Enhancing Cross-Domain Term Extraction with Neural Topic-based Models

Yuan Meng
Southeast University
NanJing, China
yuan_meng@seu.edu.cn

Xiaoqiu Lu*
Southeast University
NanJing, China
luxq823@163.com

Deqiang Yin
Southeast University
NanJing, China
gmrqiang@gmail.com

Guilin Qi†
Southeast University
NanJing, China
gqi@seu.edu.cn

Wei Song
Research Center for Intelligent Robotics, Zhejiang Lab
HangZhou, China
weisong@zhejianglab.com

ABSTRACT

Automated terminology extraction (ATE) aims to identify domain-specific specialized terms in textual data, providing crucial support for constructing knowledge graphs and facilitating information retrieval[1]. With the proliferation of text in emerging domains and the maturity of domain-specific term extraction frameworks, there is a growing interest in improving cross-domain model performance through transfer learning. However, due to the lack of high-quality labelled data in emerging domains and disparities in data distribution and domain-specific information, traditional ATE faces challenges in achieving generalization and domain adaptation when transitioning from source domains to target domains. To address this challenge, we introduce a Cross-Domain Neural Topic Model (dubbed CNTM) that employs a domain training-adaptation-inference paradigm. Specifically, this paradigm enables CNTM to adapt to specific features in the target domain while leveraging source domain knowledge to enhance performance. The approach involves training the model on data-rich source domain data, fine-tuning it with limited target data, and extracting domain-specific terminology from low-resource target text. Additionally, we introduce the Neural Topic Model (NTM) to enhance the model's capacity to identify commonalities across domains, thereby improving cross-domain generalization capabilities. Furthermore, CNTM leverages the robust representation capabilities of pre-trained language models and the discriminative optimization of the contrastive learning module to achieve precise term extraction, especially in a few resource scenarios. Extensive experiments on four real-world datasets across various domains with different sizes and term description sources validate the efficacy and resilience of CNTM in terms of overall performance.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

Term Extraction, Neural Topic Model, Cross-Domain

ACM Reference Format:
Yuan Meng, Xiaoqiu Lu, Deqiang Yin, Guilin Qi, and Wei Song. 2023. Enhancing Cross-Domain Term Extraction with Neural Topic-based Models. In *Proceedings of the International Joint Conference on Knowledge Graphs (IJCKG'24).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Term Extraction (TE) plays a crucial role in deepening our understanding of specialized domain knowledge, providing invaluable input for various Natural Language Processing (NLP) applications[2–5], such as information retrieval, thesaurus construction, question answering, and machine translation. Faced with the frequent emergence of new terms in rapidly evolving domains, traditional term extraction models typically rely on domain-specific corpora and rules, making it challenging for them to generalize to new domains or cross-domain applications. Meanwhile, emerging domains often lack sufficient quantity and quality of annotated data, making it more difficult to build high-performance TE models.

To address this challenge, a large number of previous studies have focused on the task of transfer learning[6–8], where the knowledge and features learned by the model in one domain are transferred and adapted to another domain. This shows that even in the case of low resource scenarios, existing domain knowledge can be effectively leveraged to enhance term extraction performance in new domains. For example, Zhang[7] explores the potential of fine-tuning pre-trained BERT models for Automatic Term Extraction across diverse domains and languages. Hanh[9] conducts a comprehensive study on Transformers-based pre-trained language models for multi-language, cross-domain automatic term extraction, demonstrating the superiority of monolingual models and the benefits of

*Co-first author.
†Corresponding author.

IJCKG'24, December 8 to 9, 2023, Miraikan, Tokyo, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ensemble strategies. These approaches leverage the generalization capabilities of PLM across multiple languages and domains, thereby enhancing term extraction performance in emerging domains.

Traditional PLM are typically trained on large-scale generic corpora, excelling at capturing general semantic information. However, this can lead to a decrease in performance when dealing with domain-specific terms, especially in emerging domains with limited annotated data, as they lack domain-specific knowledge and struggle to identify unique information across different domains. Terms from different domains may possess distinct meanings and contexts, making it challenging for pre-trained models to accurately capture and classify these terms. For example, "heart attack" has a clinical meaning related to cardiovascular health in medicine, while in the automotive field, it refers to an engine suddenly stopping. Traditional pre-trained language models may struggle to distinguish between these meanings due to their lack of domain-specific knowledge and context, leading to potential misinterpretations.

To address these challenges, we propose a comprehensive approach, CNTM for terminology extraction, which comprises three main stages: domain training, domain adaptation, and domain inference. During the domain training phase, we leverage a richly labeled source dataset to provide a pre-trained architecture for text in the target domain. This architecture includes an input module, an encoding module, a topic-enhanced encoding module, and a decoding module. These modules collaborate to capture the intricate relationships between terms and their context, ensuring a comprehensive representation of terminology within the domain. The whole process provides the basis for the subsequent stages of domain adaptation and domain reasoning.

In the domain adaptation stage, we employ limited annotated data from the target domain to fine-tune CNTM, thereby enhancing its adaptation to the distinctive characteristics of the target domain. This approach helps mitigate the challenge of transferring information between different domains, as the model can undergo parameter adjustments to align more effectively with the data distribution of the target domain. The adaptation also involves the use of an early stopping strategy to control gradient propagation and prevent overfitting. Additionally, a topic enhancement process is employed to capture semantic information from the target domain's documents. The domain inference phase utilizes the neural network model trained in the domain training and adaptation phases to perform terminology inference on unlabeled data in the target domain. This involves calculating the distance between domain-specific descriptors and individual words in input documents to assess domain relevance, ultimately achieving cross-domain term extraction tasks.

