-trained language model, a cross-language knowledge distillation training task is proposed based on the improved knowledge distillation training. The aims to enable the model to learn Mongolian semantic representation while maintaining its Chinese semantic representation capabilities. As shown in Figure 3.

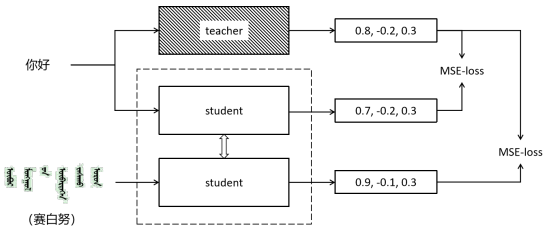

Fig. 3. Cross-language knowledge distillation.

The teacher model M is a pre-trained language model RoBERTa-WWM, which maps one or more Chinese sentences into a dense vector space. Using Mongolian-Chinese parallel corpora $\{(s_1, t_1), ..., (s_n, t_n)\}$ to train the

student model $\hat{M}$,the goal is to make $\hat{M}(s_i) \rightarrow M(s_i)$ while also $\hat{M}(s_i) \rightarrow M(t_i)$. The trained student model will possess the capabilities of a multilingual encoder and be able to learn richer semantic information. The loss function for a given mini-batch of model training is similar to MSEloss.

$$L = \frac{1}{|B|} \sum_{j \in B} \left[\left(M(s_j) - \widehat{M}(s_j) \right)^2 + \left(M(t_j) - \widehat{M}(s_j) \right)^2 \right] \quad (1)$$

By minimizing the loss function, the model parameters are adjusted to gradually learn the appropriate parameter configuration, enabling the learning model to encode Mongolian into a high-dimensional space corresponding to Chinese semantics. Furthermore, the knowledge distillation process helps the model better capture the semantic relationship between Mongolian and Chinese, enabling the model to more accurately understand and represent topic information across different languages in cross-language topic detection tasks. By encoding Mongolian into a high-dimensional space corresponding to Chinese semantics, the model can more effectively process cross-language data and improve the accuracy and generalization ability of topic detection. The advantage of this approach lies in its ability to overcome the challenges posed by linguistic differences and data scarcity.

The pre-trained language model used in this paper is RoBERTa-WWM, which is utilized to build a Mongolian semantic encoder with strong semantic understanding capabilities. Through fine-tuning the pre-trained language model, it can effectively capture the deep information contained in Mongolian texts. In the fourth part of this paper, it is compared with several Chinese pre-trained language models trained on large-scale Chinese data, testing the influence of pre-trained language models on cross-language topic detection.

Addressing the linguistic differences between Mongolian and Chinese, the PEFT module is introduced into the student model based on an improved knowledge distillation training structure, aiming to better preserve the original model's semantic representation capability for Chinese and thus enabling the model to possess cross-lingual semantic fusion capabilities.

The HDBSCAN clustering algorithm has significant advantages in topic detection. Firstly, it can automatically discover the cluster structure present in the data without needing to specify the number of clusters beforehand, making it more flexible. Secondly, HDBSCAN can effectively handle noisy data and outliers, improving the accuracy of topic detection. In addition, by adjusting parameters such as the minimum number of samples and density threshold, the tightness of clusters can be flexibly controlled, enhancing adaptability to data with different densities. Finally, the HDBSCAN algorithm is highly robust and efficient, capable of processing large-scale datasets and completing clustering tasks in a relatively short time, making it suitable for the need to quickly process a large amount of news data and identify topics. Combining the dynamic and diverse characteristics of news data with the advantages of the HDBSCAN algorithm, this paper utilizes the HDBSCAN algorithm to perform clustering operations on news texts.

IV. Experiments

A. Data

The dataset used in the experimental section of this paper is the 118,502 sentence pairs of Mongolian-Chinese parallel corpus after deduplication and correction from CWMT. These corpora cover a wide range of categories, including politics, economy, culture, entertainment, and more. Among the parallel corpora, 80% is used for knowledge distillation training, while the remaining 20% serves as a validation set to evaluate the model during training.The core processing target for Mongolian-Chinese cross-language topic detection is news data. Therefore, a Mongolian-Chinese cross-language news corpus was constructed using web crawlers, including 978 Mongolian news articles and 1,132 Chinese news articles. Other news statistics are shown in Table I.

TABLE I
Statistics of News Corpus Information

	Mongolian	Chinese
Corpus Size	10.4MB	12.6MB
Number of Texts	978	1132
Total Word Count	296875	575823
Number of Unique Words	22217	9782
Average Number of Words per Text	303.5	508.7

B. Baseline and evaluation indicators

Baseline.

