Cross-domain NER in Data-poor Scenarios for Human Mobility Knowledge

Yutong Jiang
Beijing Institute of Technology
Fusheng Jin
jfs21cn@bit.edu.cn
Beijing Institute of Technology
Mengnan Chen
Beijing Institute of Technology
Guoming Liu
Beijing Aerospace Automatic Control Institute
He Pang
Beijing Aerospace Automatic Control Institute
Ye Yuan
Beijing Institute of Technology

Research Article

Keywords: Named entity recognition, Human mobility, Few-shot, Deep learning

Posted Date: July 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3152699/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at GeoInformatica on March 5th, 2024. See the published version at https://doi.org/10.1007/s10707-024-00513-z.

Cross-domain NER in Data-poor Scenarios for Human Mobility Knowledge

Yutong Jiang[1], Fusheng Jin[1*], Mengnan Chen[1], Guoming Liu[2], He Pang[2], Ye Yuan[1]

[1*]School of Computer Science and Technology, Beijing Institute of Technology, Southern Street of Zhongguanchun, Beijing, 100081, China.
[2] Beijing Aerospace Automatic Control Institute, Yongding Rd, Beijing, 100854, China.

*Corresponding author(s). E-mail(s): jfs21cn@bit.edu.cn;

Abstract

In recent years, the exploration of knowledge in large-scale human mobility has gained significant attention. In order to achieve a semantic understanding of human behavior and uncover patterns in large-scale human mobility, Named Entity Recognition (NER) is a crucial technology. The rapid advancements in IoT and CPS technologies have led to the collection of massive human mobility data from various sources. Therefore, there's a need of Cross-domain NER which can transfer entity information from the source domain to automatically identify and classify entities in different target domain texts. In the situation of data-poor, how could we transfer human mobility knowledge over time and space is particularly significant, therefore this paper proposes an Adaptive Text Sequence Enhancement Module (at-SAM) to help the model enhance the association between entities in sentences in data-poor target domains. This paper also proposes a Predicted Label-Guided Dual Sequence Aware Information Module (Dual-SAM) to improve the transferability of label information. Experiments were conducted in domains that contain hidden knowledge about human mobility, the results show that this method can transfer task knowledge between multiple different domains in data-poor scenarios and achieve the best performance to date.

Keywords: Named entity recognition, Human mobility, Few-shot, Deep learning

1 Introduction

Named Entity Recognition (NER) is a basic and important task in semantically understanding human behavior, which aims to identify specified categories of information from text, such as basic categories like people, location, etc.[1]. Named entity recognition can be viewed as a specific sequence labeling problem, Transformer is a model proposed by Vaswani et al.[2] that does not require recurrence and convolutions. Radford et al.[3] proposed Generative Pre-trained Transformer (GPT) for language understanding tasks based on Transformer. Baevski et al.[4] proposed a novel cloze-driven pre-training regime based on a bi-directional Transformer, which is trained with a cloze-style objective and predicts the center word given all left and right context. Although methods based on pre-trained language models[5],[6],[7],[8] have made significant progress on the named entity recognition task, these models still rely heavily on large amounts of training data to ensure good performance and adapt well to target domains with scarce training data[9].

However, due to potential privacy concerns, obtaining a large amount of data may be challenging, so we often encounter situations where data availability is limited or insufficient (referred to as "data-poor" scenarios). In order to make the model better understand natural language in data-poor domains, it is very necessary to overcome the challenge of data scarcity, as the available training samples are few or even zero. Therefore, the academic community proposed cross-domain named entity recognition to alleviate this problem, which aims to learn information from source domains to enhance named entity recognition performance in target domains[9]. Cross-domain named entity recognition methods aim to learn source domain knowledge from a large number of training samples in data-rich domains and apply it to data-poor domains. In this paper, data-poor domains refer to domains with scarce annotated data.

In addition, massive human mobility data can be collected from various sources, for example connected cars, mobile phones, also location-based social networks, etc. The utilization of such data can be of great significance for transportation operations, social economics, urban planning, and more by enabling the understanding and prediction of human mobility. However, due to the difference in text types between source and target domains and the limitation of labeled data, most traditional methods trained in a specific domain (source domain) are difficult to generalize to new domains (target domain). The current cross-domain named entity recognition under data-poor scenarios still faces two main challenges: 1. Due to the large difference in background between data-rich source domain and data-poor target domain, it is difficult to learn good representations applicable to the target domain from data-poor target domain; 2. Most cross-domain named entity recognition methods focus on solving the problem of transferring task knowledge from data-rich source domain to data-poor target domain, while the transfer of valuable label information is often not explicitly considered or even ignored.

This paper focuses on solving the above challenges encountered by cross-domain named entity recognition under data-poor scenarios in the deep learning framework. Therefore, this paper's research objective is to achieve cross-domain named entity recognition under data-poor scenarios, while alleviating the problems of not being able to learn good representations applicable to this domain from data-poor target domains and valuable label information transfer often not being explicitly considered or even ignored, so as to uncover the hidden knowledge of human mobility within the data. It effectively reduces the data demand of named entity recognition models in specific domains and has high practical value. Moreover, the proposed techniques contribute to the broader goal of deep understanding and modeling of complex data, such as leveraging advanced deep learning approaches for deep modeling and understanding of various domains such as human mobility, transportation, and social economics.

The main contributions of this paper are summarized as follows:

- To address the problem that pre-trained language models cannot learn good representations from data-poor target domains, this paper uses domain adaptive pre-training to mask language modeling on unlabeled text corpora in data-poor target domains to learn background and basic knowledge representations in target domains, and uses phrase-level span masking to increase the difficulty for pre-trained models, helping BERT better understand target domain text and thus better solve the problem of domain discreteness in task knowledge transfer. In addition, this paper proposes an adaptive text sequence augmentation module at-SAM, which helps the model enhance the association between entities in sentences within data-poor target domains and learn good representations applicable to this domain.
- To address the problem that valuable label information transfer is not explicitly considered, this paper proposes a prediction label guided dual-sequence perception information module Dual-SAM. This module uses an autoregressive framework and first uses Bi-LSTM to learn sequence information of predicted labels to capture potential relationships between labels; and uses a hybrid attention module to fuse token sequence representation from domain adaptive pre-trained model and label sequence information from Bi-LSTM to obtain token-label and label-token fusion information, which learns potential relationships between characters and labels in target domain and promotes model for domain adaptation; finally combines label sequence information from Bi-LSTM, token representation information from pre-trained model, token-label and label-token fusion information together and jointly participates in entity label prediction in data-poor target domain.
- This paper also introduces the architecture of the cross-domain named entity recognition model under data-poor scenarios and analyzes the effectiveness of the model; then, this paper conducts experimental verification of the cross-domain named entity recognition model under data-poor scenarios, compares it with several existing optimal methods and conducts generalization verification. In addition, this paper compares in detail the effects of the adaptive text sequence enhancement module at-SAM and the predicted label-guided dual sequence perception information module Dual-SAM on each category in the target domain, and achieves the best results so far.