In summary, CNTM provides a comprehensive solution for cross-domain terminology extraction, offering valuable support in addressing the challenge of weak information transferability. By leveraging pre-trained models, contrastive learning, and a neural topic module, these integrated techniques significantly enhance the accuracy of term extraction, particularly in scenarios characterized by emerging domains and limited annotated data. A series of comprehensive experiments conducted on four real-world datasets spanning various domains provides strong evidence for the robustness and effectiveness of CNTM.

2 RELATED WORK

2.1 Neural topic model

Recently, the Neural Topic Model has emerged as a prominent research area, attracting significant attention in the fields of text mining and natural language processing. While traditional topic models like Latent Dirichlet Allocation (LDA) [10] have made commendable progress in text topic modeling, they often face challenges in capturing intricate semantics and implicit information embedded within textual content. The Neural Topic Model (NTM) harnesses the integration of neural networks into topic modeling, enabling more effective learning of semantic representations and underlying topics from extensive textual datasets. NTM's potential to unveil intricate topic structures and decode latent semantic structures has attracted researchers exploring its applicability across diverse domains [7, 11]. By combining neural networks with topic modeling, NTM provides a pathway to extract deeper insights from textual data, contributing to advancements in various natural language understanding tasks. Zhao [12] proposed a method for learning document topic distribution representations by directly minimizing the optimal distance between document-to-document word distributions. Wang [13] introduced the Layer-Assisted Neural Topic Model (LANTM), which enhances the encoding capability of topic representations by connecting text content with auxiliary networks. Yang [14] presented TopNet, a method that utilizes recent advances in neural topic modeling to generate high-quality backbone words, addressing the limitations of short inputs. Gupta [15] proposed a neural topic modeling framework that utilizes a multi-view embedding space, incorporating pre-trained topic embeddings and word embeddings, which not only better handles polysemy but also enhances the quality of obtained topics. While continuous advancements in neural topic modeling have been made, existing work has not fully harnessed the rich information conveyed by input documents. We introduce a topic-enhancement module to amplify the abundant thematic semantic information corresponding to the original input documents. Simultaneously, we leverage a pre-trained language model based on contrastive learning and prompt learning to acquire comprehensive label information for term labels, thereby effectively enhancing the domain adaptation capability of term extraction tasks.

2.2 Automatic term extraction

In recent years, the field of term extraction algorithms has undergone a significant transformation, transitioning from manual extraction to automated methods. Manual extraction heavily relies on domain experts and their specialized knowledge, which can be labor-intensive and inefficient. In the realm of automated term extraction, the initial focus was primarily on the intrinsic characteristics of terms and shallow linguistic analysis based on their frequency in target corpora.Methods in this early stage of development can be broadly categorized into three main approaches: linguistics-based, statistics-based, and hybrid approaches that combine linguistic and statistical elements.Linguistics-based approaches in term extraction are primarily centered around leveraging manually constructed linguistic rules to facilitate the automatic extraction of terms. For instance, researchers like Bourigault[16] employ part-of-speech tagging to annotate documents within a given corpus. Based on a

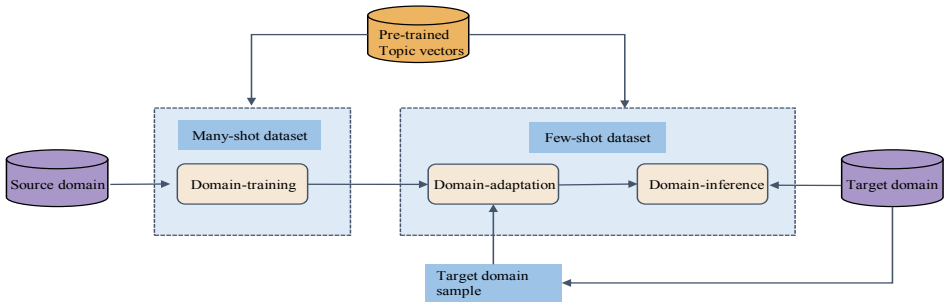

Figure 1: The overall framework of CNTM, which consists of three main components: domain training, domain adaptation, and domain inference. Domain training uses labeled data from the source domain, while domain adaptation and inference utilize limited labeled data from the target domain for fine-tuning and term inference.

pre-existing set of standard terms, they utilize finite-state annotate (FSM) to automatically deduce a collection of rules implicit within the documents. Frantzi[17]were among the pioneers to amalgamate linguistic and statistical methods in term extraction. They introduced the C-value method, which initially selects candidate terms based on certain rules and subsequently employs statistical criteria to filter these candidate terms.

