electronics

MDPI

Article

Harnessing Causal Structure Alignment for Enhanced Cross-Domain Named Entity Recognition

Xiaoming Liu [1,2,*], Mengyuan Cao [1,3], Guan Yang [1,4], Jie Liu [2,5], Yang Liu [6] and Hang Wang [1,3]

1 School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China; 2021107256@zut.edu.cn (M.C.); guanyang@zut.edu.cn (G.Y.); 2021107253@zut.edu.cn (H.W.)
2 China Language Intelligence Research Center, Beijing 100089, China; liujie@ncut.edu.cn
3 Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou 450007, China
4 Zhengzhou Key Laboratory of Text Processing and Image Understanding, Zhengzhou 450007, China
5 School of Information Science, North China University of Technology, Beijing 100144, China
6 The School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; yangl@xidian.edu.cn
* Correspondence: ming616@zut.edu.cn

Abstract: Cross-domain named entity recognition (NER) is a crucial task in various practical applications, particularly when faced with the challenge of limited data availability in target domains. Existing methodologies primarily depend on feature representation or model parameter sharing mechanisms to enable the transfer of entity recognition capabilities across domains. However, these approaches often ignore the latent causal relationships inherent in invariant features. To address this limitation, we propose a novel framework, the Causal Structure Alignment-based Cross-Domain Named Entity Recognition (CSA-NER) framework, designed to harness the causally invariant features within causal structures to enhance the cross-domain transfer of entity recognition competence. Initially, CSA-NER constructs a causal feature graph utilizing causal discovery to ascertain causal relationships between entities and contextual features across source and target domains. Subsequently, it performs graph structure alignment to extract causal invariant knowledge across domains via the graph optimal transport (GOT) method. Finally, the acquired causal invariant knowledge is refined and utilized through the integration of Gated Attention Units (GAUs). Comprehensive experiments conducted on five English datasets and a specific CD-NER dataset exhibit a notable improvement in the average performance of the CSA-NER model in comparison to existing cross-domain methods. These findings underscore the significance of unearthing and employing latent causal invariant knowledge to effectively augment the entity recognition capabilities in target domains, thereby contributing a robust methodology to the broader realm of cross-domain natural language processing.

Keywords: cross-domain named entity recognition; transfer learning; causal inference; feature interactions; causally invariant knowledge

Citation: Liu, X.; Cao, M.; Yang, G.; Liu, J.; Liu, Y.; Wang, H. Harnessing Causal Structure Alignment for Enhanced Cross-Domain Named Entity Recognition. *Electronics* **2024**, *13*, 67. https://doi.org/10.3390/electronics13010067

Academic Editors: Long Yang and Noura Al Moubayed

Received: 17 October 2023
Revised: 7 December 2023
Accepted: 15 December 2023
Published: 22 December 2023

1. Introduction

Named entity recognition (NER) is a fundamental task in natural language processing (NLP), aimed at identifying entities with specific semantic meanings from text, such as names of people, locations, organizations, and institutions. It plays a significant role in knowledge graphs, information extraction, and text understanding [1–3]. In practical applications, the considerable variance in text genres and terminologies across diverse domains presents a substantial challenge, frequently leading to a scarcity of annotated data within specific target domains. Consequently, the adaptation of named entity recognition (NER) models for cross-domain scenarios, specifically cross-domain named entity recognition (CD-NER), has garnered significant research attention in recent years. This is particularly relevant in resource-constrained environments where the availability of labeled data is limited [4].

The current research on CD-NER has primarily focused on three distinct strategies. First, some researchers [5,6] have explored multi-task joint learning approaches, enhancing cross-domain entity recognition by simultaneously training models on both source and target domains to obtain refined feature representations across tasks. Second, a group of scholars [7,8] have proposed innovative model architectures aimed at understanding the complex semantic dynamics between domains, thus improving cross-domain performance. Third, another set of researchers [9,10] have leveraged pre-trained language models (PLMs) to develop models in data-rich domains, establishing robust source domain models. They have further improved cross-domain performance by transferring feature knowledge from the source domain to the target domain through fine-tuning and domain parameter sharing techniques. A notable example of current state-of-the-art CD-NER models is Cp-NER [10], which utilizes a frozen PLM while employing collaborative domain prefix adjustments to enhance the PLM, obtaining a significant improvement in cross-domain performance, as demonstrated by its superior performance on the CrossNER benchmark. However, it is important to note that existing methodologies often depend on inter-domain generalized knowledge for cross-domain transfer, which may inadvertently introduce out-of-domain knowledge that may not align with the specific task requirements during transfer. This observation underscores the need for a more informed approach to CD-NER, a challenge our proposed Causal Structure Alignment-based Cross-Domain Named Entity Recognition (CSA-NER) model aims to address.