Through this research, we hope to further promote the development of NER echnology and contributes to the broader goal of deep understanding and modeling of complex data, with applications in various domains, including human mobility, transportation, and social economics.

2 Related Work

Named Entity (NE) refers to a word or phrase that can be clearly distinguished from other entities with similar characteristics[10], which usually include common domains such as people, places, organizations, etc., and artificial intelligence domains such as algorithms, researchers, metrics, tasks, etc.[11]. Named entity recognition is the process of locating and classifying named entities in text into predefined entity categories. In addition to being an independent tool for information extraction, NER plays an important role in various natural language processing applications, such as text understanding[12][13], information retrieval[14][15], text summarization[16], question answering systems[17], machine translation[18], etc. In recent years, due to its great success in various domains, deep learning has attracted widespread attention and research in the academic community. Starting from Collobert et al.[9], deep learning-based named entity recognition methods have been flourishing.

Although the current deep learning-based named entity recognition models have achieved good performance, they require large-scale annotated training data to adapt to different domains. Therefore, cross-domain named entity recognition has received much attention in recent years. Cross-domain algorithms alleviate the data scarcity problem and improve the model's generalization ability to the target domain[19], making it one of the research hotspots in the NLP field. At present, the academic community has proposed many methods to enhance cross-domain NER. For example, Jia et al.[20] use parameter generation networks to perform cross-domain and cross-task knowledge transfer, and use multi-task learning to combine NER and language modeling tasks to extract domain-different knowledge from raw text, thus enhancing the model. Wang et al.[21] propose a label-aware dual transfer learning framework for cross-specialty NER, which enables a medical NER model designed for one specialty to be easily applied to another specialty. Wang and Kulkarni et al.[22] study different domain adaptation settings for the NER task. Liu et al.[23] introduce a coarse-to-fine framework Coach and a cross-lingual and cross-domain parsing framework X2Parser, where Coach decomposes the representation learning process into coarse-grained and fine-grained feature learning, X2Parser simplifies the hierarchical task structure into flatness, and their experiments show that simplifying the task structure makes representation learning more effective for data-poor languages and domains. Sachan et al.[24] study injecting target domain knowledge into language models with the same architecture for fast adaptation. In addition, Jia and Zhang[25] propose a multi-unit composite LSTM structure that enhances cross-domain NER by merging entity types through separate unit states. Liu et al.[26] propose continuing pre-training language models on target domain-related corpora. Hu et al.[11] propose a model that improves label information transfer by predicting current labels together with corresponding tags and previous labels, providing a simple but effective solution for cross-domain NER.

Recent work on large-scale pre-trained language models (LLM), such as GPT-3[27], InstructGPT[28] and ChatGPT[29][30][31], shows that LLM performs well on various downstream tasks, even without adjusting parameters, using only a few examples as instructions. For example Wei et al.[32] propose a two-stage framework (ChatIE) that transforms zero-shot information extraction tasks into multi-turn question answering problems. Leveraging ChatGPT's powerful capabilities, they extensively evaluate the proposed framework on three information extraction tasks: entity relation triple extraction, named entity recognition and event extraction. Experimental results on six datasets in two languages show that ChatIE achieves impressive performance on several datasets, even surpassing some full-fledged models.

3 Cross-domain Named Entity Recognition based on Domain Adaptation

In this section, we first explored different masking methods in the pre-training process of the BERT model, which provides more challenging objectives for the model in self-supervised tasks, enabling it to generate better sequence representations and improve the quality of contextual representations. We then introduced domain-adaptive pre-training, which addresses the issue of poor performance of general pre-trained models on named entity recognition tasks in specific domains, and presented an adaptive text sequence augmentation module that helps the model enhance the relationships between entities in sentences within data-poor target domains, thereby learning good representations applicable to the domain.

3.1 Pre-training Model based on Phrase-level Span Masking

Pre-trained models such as BERT have shown strong performance improvements using self-supervised training by masking individual words or subword units. However, named entity recognition tasks involve predicting sequence information across two or more text spans. For example, in the named entity recognition problem, "Beijing Haidian District" is a longer place name. Such spans provide more challenging targets for self-supervised tasks, as in the previous example, predicting only "Beijing" when knowing the next word is "Haidian District" is much simpler than predicting "Beijing Haidian District".

Inspired by Joshi et al., this [27]paper changes the token-level masking in BERT to span-level masking. In BERT, the masked language model (MLM) first randomly masks a total of 15% of tokens, then replaces 80% of masked tokens with a special token ([MASK]), replaces 10% with random tokens, and replaces the remaining 10% with original tokens. Except for the first masking step, this paper follows the same masking strategy as BERT. In the first step, after random masking, this paper moves a single masking index position to an adjacent position adjacent to another masking index position to generate more masking vocabulary spans. Therefore, the model can learn to predict the entire masking span based on the markers observed at its boundaries. The span-based masking enforced by the word group span-based masking model only uses the context in which they appear to predict the entire word group span. For example, the randomly masked sentence:

"Yu Yuan Tan [MASK] is located in the southern part of Beijing [MASK]."

will become:

"Yu Yuan Tan Park is located in the southern part of [MASK][MASK]."

Intuitively, the word group span-based masking method provides more challenging tasks for pre-trained language models, as in the example of "Beijing Haidian District" mentioned above. Therefore, word group span-based masking can help BERT better understand domain texts, allowing the model to generate better sequence representations and better solve domain-specific problems in knowledge transfer.

3.2 Named Entity Recognition Model Based on Domain Adaptive Pre-training

Traditional pre-training models are usually pre-trained on large-scale general language corpora and then applied to specific tasks through fine-tuning. However, in specific domains, due to different language patterns and structures, these general pre-training models often cannot achieve optimal performance.

Domain adaptive pre-training refers to pre-training in a specific domain, which can combine existing domain knowledge with the general knowledge of language models to improve performance in that domain. Compared with general pre-training models, domain adaptive pre-training models can better adapt to specific domain tasks, thus achieving better performance.

3.2.1 Domain Adaptive Pre-training Model Structure

The domain adaptive pre-training model used in this article is based on the BERT model. By adapting to specific domains through pre-training on unlabeled data in that domain, an input sequence encoder suitable for a specific domain is obtained, which in turn generates distributed representations of word-level or character-level inputs. The structure of the domain adaptive pre-training model is shown in Figure1. In terms of implementation, the domain adaptive pre-training model usually adopts a two-stage training method. First, the language model is pre-trained on a general corpus, which results in a general text encoder. Then, the model is pre-trained again on unlabeled data in a specific domain. However, it should be noted that the word span masking method introduced in Section 3.1 is used in this article. This masking method improves BERT's understanding of text on the one hand, and matches the characteristics of entity spans in named entity recognition tasks on the other hand. This results in an encoder suitable for specific domain tasks, which better solves the problem of domain discreteness in knowledge transfer for tasks.