With the continuous advancement of machine learning and deep learning, an increasing number of cutting-edge technologies have been applied to term extraction tasks. Automatic terminology extraction is mostly based on methods including statistical analysis, semantic relationships, machine learning, and rule formulation. For expamele, vivaldi[18] introduced an external knowledge base, Wikipedia, into the terminology extraction task to enrich semantic information. Yang[19] introduced an adaptive unsupervised clustering algorithm that enhances the algorithm's robustness and stability by utilizing noisy data to construct pseudo-terms for training and incorporating fault tolerance mechanisms. CoAKT[20] uses deep learning to reduce the dependence of traditional artificial feature engineering methods. AutoPhrase[21] employs distant supervision techniques and phrase segmentation guided by part-of-speech tagging, effectively avoiding the need for additional manual labeling efforts and enhancing the effectiveness of terminology extraction. Hazem[22] proposed a BERT-based approach for terminology extraction, achieving favorable experimental results on multilingual datasets. Tran[9] and Hazem[23] employed pre-trained models to automatically extract terms across domains from low-resource data. They achieved cross-lingual and cross-domain transfer learning by pre-training on the source language and fine-tuning on the target language.

While pre-trained language models trained on extensive corpora have demonstrated remarkable accuracy and recall in term extraction, they encounter challenges in cross-domain term extraction tasks, particularly in nascent domains with limited labeled data. Prominent challenges involve the limited capacity to effectively transfer information across domains and the intricacy of identifying distinctive elements amid diverse domains. To address these issues, the subsequent term extraction framework introduces the topic information derived from the neural topic model into the pre-trained language model, thereby increasing the information extraction ability of the model.

3 PROBLEM DEFINITION

We adhere to the established named entity recognition BIO annotation format (Lample et al., 2016), where 'B' indicates the beginning of a term, "I" represents an intermediate term, and "O" signifies non-term entities. Each token x_i in a specific input text $X = \{x_1, ..., x_t\}$ is assigned a label $y_i \in C$, where C denotes a predefined set of domain-specific term labels. Our datasets consist of a source domain training set with abundant labeled data (H) and a low-resource target domain dataset (T), which is a union of sparsely labeled samples (S) and an unlabeled target domain test set (L).

4 METHOD

The overall framework of the neural topic model-based terminology extraction method is depicted in Figure 1. Firstly, the domain training phase involves training the neural network on a substantial amount of labeled domain-specific data, thereby enabling the model to learn a representation that adapts to the distinctive features of the domain's data. Secondly, the purpose of domain adaptation is to fine-tune model parameters using a limited amount of samples, aiming to enhance the model's adaptation to specific domain data characteristics. Finally, domain inference is performed on domain-specific datasets with limited annotations, applying the learned model to extract term information from unlabeled data. Throughout these three phases, we leverage pre-trained topic vectors to enhance semantic information within the datasets, improving cross-domain term recognition. Then we will provide a detailed elaboration of these three stages.

4.1 Domain training phase

During the domain training phase, the model is trained on a source dataset with rich labels, thus providing a pre-trained architecture for texts in the target domain. The specific model architecture, as illustrated in Figure 2, mainly comprises input Module Based on Prompt Learning, encoding Module based on contrastive learning, a topic-enhanced encoding module, and a decoding module based on BiLSTM and CRF.

4.1.1 Knowlegable context construction. The prompt-based input module aims to fully exploit the rich semantic information inherent in texts to represent domain-specific term labels. For instance, given a specfic input sentence "*[BOS] A Taxonomy of Cyber Attacks on*

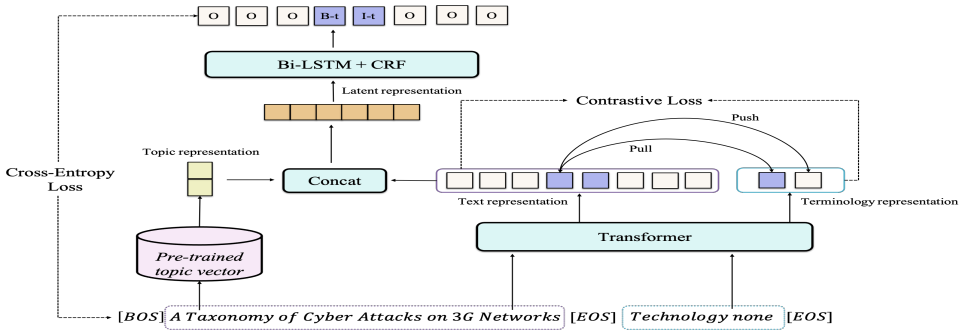

Figure 2: Model Structure in Domain training phase

3G Networks [EOS]". We can utilize predefined prompt template "*Technology none*", "*Technology none*" indicates a technology domain term, and "*none*" signifies non-terms. Alternatively, "none" can also be interpreted as an O-class word in the original BIO annotation, representing non-term entities. This approach associates input text with domain information and helps label non-term words using prompt templates. To optimally leverage the prompt information, the construction of task-specific prompts tailored to specific domain datasets is necessary.

Hence, we denote the set of prompts as M, which associates term labels with domain descriptions. Each word v_i used for describing domain information can be obtained through the mapping M and the corresponding domain label $y_i \in C$, where C represents the predefined set of term domain labels. This study employs a simple yet effective approach to customize such a mapping set M. For instance, for a specific domain term label like "Tech," intuitively using "Technology" as the domain description prompt is chosen, i.e., $M(\text{Tech}) = \text{T echnology}$. The design of prompts is usually customized for a specific term or domain. This type of prompts inherently contains general semantic information and context related to the corresponding term domain and helps mitigate the bias of limited labeled data.