To effectively harness domain-invariant knowledge, our CSA-NER model employs a strategy that extracts causal invariant knowledge between domains. This is achieved by constraining domain-invariant knowledge through causal learning, ultimately enhancing the performance of the target domain. Specifically, Figure 1 illustrates the acquisition of cross-domain causal invariant knowledge from similar syntactic structures in contexts and entities, where an ellipsis in the target domain denotes the omitted text "good way to". This process requires causal inference to learn causal relationships between entities and hidden syntactic structures. Subsequently, causal invariant knowledge hidden in syntactic structures and entities is extracted by aligning similar causal structures using GOT. This approach serves to alleviate the impact of out-of-domain knowledge on the task within the target domain. In various scientific domains, the concept of causal invariance has been extensively explored. For instance, Chevalley [11] designs a unified invariant learning framework that expertly utilizes distribution matching to enrich the acquisition of causal invariant knowledge, leading to a noteworthy enhancement in the model's performance. Chen [12] introduced causally inspired invariant graph learning to discern and leverage causally invariant knowledge pertaining to graph data. This is achieved by constructing causal graphs to represent shifts in the distribution of graphs, enabling the model to concentrate solely on the subgraphs that encapsulate the most pertinent information about the underlying causes of the labels. Furthermore, Arjovsky [13] argued that there is no causal relationship between the spurious correlation resulting from the transfer from the source domain to the target domain and the prediction target, and proposed an invariance risk minimization algorithm to mitigate the model's over-reliance on data bias by using causality tools to characterize the spurious correlation and invariance in the data. Through previous researchers' studies in these scientific domains, this paper finds that extracting domain-invariant knowledge with causal relationships can sufficiently enhance cross-domain migration, thereby constraining the introduction of extraterritorial knowledge that is inconsistent with a given task. The contributions of this paper are summarized as follows:

- This paper proposes a novel method that utilizes causally invariant knowledge between features to improve cross-domain named entity recognition (CD-NER). By leveraging the stability of causally invariant knowledge across domains, this method aids in the effective transfer of knowledge across different data environments.

- The proposed cross-domain named entity recognition model with causal structure alignment incorporates a causal alignment module in the embedding layer to build a causal feature graph by identifying causal relationships between features. Through alignment metrics via graph optimal transport (GOT), it obtains causal invariant knowledge, mitigating the negative transfer effects between domains. Additionally, Gated Attention Units (GAUs) are used in the hidden layer to enhance the utilization of causally invariant knowledge, thereby extracting more efficient feature representations in the target domain
- The proposed method and modeling approach are validated through rigorous experiments conducted on a variety of data sources, including five English datasets and a proprietary cross-domain NER dataset. The experimental results confirm the effectiveness of including causal invariant information within features, demonstrating its significant role in facilitating knowledge transfer for cross-domain named entity recognition.