3.2.2 Adaptive Text Sequence Enhancement Module

In data-rich domains, the background may differ significantly from that of data-poor target domains, making it difficult to learn good representations for the target domain from the data-poor domain. In this paper, we propose an adaptive text sequence enhancement module that selects the appropriate depth of Transformer encoder based on the length of the input sequence, enhancing the contextual correlation of the distributed representation of text that has been pre-trained for domain adaptation, and helping the model to strengthen the correlation between entities in sentences in the data-poor target domain, thereby learning good representations for the target domain. Generally, the longer the input text sequence, the more entities there are in the

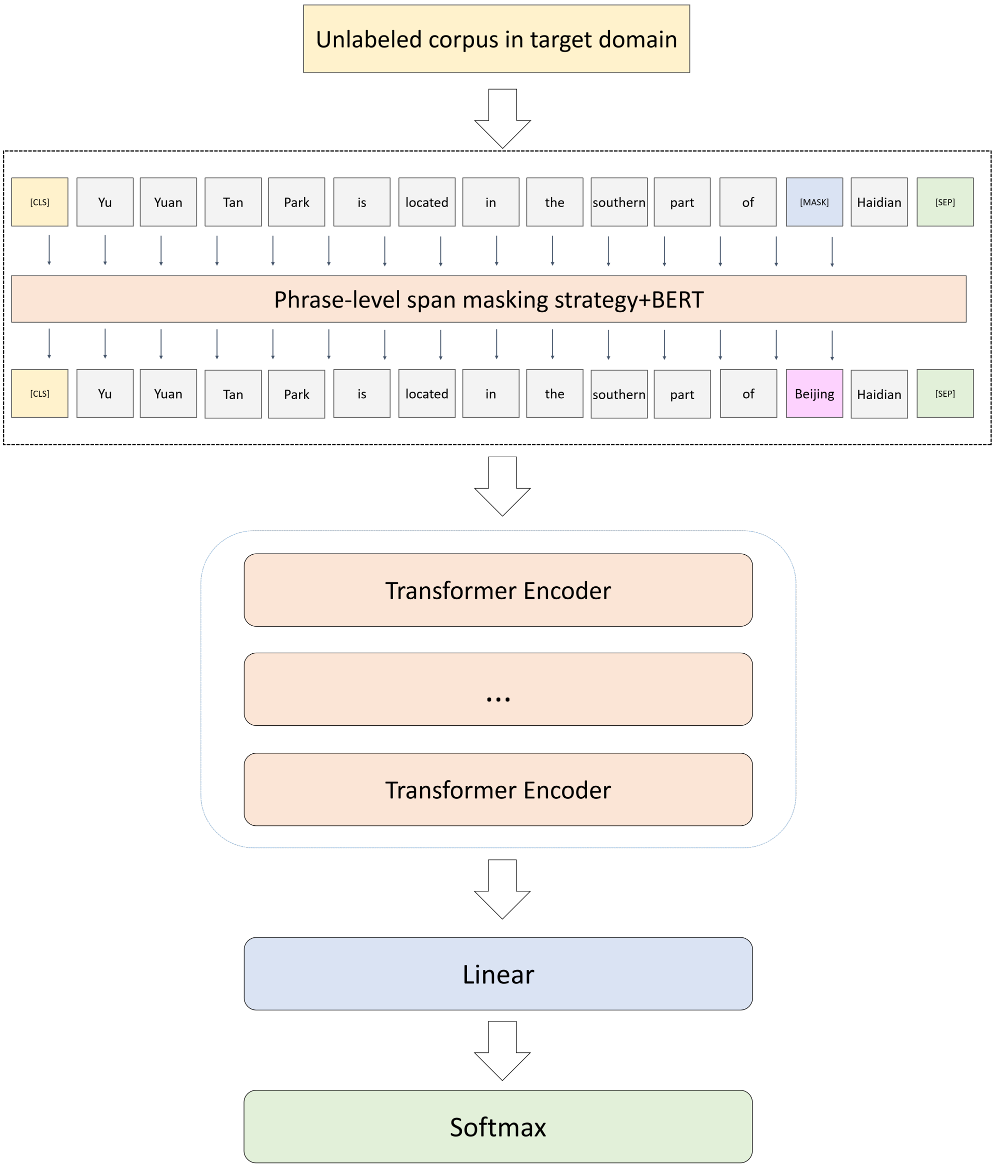

Fig. 1 Domain-adaptive pre-training architecture diagram

sequence, and the more correlation there is between entities in a single text sequence. To obtain the implicit connections between these entities, a more complex network is needed. Specifically, the proposed adaptive text sequence enhancement module sets a perception coefficient L to represent the length of the input BERT text sequence. The adaptive text sequence enhancement module is composed of m Transformer encoders stacked together, where the relationship between L and m is shown in Equation (1):

$$m = \frac{L}{128} \tag{1}$$

When the input length is long, more layers of encoders can capture higher-level semantic information between entities in longer sentences, providing reference for entity category prediction.

3.2.3 Named Entity Recognition Model Based on Domain Adaptive Pre-training and Text Sequence Augmentation

This section combines the domain adaptive pre-trained model and the adaptive text sequence augmentation module obtained in 3.2.1 and 3.2.2, and presents the model architecture based on domain adaptive pre-training and text sequence augmentation module as shown in Figure 2. Specifically, the original input sequence of the text is $X = \{x_1, x_2, ..., x_n\}$, where n represents the total number of vocabulary or subwords in the input sentence. The input sequence is fed into the large model trained through domain adaptive pre-training to obtain the distributed representation of the input sequence, denoted as $\hat{X} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_n\}$. Then, the structure of the adaptive text sequence enhancement module is determined by formula (1), and the distributed representation of the input sequence $\hat{X}$ is passed through the adaptive text sequence enhancement module, and then the distributed representation $\bar{X} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\}$ that implies the contextual connection between entities in the sentence is obtained. The transformation relationship is shown in formula (2). Next, before predicting for each type of named entity label, $\bar{X}$ needs to pass through a fully connected layer to map the vector dimension to a dimension consistent with the number of category labels, as shown in formula (3). Finally, the output vector is mapped to the most likely named entity label by Softmax layer.

$$\bar{X} = enc(\ldots enc(\hat{X})) \tag{2}$$

The depth of the $\ldots enc()$ function nesting is the number of stacked Transformer encoders in the adaptive text sequence enhancement module. This number is denoted as m.

$$\tilde{X} = W_1 \cdot \bar{X} + b_1 \tag{3}$$

Among them, W_1 and b_1 are learnable parameters.