For the term extraction task, the term domain labels are initially mapped to domain description information. Subsequently, the generated prompt information is concatenated with the input document X, resulting in an extended input sequence $X' = \{x_1, x_2, ..., x_t, v_1, v_2\}$. Here, t represents the length of the original input document, v_1 signifies the domain-specific label, and v_2 represents "none", denoting non-domain terms. The input text X' is fed through a pre-trained language model $g(.)$ to obtain corresponding latent vector representations. The final hidden layer outputs serve as the representations of each word, as shown in Equation 1:

$$H = [h_1, \ldots, h_t, h_{,1}, h_{,2}] = g([x_1, \ldots, x_t, v_1, v_2]) \tag{1}$$

4.1.2 Domain Training through Contrastive Learning. In order to enhance the model's cross-domain learning capabilities and improve its adaptation to diverse semantic information across various domains, we introduce the optimization concept of contrastive learning. The goal of this module is to minimize the distance between the input text X and its corresponding domain-specific term description, while simultaneously maximizing the distance between irrelevant domain-specific term descriptions. This strategy is designed to effectively capture the relationship between words and domain labels, ensuring that the model can precisely differentiate terms across various domains.

Specifically, each positive pair is defined as (x_p, v_p), where v_p denotes the corresponding domain term description. On the other hand, each negative pair is formed by combining x_p with unrelated domain term description from the prompt information. Thus, the contrastive loss for word x_p is given by Equation 2.

$$\ell\left(x_p\right) = -\log \frac{\exp\left(-d\left(\mathbf{h}_p, \mathbf{h}'_p\right)/\tau\right)}{\sum_{q=1}^{2} \exp\left(-d\left(\mathbf{h}_p, \mathbf{h}'_q\right)/\tau\right)} \tag{2}$$

Where τ represents the temperature hyperparameter, we employ the Euclidean distance to measure similarity of vectors, as shown in Equation 3.

$$d\left(\mathbf{h}_p, \mathbf{h}'_q\right) = \left\|\mathbf{h}_p - \mathbf{h}'_q\right\|_2^2 \tag{3}$$

The loss function based on contrastive learning can be defined as in Equation 4.

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} \ell\left(x_i\right) \tag{4}$$

Where $\mathcal{X}$represents the collection of documents.

4.1.3 Topic Enhancement Module. The key distinction between TE tasks and NER lies in the fact that the latter can identify multiple categories of fine-grained instance, providing richer and more diverse label semantics. Conversely, the descriptions in TE task is generally consistent and unique. Compared to the finer-grained label information in named entity recognition, term extraction tasks typically involve coarser-grained domain labels. To harness the semantic richness in domain-specific dataset documents and mitigate label scarcity in term extraction tasks, we introduce a topic-enhanced encoding module to enhance semantic space encoding.

Specifically, we employ a neural topic model based on mutual information[24] to obtain high-quality latent representations for each input document. These topic representations effectively capture the semantic information of the original input documents and can be utilized for text reconstruction. For a given input document

$X = \{x_1, ..., x_t\}$, its latent vector representation after being processed by a pre-trained language model is denoted as $E = \{e_1, ..., e_t\}$, and the corresponding topic vector representation is denoted as $Z = \{z_1, ..z_i\}$, where i represents the predefined number of topics.

$$\mathcal{P} = Z \oplus E \tag{5}$$

Concatenating the elements from E and Z yields P, which provides the decoding module with highly representative and reconstructive feature vectors.

4.1.4 Decoding Module and Training Objectives. The decoding module primarily decodes the comprehensive encoding vector P to obtain the predicted labels. Subsequently, the predicted label is derived by calculating the cross-entropy loss between the input document and the output document. Specifically, we employ Bi-LSTM and CRF as the decoding module, with P serving as input to the decoder to produce an output vector Q. The dimension of Q is $t \times m$, where t signifies the number of words in the input document, and m represents the count of labels. In this context, Q_{ij} represents the probability associated with the jth label for the ith word in the input document. For a predicted sequence $y = \{y_1, y_2, ..., y_t\}$, its probability is formulated as depicted in Equation 6:

$$f(X, y) = \sum_{i=0}^{t} A_{y_i y_{i+1}} + \sum_{i=1}^{t} Q_{i, y_i} \tag{6}$$

The matrix A represents the transition matrix, with A_{ij} denoting the probability of transitioning from label i to label j. Here, y_0 and y_t correspond to the starting and ending tags for the predicted sequence, respectively. The probability of generating the label sequence y based on the original input sentence X is mathematically represented by Equation 7:

$$p(y \mid X) = \frac{e^{f(X,y)}}{\sum_{\tilde{y} \in Y_X} f(X, \tilde{y})} \tag{7}$$

Where $\tilde{y}$ represents the true label. We employ the cross-entropy loss function to maximize the likelihood probability of the accurate label sequence, as illustrated by Equation 8:

$$\mathcal{L}_{\text{entropy}} = -\log(p(y \mid X)) = f(X, y) - \log\left(\sum_{\tilde{y} \in Y_X} e^{f(X, \tilde{y})}\right) \tag{8}$$

Where Y_X represents all possible label sequences corresponding to an input sequence X. Finally, the objective function can be viewed as a fusion of the contrastive learning loss function and the cross-entropy loss function, as depicted in Equation 9:

$$\mathcal{L} = \omega * \mathcal{L}_{\text{con}} + (1 - \omega) * \mathcal{L}_{\text{entropy}} \tag{9}$$

It can be observed that the domain traing phase encompasses an input module based on prompt learning, Domain Training through contrastive learning, a topic enhancement module, and decoder modules.