- The Improved Chinese-English Latent Dirichlet Allocation model (ICE-LDA) [18]: This model utilizes the Bi-LDA probabilistic topic model to derive the topics of news articles. These topics are then vectorized and mapped into the same semantic space through translation.
- Cross-Language Text Clustering Algorithm based on Latent Semantic Analysis (CLTC-LSA) [19]: This algorithm employs latent semantic analysis to extract feature words from news articles. It then utilizes the correlation of these feature words to construct a Mongolian-Chinese bilingual semantic space, enabling clustering of Mongolian and Chinese news texts.
- The Generalized Vector Space Model (GVSM) [16]: This model utilizes the alignment of Mongolian and Chinese news entities and the co-occurrence relationship in the context to calculate the similarity between elements based on bipartite graphs, thereby clustering the texts.
- Cross-Language Neural Topic Model (CL-NTM) [20]: This model trains neural topic models based on variational autoencoders for Chinese and Mongolian separately, obtaining monolingual abstract representations of topics. Then, it leverages a small-scale parallel corpus to map bilingual topics into the same semantic space. Finally, it uses the K-means method to cluster the bilingual topic representations, thus discovering topics within clusters of news events.

Evaluation indicators.

The main experimental metrics in this paper are Macro-F1 (denoted as F1), topic coherence, and topic diversity.

- Macro-F1:It is typically used as an experimental metric to evaluate the clustering results of baseline models, encompassing Macro-Precision and Macro-Recall . The specific definitions are as follows:

$$P = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TPi + FP_i} \tag{2}$$

$$R = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Where TP_i represents the number of documents correctly classified into the i-th cluster, the number of correctly clustered documents; FP_i represents the number of documents wrongly classified into the i-th cluster; FN_i represents the number of documents belonging to the i-th cluster but wrongly classified into other clusters; F1 is an overall evaluation of the clustering performance across all clusters.

- The Topic Coherence: The evaluation metric [21] aims to assess whether a topic model can generate semantically consistent and easily understandable topics. This coherence measure has been proven to simulate human judgment with reasonable performance [22]. The metric ranges from [-1, 1], where 1 represents perfect coherence. The method commonly used to evaluate the topic coherence of a baseline model is Normalized Pointwise Mutual Information (NPMI) .
- The Topic Diversity [23]: The percentage of unique words within the first 25 words of all topics. This metric ranges from [0, 1], where 0 represents a redundant topic and 1 indicates a more diversified topic.

C. Parameter settings

Model parameter settings

During the knowledge distillation training process, the number of iterations was set to num_epoch=100, batch_size = 32, num_warmup_steps=1000. The model was evaluated on the validation set every 1000 steps. RoBERTa-WWM was used as the pre-trained model, which consists of 12 layers of Transformer-Encoder and outputs 768-dimensional text vectors. The maximum text length was set to 512, the learning rate was set to 5E-5, the number of training epochs was set to 10, the similarity threshold was set to 0.88, and the time decay factor was set to 1E-7.

Baseline model parameter settings

- For the ICE-LDA model, the number of clustering topics K is set to 20, the prior parameter a is set to 0.5, the prior parameter b is set to 0.01, and the number of sampling iterations is set to 1000. The translation interface adopted in this paper is the Youdao online translation interface.
- For the CLTC-LSA model, this paper refers to the original model parameter settings and, considering the relatively small number of topics in the current dataset, sets the number of clustering topics K to 20, retaining 15 feature words for each text.
- For the GVSM model, based on the conclusions of the original paper, this paper specifies the parameters as $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$.
- For the CL-NMT model, the word embedding dimension in the variational autoencoder is set to 300, the topic vector is set to 20 dimensions, the training batch size is set to 100, Adam is used as the optimization function, the learning rate is set to 0.002, gradient clipping is used to prevent gradient explosion, and the input BOW model is normalized. The K value in K-means clustering is set to 20.

D. Experiment results and analysis

As shown in Table II, RoBERTa-WWM, as a pre-trained language model, scores higher compared to other models, as RoBERTa pre-training tasks are more relevant to the topic detection task. HDBSCAN significantly improves topic consistency and diversity while ensuring sufficient accuracy in clustering. Through experimental analysis, it has been found that pre-trained language models can map news into a high-dimensional vector space, where the distribution of news samples is more suitable for the HDBSCAN clustering algorithm.

TABLE II
COMPARISON OF TOPIC DETECTION PERFORMANCE ACROSS DIFFERENT MODULES

Pre-trained Language Model	Clustering Algorithm	F1	NPMI	Diversity
ALBERT	K-means	0.733	0.128	0.657
	DBSCAN	0.779	0.141	0.712
	HDBSCAN	0.786	0.154	0.802
MacBERT	K-means	0.749	0.121	0.840
	DBSCAN	0.791	0.149	0.722
	HDBSCAN	0.803	0.165	0.850
RoBERTa-WWM	K-means	0.755	0.130	0.675
	DBSCAN	0.782	0.153	0.732
	HDBSCAN	**0.805**	**0.170**	**0.854**