During the training process, this paper first conducts domain adaptive pre-training on the domain-related unlabeled dataset.Then, the labeled data in the source domain is fine-tuned to enable the model to learn general knowledge in named entity tasks. Finally, the labeled data in the target domain is fine-tuned to obtain a named entity recognition model that focuses on entities in the target domain.

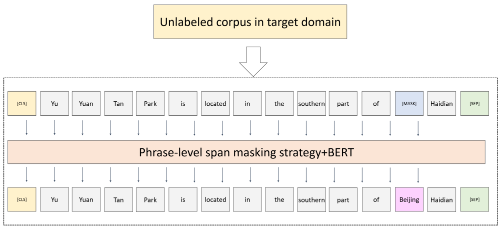

Fig. 2 Model architecture based on domain adaptive pre-training and text sequence enhancement module

4 Cross-domain Named Entity Recognition with Autoregressive Model based on Predicted Labels

In this section, we first introduce the autoregressive model used in the cross-domain named entity recognition model that combines predicted labels, including the representation method of predicted label information and the architecture of the autoregressive model. Next, we introduce the cross-domain named entity recognition model that combines predicted labels and provide a detailed explanation of the proposed Dual-SAM, as well as further explanation of the fused information from multiple perspectives obtained by input sequence and predicted label sequence mining. Finally, we conduct experimental verification and result analysis of the proposed method, and analyze the effectiveness of the improvement method. The experimental results show that the model proposed in this paper is effective in integrating predicted label information.

4.1 A self-regressive cross-domain named entity recognition model combined with predicted labels

In cross-domain named entity recognition, label information is particularly crucial because predicted label information can help the model distinguish differences and similarities between different domains[11]. For example, in the general domain, ”Xu Zechen” is usually a ”person” entity, while in the literature domain, for the sentence ”The novel ’Beishang’ won the 10th Mao Dun Literature Award, written by the writer Xu Zechen,” ”Xu Zechen” is a ”writer” entity. Therefore, if the NER model recognizes that the phrase ”Beishang” is a literature domain entity, the model can pay more attention to this phrase, making the model more likely to predict ”Xu Zechen” as a ”writer” label in the literature domain. This previously predicted label information can help the model enhance the relationship between the predicted entity and the specific domain, rather than relying solely on the encoder itself. On the other hand, the label-label relationship between predicted labels and predicted labels also supports cross-domain named entity recognition, similar to the implicit semantic relationship between entities in the token sequence. For example, for the sentence ”Jack Sully in the movie ’Avatar’ is played by Sam Worthington,” when the model identifies ”Jack Sully” as a ”character name” label in the movie domain, the model will be more concerned with the ”actor” category entity, making it easier to predict ”Sam Worthington” as an ”actor” because the entity labels ”character name” and ”actor” have semantic similarity, and the model will pay more attention to the ”actor” entity through attention mechanism. This is very useful for named entity recognition tasks within specific domains.

To address the issue of the transfer of valuable label information in previous research that has not been explicitly considered, this paper proposes a predicted label-guided dual sequence-aware information module that uses an autoregressive framework to capture the self-sequence information of predicted labels through Bi-LSTM to learn the potential relationship between labels. It then uses a hybrid attention module to fuse token representations from domain-adaptive pre-training models and label sequence information from Bi-LSTM to obtain token-label and label-token fusion information to learn the potential relationship between characters and labels in the target domain and promote domain adaptation of the model. Finally, the label sequence information from Bi-LSTM, token representation information from pre-training models, and token-label and label-token fusion information are combined to jointly participate in predicting entity labels in data-poor target domains.

4.1.1 Model Architecture

The autoregressive cross-domain named entity recognition model proposed in this article combines the domain-adaptive pre-training and text sequence enhancement based named entity recognition model in Section 3.2, and uses the domain-adaptive pre-training based on word-level span masking and text sequence enhancement module to learn knowledge representation applicable to the target domain. Then, this article proposes a prediction label-guided dual sequence perception information module to capture and merge four parts of information: (1) token representation information from the pre-training model; (2) label sequence information from Bi-LSTM; (3) fusion information between token and label; and (4) fusion information between label and token. The specific architecture of the model is shown in Figure 3:

Specifically, when predicting the entity label for the i token, the model has two inputs: the first part is the input sequence $X = \{x_1, x_2, \ldots, x_N\}$ of the target domain, and the second part is the predicted label sequence $Y = \{y_1, y_2, \ldots, y_{i-1}\}$. The input sequence of the target domain is pre-trained with domain adaptation BERT to obtain the semantic vector representation of each input token, denoted as $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$, which is $T = \{t_1, t_2, \ldots, t_N\}$ in Figure 3. For the predicted label sequence, the label encoding table needs to be queried to obtain the label encoding vector of each predicted label, which is then input into the bidirectional LSTM to capture the long-term dependency relationship between entities in the label sequence, and output a sequence $L = \{l_1, l_2, \ldots, l_{i-1}\}$ containing the dependency relationship and semantic information of entities in the label.

Next, in order to obtain the enhanced information about the relationship between the hidden predicted entity and the specific domain background (such as the relationship between "writer" and the Literature domain mentioned in Section 4.1) and the enhanced information about the relationship between the hidden predicted entity and the predicted label (such as the relationship between "character name" and "actor" mentioned in Section 4.1), this article uses a hybrid attention module for querying. For the enhanced information about the relationship between the hidden predicted entity and the specific domain background, the last layer output l_{i-1} in sequence L

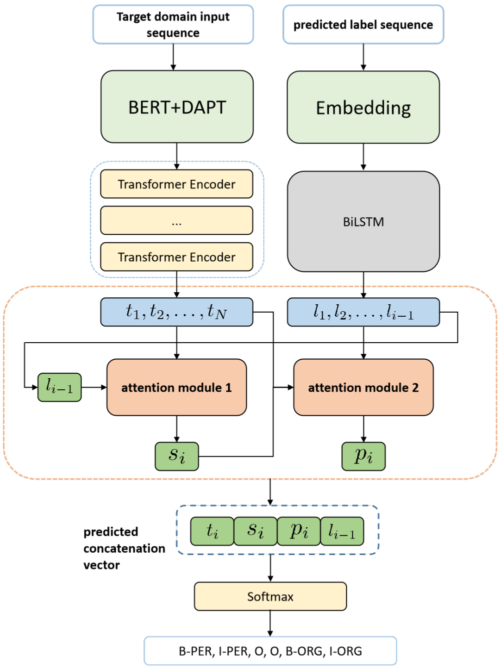

Fig. 3 Architecture diagram of self-regressive cross-domain named entity recognition model combined with predicted labels

needs to serve as the query vector in attention module 1, which is first projected to the same dimension as t_i, and then the attention weight w_i^1 is calculated between l_{i-1} and $T = \{t_1, t_2, \ldots, t_N\}$.