4.2 Domain Adaptation

The domain adaptation phase aims to fine-tune the model trained on the source domain using a small amount of labeled data from the target domain, enabling cross-domain term inference tasks. The model structure of the domain adaptation phase is similar to that of the domain adaptation phase, with most modules resembling those from the domain training phase. Furthermore, to enhance the model's domain adaptation capability for different label spaces, we have improved the topic-enhanced module during the domain adaptation phase. This enhancement aims to fully leverage the rich semantic information present in the unlabeled target domain dataset.

In the term extraction task, the primary distinction between the target domain and the source domain lies in the variation of term labels. Given that each dataset corresponds to a specific document within the domain, the domain-specific descriptions and domain labels used in hint learning vary. To enhance the model's domain adaptation capability, we fine-tune a pre-trained neural network model that has been trained in the source domain using a small amount of annotated data from the target domain. In this process, we use an early stopping strategy[25] to control the degree of gradient propagation, which works by monitoring the model's performance on the validation data set during training and stopping the training process when performance starts to decline. This indicates that the model has started to overfit the training data, thereby preventing overfitting and improving the model's generalization.

4.2.1 Topic Enhancement based on ElasticSearch. When dealing with text data associated with a specific domain, the textual information related to the target domain usually contains diverse semantic details that can effectively capture the essence of domain-specific terms. Additionally, these meticulously structured and constrained vocabularies often demonstrate a notable extent of word overlap. Motivated by the insights from reference[26], we employ an approach in which the provided text functions as a query, aiming to retrieve the most analogous samples from the target domain's database. In this study, we leveraged the existing Elasticsearch[27]retrieval engine and employed the efficient BM25[28] retrieval method. We constructed an Elasticsearch index using the built-in standard analyzer, specifically tailored for large-scale unlabeled domain corpora. This index enables real-time retrieval of the top K samples based on the BM25 scores computed from the input text. Subsequently, the top K samples with the closest similarity are extracted for further topic enhancement. Specifically, we perform a weighted average of the pre-trained topic vectors corresponding to these documents,resulting in an average topic vector that corresponds to the original input document. This process serves to enhance the semantic information encapsulated within the original input document.

4.3 Domain Inference Phase

The purpose of domain inference is to utilize the neural network model trained during the domain training and domain adaptation phases. This aims to perform terminology inference on a large volume of unlabeled data in the target domain, determining whether the input document contains specific terms from the domain. The model structure is illustrated in Figure 3.

We calculate the distance between domain-specific descriptors and each individual word to assess the domain relevance of words. For the input documents in the test set L_{test}, they are fed into the pre-trained language model to obtain their corresponding latent vector representations, as shown in Equation 10:

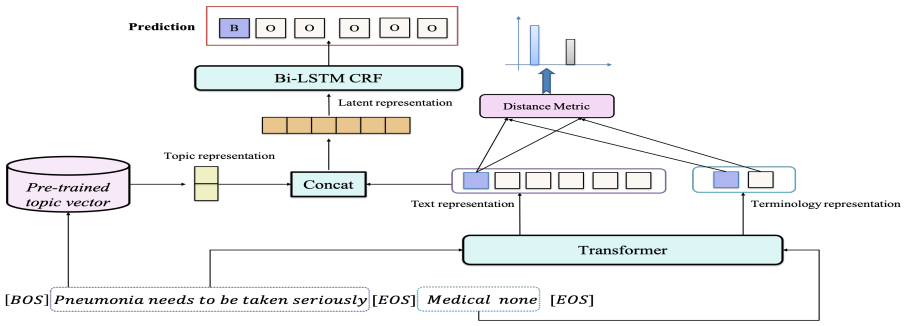

Figure 3: Model Structure in the Domain Inference Phase

$$L_{\text{test}} = \left[\mathbf{h}_1, \ldots, \mathbf{h}_t, \mathbf{h}'_1, \mathbf{h}'_2\right] \tag{10}$$

For each word x_i in the input document, we find the closest anchor embedding and assign its corresponding domain label information c_j to the respective word token, as shown in Equation 11.

$$y_i^{\text{test}} = \underset{c_j}{\operatorname{argmin}} \left\|\mathbf{h}_i - \mathbf{h}'_j\right\|_2^2 \tag{11}$$

Where h_i, h'_j represents the latent representation of word,domain label, respecially. In the process of term inference, the Viterbi algorithm is still used to predict the output sequence with the largest score, so as to obtain the BIO prediction sequence representation corresponding to the input document, as shown in Equation 12:

$$y^* = \arg_{\tilde{y} \in Y_X} \max f(X, \tilde{y}) \tag{12}$$

The final term inference results can be obtained through the integration of distance-based contrastive learning predictions and probability predictions based on Viterbi decoding, thereby achieving cross-domain term extraction tasks.

5 EXPERIMENT

5.1 Experiment Set

Dataset. To provide a comprehensive evaluation of the performance of the TNE for cross-domain term extraction tasks, we conducts experiments on four real-world datasets. Table 1 provides statistics for each dataset, including the number of documents, average document length, and the count of standard terms.

ACLVer2: This dataset comprises abstracts from 300 articles in the field of computer science, sourced from the Association for Computational Linguistics (ACL) index publications.