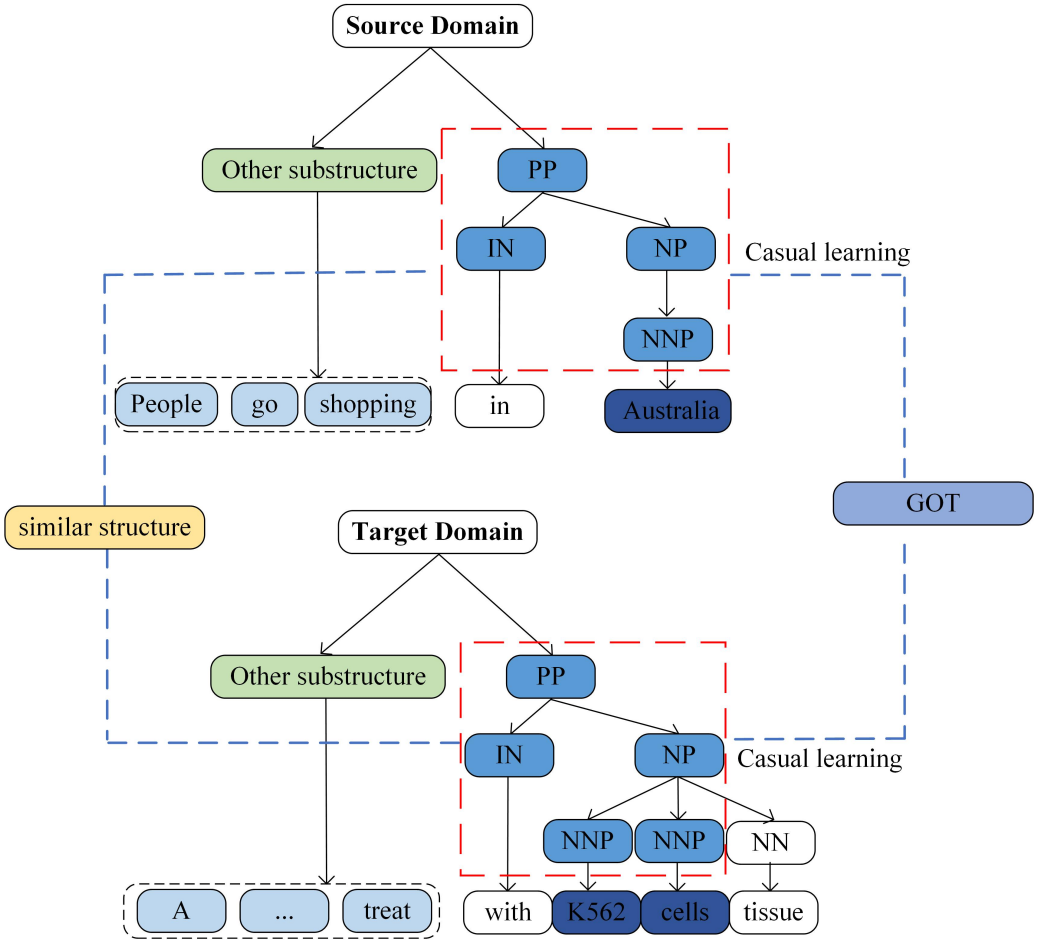

Figure 1. Cross-domain causal similarity structure.

2. Related Work

2.1. Cross-Domain Named Entity Recognition

Cross-domain named entity recognition, which aims to utilize knowledge learned from resource-rich source domains to improve entity recognition in target domains, has received increasing research attention because it can alleviate the problems of data dependency and insufficient training data. Zhang [5] proposed a Multi-Cell Composition LSTM structure that models each entity type as a separate cell state, thus solving the problems of data annotation scarcity and entity ambiguity. These methods need to be trained on a large amount of source domain data to adapt to each domain, making them time consuming

and inefficient. Hu [8] proposed a new auto-regressive modeling framework that exploits semantic relationships between domains to migrate semantic features with the same label in the source domain to the target domain to jointly predict entity labels. Zheng [9] constructed a labeled graph by pre-training a language model and solved the cross-domain label semantic feature mismatch problem by dynamic graph matching. Chen [10] utilized frozen PLMs and conducted collaborative domain-prefix tuning to stimulate the potential of PLMs to handle NER tasks across various domains. In contrast, previous methods based on the transfer of semantic feature knowledge do not solve the negative transfer problem well and thus fail to produce more stable predictions by exploiting the causally invariant knowledge present in the source domain. Therefore, the method in this paper constructs causal feature graphs using causal relationships between features in the source and target domains, and performs graph matching through GOT to learn the causal invariant knowledge in the source domain to mitigate the possible negative effects of using the source task knowledge in the target task. In addition, the model has fewer training parameters, takes less time, and can be combined with different backbone models for better adaptability.

2.2. Few-Shot Named Entity Recognition

Few-shot named entity recognition (FS-NER) aims to identify new classes in resource-poor scenarios and also highlights good cross-domain capabilities. Fritzler [14] used prototype networks to achieve entity recognition for few-shot. Tong [15] proposed mining undefined classes to improve the robustness of the model and thus better adapt to few-shot learning. Cui [16] combined prompted learning templates and BART models for guided entity recognition to improve model performance and cross-domain applications. The authors of [17] do not even need a richly resourced source domain to accomplish small-sample learning without template tuning using prompted learning. The authors of [9] improve domain adaptation in low-resource domains by extracting semantic information of labels in resource-rich source domains. Although the above methods have been significantly improved in small-sample learning, they only improve the model domain adaptation [18] and generalization ability through few-shot training, but do not take into account the fact that the migrated causally invariant knowledge plays a key role in the downstream task. Therefore, the main difference between this paper's method and the above methods is that it is not only applicable to both resource-rich and resource-poor domains, but also more effectively utilizes the causal invariant knowledge to improve the recognition ability of few-shot.