$$\bar{l}_{i-1} = W_2 l_{i-1} + b_2 \tag{4}$$

$$w_i^1 = \text{Softmax}(\frac{\bar{l}_{i-1} T^T}{\sqrt{d_t}}) \tag{5}$$

Among them, d_t represents the dimension of vector t_i. W_2 and b_2 are both learnable parameters, and then an enhanced representation of the relationship between the hidden entity to be predicted and the specific domain background needs to be calculated through attention weight.

$$s_i = \sum_{j=1}^{N} w_{i,j}^1 T_j \tag{6}$$

For the enhanced information that involves the relationship between the implicit entity to be predicted and the predicted label, it is necessary to first concatenate the enhanced representation s_i of the implicit entity to be predicted with the representation t_i of the i-th input token that corresponds to the specific domain context, and project it to the same dimension as vector l_{i-1} for subsequent querying in attention module 2, as shown in formulas 7-9.

$$\bar{s}_i = W_3[s_i|t_i] + b_3 \tag{7}$$

Both W_3 and b_3 are learnable parameters, and the calculation method is similar to that of the attention module. However, in this case, $\bar{s}$ is used as the query vector, and the attention weight w_i^2 is calculated by $\bar{s_i}$ and $L = \{l_1, l_2, \ldots, l_{i-1}\}$:

$$w_i^2 = \text{Softmax}(\frac{\bar{s}_i L^T}{\sqrt{d_l}}) \tag{8}$$

$$p_i = \sum_{j=1}^{i-1} w_{i,j}^2 L_j \tag{9}$$

Among them, d_l represents the dimension of vector l_{i-1}, and p_i represents the enhanced representation of the relationship between the hidden entity to be predicted and the predicted label. Finally, the prediction label-guided bidirectional sequence perception information module concatenates the enhanced representations s_i of the hidden entity to be predicted and its specific domain background relationship, the enhanced representation p_i of the relationship between the hidden entity to be predicted and the predicted label, the sequence information representation t_i of the token, and the sequence representation l_{i-1} of the predicted label, and maps them to the dimension of all labels in the domain, as shown in formula 10:

$$y_i = \text{Softmax}(W_4[t_i|s_i|p_i|l_{i-1}]^T + b_4) \tag{10}$$

W_4 and b_4 are both learnable parameters, from which specific named entity labels in each input token's predicted domain can be obtained.

4.1.2 Analysis of Dual Sequence-Aware Information Module Guided by Predicted Labels

As entity label prediction is an autoregressive process, this paper uses all predicted labels before the current prediction stage, so there is no issue of label information leakage. Regarding the four parts of information involved in the final classification in the autoregressive model, they can be understood from the following perspectives: compared to some of the latest methods that use retrieved external resources for information supplementation [33], this method focuses more on mining internal information in the input sequence, combined with the predicted label sequence. The model can utilize two types of sequence information, namely, the original input token sequence in the target domain and the already predicted label sequence before the current prediction position. Therefore, four parts of information can be mined: (1) the representation information of the token sequence obtained through domain-adaptive pre-training of BERT; (2) the sequence information captured by Bi-LSTM in the already predicted label sequence; (3) the enhanced information about the relationship between the target entity and the specific domain background (the representation of the already predicted label sequence is queried in the token sequence through attention mechanism); (4) the enhanced information about the relationship between the target entity and the already predicted label (the representation of the token sequence is queried in the already predicted label sequence through attention mechanism).

5 Experiments

5.1 Settings

In order to verify the cross-domain named entity recognition model proposed in this article under data-poor scenarios, experiments were conducted in five domains: Politics, Science, Music, Literature, and AI. Generalization experiments were also conducted on the Movie and Restaurant datasets. However, it should be noted that since a large amount of domain-specific data was not collected in the Movie and Restaurant domains, domain-adaptive pre-training was not used in the experiments in these two domains. The specific parameters used in the experiments in this chapter are consistent with Table 1. In addition, Adam optimizer was used to optimize all training parameters in this article. Figure 4 shows the change of model loss during the model training process. It can be seen that the cross-domain named entity recognition model proposed in this article has good convergence and can fit the target domain data well under data-poor scenarios.

Table 1 Training parameters of cross-domain named entity recognition model with autoregressive prediction labels

Parameters	parameter values
batch size	16
early stop	30
maximum training batch size	300
learning rate	$5e-5$
BERT model name	bert-base-cased
vocabulary distributed representation dimension	768
label embedding dimension	100
LSTM hidden layer dimension	100

5.2 Results and Analysis

As shown in Table 2, the experimental results indicate that † denotes data from Hu et al. [12], and ‡ denotes data from Tang et al. [30]. The table presents the actual performance of the data-poor cross-domain named entity recognition model LANER+at-SAM+Dual-SAM, which integrates the adaptive text sequence enhancement module at-SAM and the dual-sequence perception information module Dual-SAM with label-guided prediction, in seven different target domains related with human mobility(the last two domains, Movie and Restaurant, are used for generalization verification), and compares it with existing excellent cross-domain named entity recognition models. The entire experiment is divided into two parts: one group is models that do not use domain adaptive pre-training, including FLAIR [31], BARTNERBASE [32], DoSEA [30], and LANER [12], where FLAIR uses the internal state of the character language model to generate context string embeddings and integrates them into the NER model; BARTNERBASE formalizes the NER task as an entity span

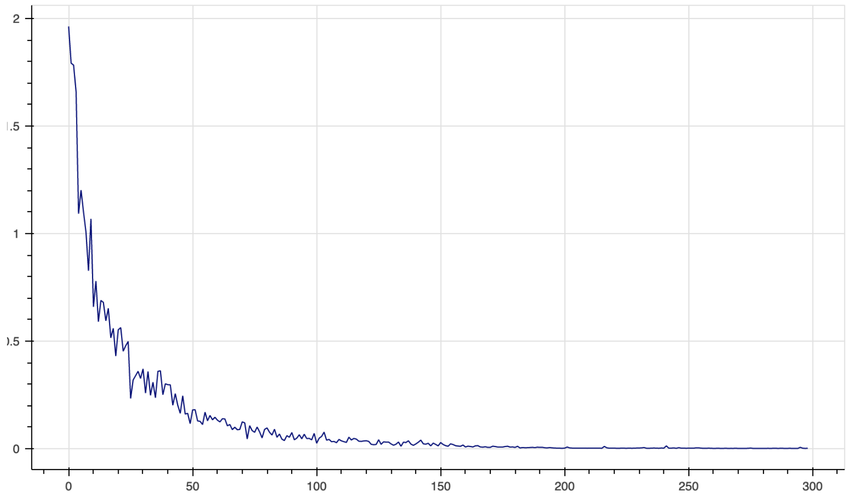

Fig. 4 Model training loss function change chart

sequence generation problem and uses BART [34] as its backbone network; DoSEA proposes a new framework based on machine reading comprehension, which introduces entity existence discrimination tasks and entity perception training settings to identify inconsistent entity annotations in the source domain and bring additional references for better cross-domain information sharing. The other group of experiments uses domain adaptive pre-training models, including MULTI-CELL LSTM [25] and LANER, where MULTI-CELL LSTM uses a multi-unit combined LSTM structure to enhance NER domain adaptability, and LANER is a cross-domain named entity recognition framework based on a label-aware autoregressive model to solve the label transfer problem in cross-domain scenarios. It is worth noting that DoSEA and LANER are currently the best models in the field of cross-domain named entity recognition. In addition, the performance of each model is calculated by the F1-score (%) of each model in each domain, and the sixth column (Avg.) is the average F1-score of each model in the first five domains.