GENIA: The GENIA dataset is designed for semantic annotation in biomedical text mining, consisting of 2,000 abstracts covering various domains in biomedicine, such as human biology, blood cells, and transcription factors.

TTCM: Comprising 37 articles focused on mobile technology, this dataset was constructed by web scraping and includes a manually filtered "standard terminology list," making it a valuable resource for research and analysis in the field.

TTCW: This collection of 103 articles, centered around wind energy, offers invaluable data concerning wind turbine performance and environmental conditions.

Table 1: Terminology Extraction Statistics for the Four Datasets

Dataset	#Terms	Average Document Length	#Documents
ACL	3059	736	300
GENIA	33396	1498	2000
TTCM	255	55652	37
TTCW	288	49737	103

5.2 Baseline Models

To evaluate the practical effectiveness of our framework, we compare our model with the most prominent methods:

- **ComboBasic**[29] is a terminology extraction algorithm, places its primary focus on evaluating the frequency of candidate terms in documents and analyzing their contextual relationships. Additionally, it utilizes domain-specific vocabularies and incorporates multi-positional features to effectively filter out irrelevant terms.
- **TermExtractor**[30] is a graph-based terminology extraction algorithm that builds a co-occurrence matrix and a corresponding weighted graph through the fusion of textual data. The edges within the weighted graph signify co-occurrence associations between candidate terms, enabling the extraction of domain-specific terms through graph analysis.
- **CoAKT**[20] is an automatic key term extraction method based on deep learning that operates without explicit features. Its goal is to decrease reliance on traditional feature engineering methods through deep learning techniques, leveraging a large-scale corpus for training. By employing unsupervised learning, it enhances the representational capacity of text features, ultimately reducing the time and effort needed for manual feature construction.
- **TTE**[31] is a multi-language terminology extraction model based on Transformers. It employs multiple Transformer layers to extract latent text features and utilizes self-attention mechanisms to capture relationships among words. Furthermore, the model incorporates an innovative filtering mechanism for efficient candidate term selection.
- **BERT-biLSTM-CRF**[23] is a cross-domain terminology extraction model. Based on BERT, it captures shared contextual

Table 2: Performance of Domain Terminology Extraction

Method	ACL		GENIA		TTCM		TTCW	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
ComboBasic	0.15	0.11	0.52	0.09	0.27	0.34	0.36	0.64
TermExtractor	0.23	0.15	0.41	0.10	0.23	0.17	0.20	0.32
CoAKT	0.38	0.19	0.57	0.15	0.43	0.38	0.47	0.58
TTE	0.42	0.40	0.60	0.45	0.53	0.51	0.61	0.52
BERT-biLSTM-CRF	0.51	0.43	0.63	0.41	0.52	0.42	0.63	0.54
TNE	0.53	0.50	0.65	0.52	0.58	0.45	0.67	0.56

information for terms across domains and languages. Leveraging BERT, this model conducts cross-language and cross-domain transfer learning, resulting in enhanced extraction of both single-word and multi-word terms. This approach significantly improves the efficiency of term extraction.

5.3 Evaluation Metrics

Term extraction models typically extract candidate terms from domain-specific corpora as output. This study employs the complete term list obtained from the corpus as the ground truth. Evaluation of model-generated candidate terms is performed using Precision, Recall, and F1-score.

- **Precision(Pre)** is calculated by determining the proportion of correctly extracted terms among all candidate terms predicted by the model, as Equation 13.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{13}$$

- **Recall(Rec)** is calculated by determining the proportion of correctly extracted terms among all standard terms within the domain corpus, as Equation 14.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{14}$$

- **F1-score(F1)** is obtained by computing the harmonic mean of precision and recall, providing a comprehensive assessment of the term extraction model's performance, as Equation 15.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}} \tag{15}$$

5.4 Implementation Details

Hyperparameters for ComboBasic, TermExtractor, CoAKT, TTE, and BERT-biLSTM-CRF were configured based on the optimal settings outlined in their respective works. The ratios of domain training, domain adaptation, and domain inference in the same domain dataset are set to 8:1:1, respectively. During model training, a learning rate of 10^{-4} was employed in conjunction with the Adam optimizer across all datasets. The ω parameter in the overarching loss function described in Equation 4.8 was empirically set to $\omega = 0.4$. All experiments consisted of 100 training epochs and employed a batch size of 128. In the content of cross-domain term extraction, this paper employed 1-shot and 5-shot settings during the domain adaptation phase to assess the performance of neural topic model-based term extraction methods and other baseline models in cross-domain term extraction tasks.

6 EXPERIMENT ANALYSIS

6.1 Experimental results in a domain

The overall performance of compared methods and the proposed framework is indicated in Table 2.

Firstly, when we evaluate the traditional term extraction frameworks, including ComboBasic, TermExtractor, and CoAKT, we observe that their precision (Pre) and recall (Rec) values are relatively modest across all datasets. Simultaneously, it is evident that the two pre-trained language models, TTE and BERT-biLSTM-CRF, outperform the conventional term extraction frameworks, ComboBasic, TermExtractor, and CoAKt, underscoring the substantial capabilities of pre-trained language models in terms of domain-specific semantic information and their ability to harness domain knowledge effectively. We observe that the TNE yields optimal results in terms of precision and recall on the ACL, GENIA, TTCM, and TTCW datasets. This suggests that the fusion of the neural topic model and pre-trained language models facilitates the acquisition of domain-specific thematic information, effectively enhancing the representation capabilities of the topic-enhancement module within the term extraction model.