The results of the comparison experiment with baseline models are shown in Table III. The CLTC-LSA method among the following models is a non-probabilistic topic model. It mainly measures the similarity between news texts by calculating the semantic similarity between words, thus performing document clustering. However, in the Mongolian-Chinese cross-language topic detection task, due to poor translation quality, CLTC-LSA is unable to accurately calculate the similarity between Chinese and Mongolian words, resulting in poor alignment of the Mongolian semantic space, which further affects the topic detection effect. In contrast, the topic detection effect of ICE-LDA outperforms CLTC-LSA, as ICE-LDA employs a probabilistic topic model that can better extract topics from news texts. However, ICE-LDA's approach relies on translated news topic keywords to establish bilingual connections, which is also heavily influenced by the performance of translation tools under low-resource Chinese-Vietnamese

conditions, resulting in poor cross-language topic detection results. Regarding the clustering effect of the GVSM (EUB) method, it depends on the number of annotated news entities and is only suitable for clustering texts, failing to adequately express the core topics of such texts. In contrast, the CL-NMT method also utilizes pre-trained language models, resulting in relatively satisfactory performance. However, the method proposed in this paper adopts pre-trained language models combined with knowledge distillation, successfully integrating Mongolian-Chinese cross-language semantics, thereby achieving better clustering results and being less affected by the scarcity of Mongolian-Chinese resources.

TABLE III
THE COMPARATIVE EXPERIMENTAL RESULTS OF CROSS-LINGUAL TOPIC DETECTION BETWEEN MONGOLIAN AND CHINESE USING DIFFERENT METHODS

model	F1	NPMI	Diversity
ICE-LDA	0.632	0.023	0.512
CLTC-LSA	0.598	0.023	0.608
GVSM	0.634	0.041	0.714
CL-NMT	0.716	0.052	0.793
Based on the PEFT method	**0.759**	**0.055**	**0.832**

V. CONCLUSION

Due to the scarcity of Mongolian-Chinese parallel corpora, there is no corresponding pre-trained language model available, and there is also a dearth of research on Mongolian-Chinese cross-language topic detection. Moreover, traditional topic detection methods have certain limitations in Mongolian. Therefore, this article first improves upon traditional knowledge distillation and further proposes an improved cross-language knowledge distillation method. Secondly, a parameter-efficient fine-tuning method is introduced into the model, which can not only allow the model to retain the ability to represent Chinese semantics but also prevent catastrophic forgetting of the model. Finally, a density-based hierarchical clustering algorithm is used to cluster news texts. Experimental results show that the method proposed in this article can indeed effectively improve the accuracy of Mongolian-Chinese cross-language topic detection.

REFERENCES

[1] H. Yu, Z. Yu, L. Ting, and L. Sheng, "Topic detection and tracking review," *Journal of Chinese information processing*, vol. 21, no. 6, pp. 71–87, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[3] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," *Advances in neural information processing systems*, vol. 17, 2004.

[4] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445–1456.

[5] H. Sayyadi and L. Raschid, "A graph analytical approach for topic detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 13, no. 2, pp. 1–23, 2013.

[6] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[7] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, M.-A. Balafar, C. Motamed *et al.*, "Topicbert: A cognitive approach for topic detection from multimodal post stream using bert and memory–graph," *Chaos, Solitons & Fractals*, vol. 151, p. 111274, 2021.

[8] M. Asgari-Chenaghlu, N. Nikzad-Khasmakhi, and S. Minaee, "Covid-transformer: Detecting covid-19 trending topics on twitter using universal sentence encoder," *arXiv preprint arXiv:2009.03947*, 2020.

[9] T. Leek, H. Jin, S. Sista, and R. Schwartz, "The bbn crosslingual topic detection and tracking system," in *Working Notes of the Third Topic Detection and Tracking Workshop*. Citeseer, 2000.

[10] W. Yang, J. Boyd-Graber, and P. Resnik, "A multilingual topic model for learning weighted topic links across corpora with low comparability," 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)*, 2019, pp. 1243–1248.

[11] B. Mathieu, R. Besançon, and C. Fluhr, "Multilingual document clusters discovery." in *RIAO*. Citeseer, 2004, pp. 116–125.

[12] B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, and I. Temnikova, "Multilingual and cross-lingual news topic tracking," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 959–965.

[13] C.-H. Chang, S.-Y. Hwang, and T.-H. Xui, "Incorporating word embedding into cross-lingual topic modeling," in *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2018, pp. 17–24.

[14] D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 880–889.

[15] S. Hao and M. Paul, "Learning multilingual topics from incomparable corpora," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 2595–2609.

[16] X. Hong, Z. Yu, M. Tang, and Y. Xian, "Cross-lingual event-centered news clustering based on elements semantic correlations of different news," *Multimedia Tools and Applications*, vol. 76, pp. 25 129–25 143, 2017.

[17] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," *arXiv preprint arXiv:2004.07737*, 2020.

[18] X. Chen, L. Luo, H. Wang, W. Wang, and Y. Gao, "Analysis and research on cross language topic discovery in chinese and english," *Advanced Engineering Sciences*, vol. 49, no. 2, pp. 100–106, 2017.

[19] H. Lan and J. Huang, "Chinese-english cross-lingual text clustering algorithm based on latent semantic analysis," *Proceedings of Science*, pp. 1–7, 2017.

[20] W. YANG, Z. YU, S. GAO, and R. SONG, "Chinese-vietnamese news topic discovery method based on cross-language neural topic model," *Journal of Computer Applications*, vol. 41, no. 10, p. 2879, 2021.

[21] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, vol. 30, pp. 31–40, 2009.

[22] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.

[23] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.