From Table 2, it can be seen that in all model settings, the proposed model LANER+at-SAM+Dual-SAM outperforms existing models in most domains in terms of F1-score. Compared with the best LANER model, it improves the average performance in all domains by 1.32%. For DoSEA, which achieves the best results in the Politics domain, this model improves the average performance in all domains by up to 2.64%.

Specifically, for the first set of experiments without using domain-adaptive pre-training, the model LANER+at-SAM+Dual-SAM outperforms FLAIR and BARTNER$_{BASE}$ in all domains with absolute advantages: the improvement over FLAIR ranges from 2.54% to 10.64%, and the improvement over BARTNER$_{BASE}$ ranges from 2.18% to 11.32%. For the DoSEA model with different architectures, even without domain-adaptive pre-training, the LANER+at-SAM+Dual-SAM model still has advantages in the Music and Literature domains. For the LANER model

Table 2 Experimental Results and Comparison of Cross-Domain Named Entity Recognition in data-poor Scenarios

Method	Poli.	Scie.	Music	Lite.	AI	Avg.	Movie	Rest.
not using DAPT								
FLAIR† [31]	69.54	64.71	65.60	61.35	52.48	62.73	-	-
BARTNER$_{BASE}$† [32]	69.90	65.14	65.35	58.93	53.00	62.46	71.55	79.53
DoSEA‡ [35]	**75.52**	71.69	73.10	68.59	66.03	70.99	-	-
LANER†	71.65	69.29	73.07	67.98	61.72	68.74	72.41	80.55
LANER+at-SAM+Dual-SAM	72.08	70.16	75.14	70.25	63.12	70.15	**73.12**	**81.07**
using DAPT								
MULTI-CELL LSTM† [25]	71.45	67.68	74.19	68.63	61.64	68.71	-	-
LANER†	74.06	71.83	78.78	71.11	65.79	72.31	-	-
LANER+at-SAM+Dual-SAM	75.12	**73.58**	**79.39**	**72.93**	**67.15**	**73.63**	-	-

with similar architecture, this model has different degrees of F1 value improvement in all domains, indicating that the LANER+at-SAM+Dual-SAM model can still play the advantages of the adaptive text sequence enhancement module at-SAM and the dual-sequence perception information module Dual-SAM without domain-adaptive pre-training, which can learn good representations applicable to the target domain from data-poor target domains and alleviate the problem of valuable label information transmission that has not been explicitly considered.

In the second set of experiments using domain-adaptive pre-training, compared with models that combine external resources or introduce complex training designs, the proposed model LANER+at-SAM+Dual-SAM has the advantage of practical simplicity. For example, MULTI-CELL LSTM combines entity type prediction and attention score guidance with NER tasks and applies multi-task learning. In contrast, the LANER+at-SAM+Dual-SAM model can achieve better results using a simpler method: the LANER+at-SAM+Dual-SAM model outperforms MULTI-CELL LSTM in the average F1-score in all domains by up to 4.92%, but the LANER+at-SAM+Dual-SAM model only needs to train the model in the source domain and then fine-tune it to the target domain, thereby reducing the demand for additional resources. Compared with the LANER model with similar architecture, the LANER+at-SAM+Dual-SAM model achieves F1-score improvement in all domains, with an average F1-score improvement of 1.32% in all domains, further demonstrating that this model still has significant advantages over existing optimal models in the cross-domain named entity recognition field. Comparing the first and second sets of experiments, when domain-adaptive pre-training is introduced, both the LANER model and the LANER+at-SAM+Dual-SAM model further improve their performance, indicating that domain-adaptive pre-training can help the model better understand the target domain by providing rich domain background information contained in domain-specific corpora, thereby narrowing the difference between the source and target domains.

In the generalization experiments of the model, it can be seen that the LANER+at-SAM+Dual-SAM model achieves the best performance in both the Movie and

Restaurant domains, indicating that the proposed model can effectively transfer source domain knowledge to the target domain. Even in cases where there may be significant differences between the data-rich source domain and the data-poor target domain, the model can still learn representations that are applicable to the target domain and effectively promote the transfer of valuable label information to the target domain. However, the improvement of the LANER+at-SAM+Dual-SAM model in the Movie and Restaurant domains compared to LANER is not as significant as in the other five domain datasets, because the Movie and Restaurant domains do not share common entity types with the source domain dataset CoNLL-2003, resulting in greater differences in label information between the source and target domains, making label information transfer more difficult.

5.3 Fine-grained Experimental Results and Analysis

The detailed performance of various categories in the Music and Literature fields is shown in Tables 3 and 4.

Table 3 Detailed experimental results of various categories in the field of music.

Category	Precision	Recall	F1-score	Sample size
album	0.7018	0.8223	0.7573	332
award	0.8421	0.8615	0.8517	260
band	0.8575	0.8593	0.8584	462
country	0.8232	0.9313	0.8739	160
event	0.5658	0.6232	0.5931	69
location	0.8550	0.8426	0.8488	343
miscellaneous	0.4396	0.2339	0.3053	171
musical artist	0.8043	0.9495	0.8709	515
musical instrument	0.8000	0.1818	0.2963	22
music genre	0.7922	0.8378	0.8144	487
organization	0.8502	0.8462	0.8482	208
person	0.5000	0.1125	0.1837	80
song	0.7273	0.7401	0.7336	227

Firstly, from Table 3, it can be seen that the categories with better performance in the Music field are **award**, **band**, **country**, **location**, **musical artist**, **music genre**, and **organization**, with F1-scores all above 0.8. The categories with poorer performance are **event**, **miscellaneous**, **musical instrument**, and **person**, with F1-scores all below 0.6. The performance of the categories **album** and **song** is moderate, with F1-scores of 0.7573 and 0.7366, respectively.

Secondly, by analyzing the relationship between each category and its sample size, it can be found that categories with fewer samples have poorer performance, while categories with more samples have better performance. For example, except for the category **country**, the number of other category labels in the test set is above 260 for F1-scores above 0.8, while the number of category labels in the test set is below

Table 4 Detailed experimental results of various categories in the field of literature.