6.2 Cross-domain results on Low Resource Data

To validate the domain training, domain adaptation, and domain inference paradigms mentioned in CNTM, we focus on assessing the performance of the model specifically in cross-domain terminology extraction tasks. Specifically, in this study, we employ the ACL dataset as the source domain dataset. We then evaluate the experimental results when using GENIA, TTCM, and TTCW as target domain datasets. During the domain adaptation phase, we conduct experiments with small-sample adaptations set at 1-shot and 5-shot scenarios. The cross-domain term extraction F1 scores are presented in Table 3

We can find our model consistently achieves the highest F1 scores. This indicates that our model exhibits superior domain transferability compared to other baseline models, ultimately enhancing the model's scalability. In contrast to BERT-biLSTM-CRF, TNE performs better in the few data scenario. This improvement can be attributed to domain adaptation phase, which incorporates an ElasticSearch-based topic enhancement module. This module effectively leverages unsupervised topic information to identify domain-specific thematic information, leading to enhanced term extraction performance.

Table 3: Performance of cross-domain terminology extraction

Method	1 shot			5 shot		
	GENIA	TTCW	TTCM	GENIA	TTCW	TTCM
ComboBasic	0.08	0.04	0.03	0.09	0.04	0.02
TermExtractor	0.10	0.09	0.08	0.12	0.10	0.11
CoAKT	0.15	0.12	0.16	0.17	0.13	0.16
TTE	0.31	0.25	0.24	0.33	0.25	0.26
BERT-biLSTM-CRF	0.33	0.27	0.26	0.35	0.28	0.29
Ours	0.43	0.41	0.39	0.45	0.42	0.43

6.3 Ablation experiment

In this section, we conducted ablation experiments to analyze the impact of major components of TNE framework: 1) **CNTM-T**:we have removed the Topic Enhancement Module from the TNE framework.; (2)we have excluded the Contrastive Learning from TNE to evaluate whether it imposes limitations on the extraction of domain-specific terms;

Table 4: Performance with Different Adaptation Shots

Method	GENIA	TTCW	TTCM
CNTM-T	0.25	0.21	0.19
CNTM-C	0.31	0.28	0.29
CNTM	0.43	0.41	0.39

As Table 4 shows, when we remove the Topic Enhancement Module in "CNTM-T," or Contrastive learning Module in "TNE-C", there are noticeable drop in performance. Specially, TNE integrating both modules, attained the highest performance with F1-scores of 0.43, 0.41, and 0.39. These results emphasize the importance of both modules and their synergy for effective terminology extraction.

6.4 Template Prompt Comparison Experiment

To explore the influence of different templates on the cross-domain term extraction task, we attempted to create distinct template prompts for various datasets. For the four datasets, AC, GENIA, TTCM, and TTCW, the corresponding template are shown in Table 5 as follows: Initial templates included a single domain keyword, while expanded templates enriched domain information from the initial templates. The resulting term extraction F1 scores are illustrated in Figure 4

Figure 4 clearly shows that the extended template outperforms the initial template in cross-domain experiments. This improvement

Table 5: Term Extraction Statistics for the Four Datasets

Dataset	Initial template	Extended template
ACL	Linguistics	Computational Linguistics
GENIA	Biomedical	Biomedical Science
TTCM	Technology	Mobile Technology
TTCW	Energy	Wind Energy

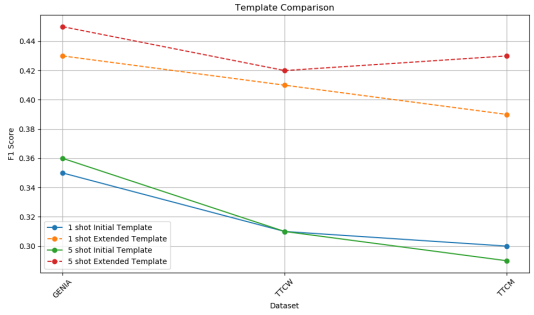

Figure 4: Experimental performance based on different templates

can be attributed to the enhanced semantic content provided by the extended template for term labels.

7 CONCLUSION

In this study, we proposed a cross-domain neural topic model(dubbed CNTM) to address low-resource term extraction challenges. CNTM showed substantial advancements in recognizing domain-specific terms, adapting to diverse domains, and has the potential to enhance applications like knowledge graph construction and information retrieval. The incorporation of topic-based enhancements and transfer learning techniques enhances the capability to capture semantic nuances, making it more adaptable to different domains, especially resource-poor emerging domains. This paves the way for future research to explore the framework's further enhancements and applications.

ACKNOWLEDGEMENT

This work is partially supported by National Nature Science Foundation of China under No. U21A20488. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

REFERENCES

[1] T. T. H. Hanh, M. Martinc, J. Caporusso, A. Doucet, S. Pollak, The recent advances in automatic term extraction: A survey, CoRR abs/2301.06767 (2023). arXiv:2301.06767, doi:10.48550/arXiv.2301.06767.
URL https://doi.org/10.48550/arXiv.2301.06767
[2] M. T. Pazienza, M. Pennacchiotti, F. M. Zanzotto, Terminology extraction: an analysis of linguistic and statistical approaches, in: Knowledge mining: Proceedings of the NEMIS 2004 final conference, Springer, 2005, pp. 255–279.
[3] A. Peñas, F. Verdejo, J. Gonzalo, et al., Corpus-based terminology extraction applied to information access, in: Proceedings of corpus linguistics, Vol. 2001, 2001, p. 458.