2.3. Causal Invariant Learning

Causal invariant learning is a common solution for domain adaptation and domain generalization in solving cross-domain migration problems, where domain generalization is crucial for learning causal invariant knowledge in the domain. For example, Li [19] introduces a method called Distortion Invariant representation Learning (DIL) to enhance the generalization ability of deep neural networks in image restoration by addressing various types and degrees of image degradation from a causal perspective. Rojas-Carulla [20] proposed a transfer learning method based on causal modeling, which aims to find predictors that lead to invariant conditions through tasks with known underlying causal structure and tasks involving interventions on variables other than the target variable. Yang [21] proposed a causal self-encoder that learns causal representations by integrating them into a unified model using self-encoder and causal structure learning in the source domain, and utilizes this medium causal representation in the target domain for prediction. However, the method lacks the extraction and utilization of causal invariant knowledge. In this paper, we argue that cross-domain transfer takes full advantage of the causal relationships that exist between different features, such as words, utterances, and chapters, to design a causal structural alignment mechanism using causal differences between features and between domains, and to improve cross-domain entity recognition by GOT learning causally invariant knowledge in the source domain in the target domain.

3. Methodology

3.1. Model Architecture

The CSA-NER architecture consists mainly of 2 key modules, as shown in Figure 2. The first module is the causal alignment module, which constructs causal feature maps through causal learning and statistical analysis, and extracts causal invariant knowledge in the causal feature graph using GOT. The second module is the feature interaction module, which captures the correlation information between domains through GAU to further strengthen the causal invariant knowledge learned in the target domain.

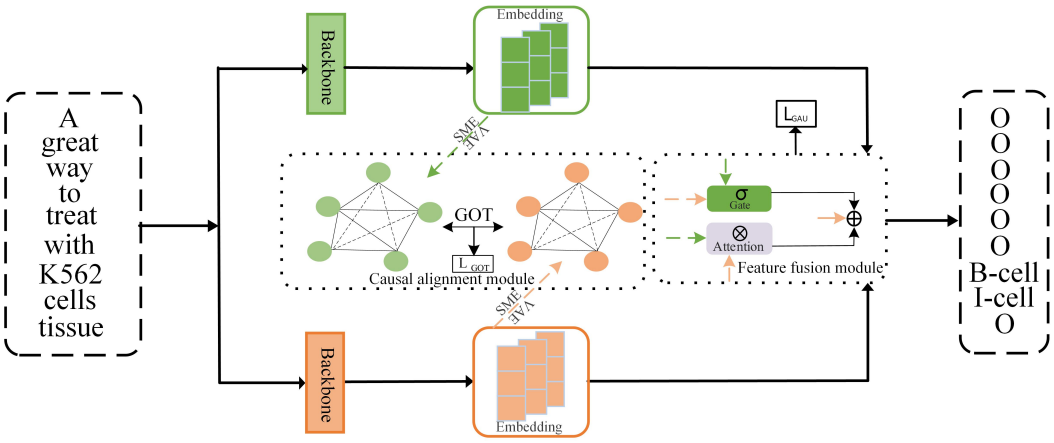

Figure 2. The framework of our CSA-NER model.

3.2. Causal Alignment Module

In order to precisely identify and effectively transfer causal invariant knowledge embedded in feature representations, CSA-NER incorporates a sophisticated causal alignment mechanism. This mechanism serves the purpose of meticulously constructing and aligning causal feature graphs within the embedding layer. The utilization of this module is driven by the imperative to enhance the model to capture and transfer causal relationships, thereby contributing to the overall robustness and performance of the NER system.

3.2.1. Causal Feature Graph

In this section, we introduce the construction of a causal feature graph to elucidate the causal relationships among features. The primary objective behind building the causal feature graph is to discern and leverage the inherent causal connections between entity features and contextual features within a sentence. Conceptually, a causal graph is represented as a directed graph, where nodes corresponding to features and edges signify dependencies between these features. These dependencies can be approximated by the existence of causal relationships [21,22].

The causal feature graph is denoted as $G = \{V, F\}$, $V = \{v_1, v_2, \ldots, v_n\}, V \in \mathbb{R}^{N \times h}$, and $F = \{f_1, f_2, \ldots, f_n\}$, where h represents the feature dimension. Each node $v_i \in V$ signifies a feature representation of the entity and context in the sentence, while each directed edge $f_i \in F$ represents a causal relationship between two nodes. To encode these causal relations within G, we use the adjacency matrix $W^{(f)} \in \mathbb{R}^{N \times N}$. The construction of the causal feature graph is pivotal for understanding and capturing the intricate causal dependencies between features, providing a foundational structure for subsequent stages of the proposed methodology.