Category	Precision	Recall	F1-score	Sample size
award	0.8784	0.9220	0.8997	141
book	0.6842	0.8397	0.7540	418
country	0.7500	0.9505	0.8384	101
event	0.6970	0.5111	0.5897	45
literary genre	0.6175	0.5825	0.5995	194
location	0.7079	0.6364	0.6702	99
magazine	0.9057	0.8421	0.8727	57
miscellaneous	0.5307	0.3975	0.4545	239
organization	0.7018	0.7273	0.7143	110
person	0.6154	0.4114	0.4932	175
poem	0.7011	0.5083	0.5894	120
writer	0.8147	0.9383	0.8721	567

171 for F1-scores below 0.6. This indicates that the model's performance is affected by both data imbalance and the information that can be conveyed by the labels. If the number of samples in a certain category is relatively small, it can be reasonably inferred that the unlabeled corpus and the training set also have relatively few data for that category label, which will reduce the label information transmitted from the source domain to the target domain and affect the prediction of that category label.

Furthermore, by analyzing the relationship between each category, it can be seen that the sample size and F1-score of the **musical artist**, **music genre**, and **band** categories are relatively close because they are semantically related. In sentences involving a **musical artist**, it is common to mention the **music genre** and **band** to which the artist belongs. Similarly, there is a close semantic relationship between **album** and **song**, so the F1-scores of these two categories are also relatively close. In addition, from Table 4, it can be seen that due to the limitations of the dataset, the model's performance on non-shared entity types between the source and target domains is not ideal. In cross-domain named entity recognition, there may be overlapping label ranges in the dataset, such as the **person** category in CoNLL-2003 and the **writer** category in the Literature field. "Shakespeare" is a **person** category in the CoNLL-2003 dataset, while it is a **writer** category in the CrossNER dataset. Therefore, the label information learned from the source domain may lead to misclassification of **person** in the target domain. Similarly, the categories **country** and **location** are easily confused, which explains the poor performance in predicting the entity category **location**.

Finally, this paper provides a detailed comparison of the effects of various ablation experimental models in the Science field on specific labels, as shown in Figure 5. Since the model proposed in this paper, LANER+at-SAM+Dual-SAM, has a similar autoregressive architecture to the baseline model LANER, and LANER is the best model in current cross-domain named entity recognition, this part of the experiment is only compared with LANER. Specifically, this figure shows the actual performance of the cross-domain named entity model LANER+at-SAM+Dual-SAM with the data-poor adaptive text sequence enhancement module at-SAM and the dual-sequence perception information module Dual-SAM guided by predicted labels in the Science field, and compares it with the models LANER+at-SAM, LANER+Dual-SAM, and the baseline model LANER, which use only the adaptive text sequence enhancement module at-SAM and the dual-sequence perception information module Dual-SAM guided by predicted labels, respectively. The horizontal axis of the figure represents each specific category in the Science field, including **academic**, **journal**, **astronomical object**, **award**, **chemical compound**, **chemical element**, **country**, **discipline**, **enzyme**, **event**, **location**, **miscellaneous**, **organization**, **person**, **protein**, **scientist**, **theory**, and **university**, and the vertical axis represents the F1-score. In addition, the performance of each model is calculated by the F1-score of each model in each category. From Figure 5, it can be seen that for the model LANER+at-SAM, its F1-score

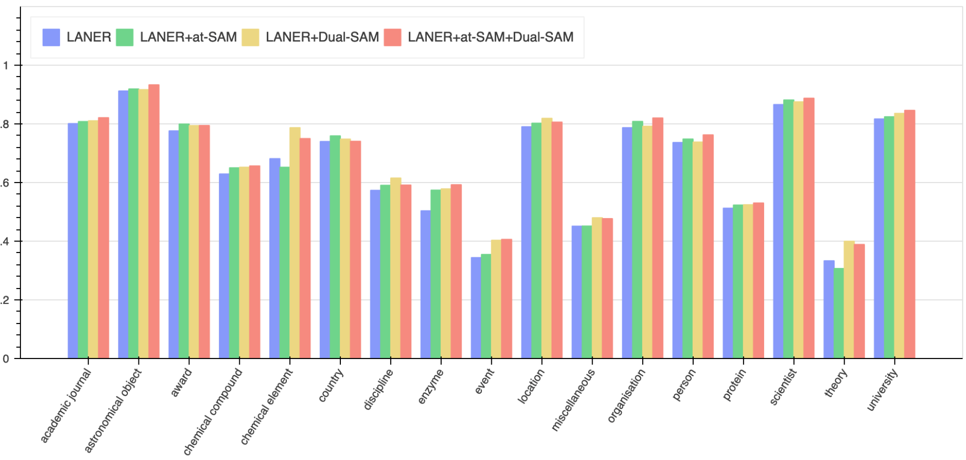

Fig. 5 Comparison chart of the actual effects of various innovation points on specific labels in the field of Science.

in most categories (except for the **chemical element** and **theory** categories) is higher than that of LANER, indicating that the adaptive text sequence enhancement model at-SAM can help the model enhance the correlation between entities in sentences in the data-poor target domain, help capture higher-level semantic information between entities in sentences, and learn good representations applicable to this field, providing reference for entity category prediction. For the model LANER+Dual-SAM, its effect is better than that of LANER in all categories, and the improvement of the module Dual-SAM for categories with poor original effects is more obvious, such as **chemical element**, **discipline**, **enzyme**, **event**, and **theory** categories. This indicates that by guiding with predicted label information, more fusion information beneficial to cross-domain named entity recognition can be mined from token sequences and predicted label sequences, and these information is more beneficial for predicting categories with generally poor original effects. In addition, in all categories, the effect of the model LANER+at-SAM+Dual-SAM is better than that of the baseline model LANER, which indicates the effectiveness and compatibility of the modules at-SAM and Dual-SAM integrated into the named entity model, and for some special entity categories, such as **chemical element**, **location**, **miscellaneous**, and **theory**, it can be seen that when the module at-SAM cannot play a positive role, the module Dual-SAM can supplement the model through additional information mined from the label sequence, thus making the effect of the model LANER+at-SAM+Dual-SAM ultimately better than that of the baseline model LANER.

By focusing on cross-domain named entity recognition in data-poor scenarios, this study explores the application of deep learning techniques, such as adaptive text sequence enhancement and label-guided prediction, to enhance the performance of named entity recognition models in different target domains. The proposed LANER+at-SAM+Dual-SAM model not only demonstrates state-of-the-art performance but also provides insights into the effective transfer of knowledge in data-poor environments. These findings contribute to the broader goal of leveraging deep learning technologies, such as graph neural networks, recurrent neural networks, and transformers, to better model and understand human mobility data.