[4] H. Du, Z. Le, H. Wang, Y. Chen, J. Yu, Cokg-qa: Multi-hop question answering over covid-19 knowledge graphs, Data Intelligence 4 (3) (2022) 471–492.
[5] H. Chen, N. Hu, G. Qi, H. Wang, Z. Bi, J. Li, F. Yang, Openkg chain: A blockchain infrastructure for open knowledge graphs, Data Intelligence 3 (2) (2021) 205–227.
[6] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.
[7] M. Peng, Q. Zhang, Y. Jiang, X. Huang, Cross-domain sentiment classification with target domain specific information, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2505–2513.
[8] Y. A. Winatmoko, A. A. Septiandri, A. P. Sutiono, Aspect and opinion term extraction for hotel reviews using transfer learning and auxiliary labels, arXiv preprint arXiv:1909.11879 (2019).
[9] H. T. H. Tran, M. Martinc, A. Pelicon, A. Doucet, S. Pollak, Ensembling transformers for cross-domain automatic term extraction, in: International Conference on Asian Digital Libraries, 2022, pp. 90–100.
[10] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.
[11] Y. Miao, L. Yu, P. Blunsom, Neural variational inference for text processing, in: M. Balcan, K. Q. Weinberger (Eds.), Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, Vol. 48 of JMLR Workshop and Conference Proceedings, JMLR.org, 2016, pp. 1727–1736.
[12] H. Zhao, D. Phung, V. Huynh, T. Le, W. Buntine, Neural topic model via optimal transport, arXiv preprint arXiv:2008.13537 (2020).
[13] Y. Wang, X. Li, J. Ouyang, Layer-assisted neural topic modeling over document networks., in: IJCAI, 2021, pp. 3148–3154.
[14] Y. Yang, B. Pan, D. Cai, H. Sun, Topnet: Learning from neural topic model to generate long stories, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1997–2005.
[15] P. Gupta, Y. Chaudhary, H. Schütze, Multi-source neural topic modeling in multi-view embedding spaces, arXiv preprint arXiv:2104.08551 (2021).
[16] K.-h. Chen, H.-H. Chen, Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation, arXiv preprint cmp-lg/9405034 (1994).
[17] K. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms:. the c-value/nc-value method, International journal on digital libraries 3 (2000) 115–130.
[18] J. Vivaldi, L. A. Cabrera-Diego, G. Sierra, M. Pozzi, et al., Using wikipedia to validate the terminology found in a corpus of basic textbooks., in: LREC, Citeseer, 2012, pp. 3820–3827.
[19] Y. Yang, H. Yu, Y. Meng, Y. Lu, Y. Xia, Fault-tolerant learning for term extraction, in: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, 2010, pp. 321–330.
[20] K. Khosla, R. Jones, N. Bowman, Featureless deep learning methods for automated key-term extraction (2019).
[21] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, J. Han, Automated phrase mining from massive text corpora, IEEE Transactions on Knowledge and Data Engineering 30 (10) (2018) 1825–1837.
[22] C. Lang, L. Wachowiak, B. Heinisch, D. Gromann, Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains, in: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, Vol. ACL/IJCNLP 2021 of Findings of ACL, Association for Computational Linguistics, 2021, pp. 3607–3620.
[23] A. Hazem, M. Bouhandi, F. Boudin, B. Daille, Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC , Marseille, France, 20-25 June, European Language Resources Association, 2022, pp. 648–662.
[24] K. Xu, X. Lu, Y. fang Li, T. Wu, G. Qi, N. Ye, D. Wang, Z. Zhou, Neural topic modeling with deep mutual information estimation, Big Data Research 30 (2022) 100344.
[25] Y. Huang, K. He, Y. Wang, X. Zhang, T. Gong, R. Mao, C. Li, Copner: Contrastive learning with prompt guiding for few-shot named entity recognition, in: Proceedings of the 29th International conference on computational linguistics, 2022, pp. 2515–2527.
[26] X. Zhang, Y. Jiang, X. Wang, X. Hu, Y. Sun, P. Xie, M. Zhang, Domain-specific ner via retrieving correlated samples, arXiv preprint arXiv:2208.12995 (2022).
[27] B. Elasticsearch, Elasticsearch: The official distributed search & analytics engine (2020).
[28] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: SIGIR'94, Springer, 1994, pp. 232–241.
[29] N. A. Astrakhantsev, D. G. Fedorenko, D. Y. Turdakov, Methods for automatic term recognition in domain-specific text collections: A survey, Programming and Computer Software 41 (2015) 336–349.
[30] F. Sclano, P. Velardi, Termextractor: a web application to learn the shared terminology of emergent web communities, in: Enterprise Interoperability II: New Challenges and Approaches, Springer, 2007, pp. 287–290.
[31] C. Lang, L. Wachowiak, B. Heinisch, D. Gromann, Transforming term extraction: transformer-based approaches to multilingual term extraction across domains, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 3607–3620.