6 Conclusion

This article focuses on the research of cross-domain named entity recognition methods in data-poor scenarios, aiming to alleviate the limitations of existing methods caused by differences in text types between source and target domains, as well as limited labeled data, which often lead to ineffective learning of knowledge representation in data-poor environments and neglect of valuable label information transmission in most cross-domain named entity recognition methods. This article mainly completes the following work:

1) To address the problem that pre-trained language models cannot learn good representations from data-poor target domains, this article uses domain adaptive pre-training to mask language modeling on unlabeled text corpus in the target domain to learn the background and basic knowledge representation in the target domain. It also increases the difficulty of pre-training models through word-level span masking to help BERT better understand target domain text, thereby better solving the domain discreteness problem in task knowledge transfer. In addition, this article proposes an adaptive text sequence enhancement module, at-SAM, to help the model enhance the association between entities in sentences in the data-poor target domain and learn representations suitable for this domain.
2) To address the problem that valuable label information transmission is not explicitly considered, this article proposes a dual-sequence perception information module, Dual-SAM, guided by predicted labels. This module uses a self-regressive framework to first learn the sequence information of the predicted labels themselves through Bi-LSTM to capture the potential relationship between labels. It then uses a hybrid attention module to fuse token sequence representations from domain adaptive pre-training models and label sequence information from Bi-LSTM to obtain token-label and label-token fusion information, thereby learning the potential relationship between characters and labels in the target domain and promoting domain

adaptation. Finally, the label sequence information from Bi-LSTM, token representation information from pre-training models, token-label, and label-token fusion information are combined to predict entity labels in the data-poor target domain.

Declarations

Competing Interests

The authors declare that they have no conflict of interest regarding the publication of this manuscript.

Authors' contributions

Yutong Jiang: Methodology, Conceptualization ,Investigation, Visualization, Writing, Resources.
Fusheng Jin: Funding acquisition, Conceptualization, Supervision, Data curation, Methodology, Writing review & editing.
Mengnan Chen: Data curation, Methodology, Validation, Visualization, Writing.
Guoming Liu: Writing, Validation.
He Pang: Writing, Supervision, review & editing.
Ye Yuan: Writing review & editing.

Funding

This study was funded by National Natural Science Foundation of China(No. 62272045).

Availability of data and materials

The data that support the findings of this study are openly available at *https : //github.com/jinpeng01/LANER/tree/main/ner_data*.

References

[1] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[3] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[4] Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven pretraining of self-attention networks. CoRR **abs/1903.07785** (2019) 1903.07785

[5] Liu, T., Yao, J.-G., Lin, C.-Y.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307 (2019)

[6] Jie, Z., Lu, W.: Dependency-guided lstm-crf for named entity recognition (2019)

[7] Xia, C., Zhang, C., Yang, T., Li, Y., Du, N., Wu, X., Fan, W., Ma, F., Yu, P.: Multi-grained named entity recognition. arXiv preprint arXiv:1906.08449 (2019)

[8] Liu, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., Zhou, J.: Gcdt: A global context enhanced deep transition architecture for sequence labeling. arXiv preprint arXiv:1906.02437 (2019)

[9] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)

[10] Sharnagat, R.: Named entity recognition: A literature survey. Center For Indian Language Technology, 1–27 (2014)

[11] Hu, J., Zhao, H., Guo, D., Wan, X., Chang, T.-H.: A label-aware autoregressive framework for cross-domain ner. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 2222–2232 (2022)

[12] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)

[13] Cheng, P., Erk, K.: Attending to entities for better text understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7554–7561 (2020)

[14] Cowan, B., Zethelius, S., Luk, B., Baras, T., Ukarde, P., Zhang, D.: Named entity recognition in travel-related search queries. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, pp. 3935–3941 (2015)

[15] Brandsen, A., Verberne, S., Lambers, K., Wansleeben, M.: Can bert dig it? named entity recognition for information retrieval in the archaeology domain. Journal on Computing and Cultural Heritage (JOCCH) **15**(3), 1–18 (2022)

[16] Khademi, M.E., Fakhredanesh, M.: Persian automatic text summarization based on named entity recognition. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 1–12 (2020)

[17] Mollá, D., Van Zaanen, M., Smith, D.: Named entity recognition for question answering. In: Proceedings of the Australasian Language Technology Workshop 2006, pp. 51–58 (2006)

[18] Li, Z., Qu, D., Xie, C., Zhang, W., Li, Y.: Language model pre-training method

in machine translation based on named entity recognition. International Journal on Artificial Intelligence Tools **29**(07n08), 2040021 (2020)

[19] Liu, Z., Winata, G.I., Fung, P.: Zero-resource cross-domain named entity recognition. arXiv preprint arXiv:2002.05923 (2020)

[20] Jia, C., Liang, X., Zhang, Y.: Cross-domain ner using cross-domain language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2464–2474 (2019)

[21] Wang, Z., Qu, Y., Chen, L., Shen, J., Zhang, W., Zhang, S., Gao, Y., Gu, G., Chen, K., Yu, Y.: Label-aware double transfer learning for cross-specialty medical named entity recognition. arXiv preprint arXiv:1804.09021 (2018)

[22] Wang, J., Kulkarni, M., Preoţiuc-Pietro, D.: Multi-domain named entity recognition with genre-aware and agnostic inference. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8476–8488 (2020)

[23] Liu, Z., Winata, G.I., Xu, P., Fung, P.: Coach: A coarse-to-fine approach for cross-domain slot filling. arXiv preprint arXiv:2004.11727 (2020)

[24] Sachan, D.S., Xie, P., Sachan, M., Xing, E.P.: Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In: Machine Learning for Healthcare Conference, pp. 383–402 (2018). PMLR

[25] Jia, C., Zhang, Y.: Multi-cell compositional lstm for ner domain adaptation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5906–5917 (2020)

[26] Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., Fung, P.: Crossner: Evaluating cross-domain named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 13452–13460 (2021)

[27] OpenAI: Introducing ChatGPT (2022). https://openai.com/blog/chatgpt Accessed 2023.04.03

[28] Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020)

[29] Liu, J., Pasupat, P., Wang, Y., Cyphers, S., Glass, J.: Query understanding enhanced by hierarchical parsing structures. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 72–77 (2013). IEEE

[30] Liu, J., Pasupat, P., Cyphers, S., Glass, J.: Asgard: A portable architecture for multilingual dialogue systems. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8386–8390 (2013). IEEE

[31] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649 (2018)

[32] Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., Qiu, X.: A unified generative framework for various ner subtasks. arXiv preprint arXiv:2106.01223 (2021)

[33] Wu, Y., Jiang, M., Lei, J., Xu, H.: Named entity recognition in chinese clinical text using deep neural network. Studies in health technology and informatics **216**, 624 (2015)

[34] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)

[35] Tang, M., Zhang, P., He, Y., Xu, Y., Chao, C., Xu, H.: Dosea: A domain-specific entity-aware framework for cross-domain named entity recogition. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 2147–2156 (2022)