The causal structure [23] present in the feature space is represented by the existing causal learning framework in using directed acyclic graphs (DAGs), under the condition of independent homogeneous distribution, given the sequence $X \in R^{N \times h}$, where N is the number of node features. The goal of causal learning is to infer the causal structure of the node feature space, and in particular to extract causal relationships between entities and syntactic structures hidden between nodes. This process requires the construction of a

causal feature graph in order to understand the causal links and thus obtain the acquisition of higher order feature representations between nodes. In this paper, by extending the Structural Equation Model (SEM) [24,25] to causal learning so as to better understand the causal relationships between variables in order to facilitate the recovery of $W^{(f)}$, a DAG structure, from the sequence X, the calculations are as follows:

$$f^{-1}(X) = W^{(f)T} f^{-1}(X) + g(Z) \tag{1}$$

where $Z \in R^{N \times h}$ denotes the corresponding noise matrix, and f and g are parameterized functions of X and Z.

Causal learning based on variational Bayes [26] approximates the true posterior distribution $q_\psi(Z \mid X)$ by minimizing the Kullback–Leibler (KL) dispersion of the variational posterior distribution $p_\theta(Z \mid X)$ to approximate the true posterior distribution. In this context, KL scatter is a measure of the difference between two probability distributions. By minimizing the KL scatter, the variational posterior distribution $p_\theta(Z \mid X)$ can be made as small as possible in terms of the difference between it and the true posterior distribution $q_\psi(Z \mid X)$. This means that we want the variational posterior to be as close as possible to the true posterior in order to better obtain the approximate causality, which is computed as follows:

$$\begin{aligned} &\operatorname{argmin}_{\theta,\phi} D_{KL}\left[q_\phi(Z|X) \| p_\theta(Z|X)\right] \\ &= \arg\min_{\theta,\phi} \int q_\phi(Z|X) log \frac{q_\phi(Z|X)}{p_\theta(Z) p_\theta(X|Z)} dZ \\ &= \operatorname*{argmin}_{\theta,\phi} D_{KL}\left[q_\phi(Z|X) \| p_\theta(Z)\right] - E_{q_\rho(Z|X)}\left[\log p_\theta(X|Z)\right] \end{aligned} \tag{2}$$

where D_{KL} is the KL scatter, and $\theta = (M_X, S_X)$ and $\phi = (M_z, S_z)$ are the generating and variational parameters, respectively.

The cost function obtained in Equation (2) is known as the expected lower bound, and its negative form is known as the variational lower bound or evidence lower bound, and, given a distribution of Z and a set of sequences $X^1, X^2, \ldots, X^N$, the loss can be defined as an average negative lower bound, which can be expressed as:

$$\begin{aligned} L_{ELBO} &= D_{ELBO} - E_{ELBO} \\ D_{ELBO} &= -\frac{1}{n} \sum_{k=1}^{n} D_{KL}\left(q_\phi\left(Z \mid X^k\right) \| p_\theta(Z)\right) \\ E_{ELBO} &= E_{q_\theta(Z|X^k)}\left[\log p_\theta\left(X^k \mid Z\right)\right] \end{aligned} \tag{3}$$

By using the probabilistic encoder and decoder of Bayesian Neural Networks (BNNs), the density functions $q_\phi(Z \mid X^k)$ and $p_\theta(X^k \mid Z)$ can be obtained. Specifically, the uncertainty distribution of the latent variable Z is obtained by mapping the input data X^k into the latent space Z and obtaining the parameters M_Z and S_Z of the variational posterior distribution parameter $q_\phi(Z \mid X^k)$ when the encoder uses the BNNs to instantiate f^{-1} and the constant mapping g. The decoder decodes the latent variable Z into the generated or reconstructed input data X using the inverse functions of f and g, and obtains the parameters M_Z and S_Z of the true posterior distribution $p_\theta(X^k \mid Z)$. The true posterior distribution represents the uncertainty distribution of the generated or reconstructed input data X, given the latent variable Z.

Considering the acyclicity constraints, causal learning [27] is transformed into the following optimization problem by using the augmented Lagrangian method with L_1 regularization, which is computed as follows:

$$\left(W^{(f)}, \Theta\right) = \arg\min_{W^{(0)},\Theta} \left(-L_{ELRO} + L_1\right) \tag{4}$$

$$L_1 = \lambda \left\| W^{(f)} \right\|_1 + \alpha h\left(W^{(f)}\right) + \frac{\rho}{2}\left|h\left(W^{(f)}\right)\right|^2 \tag{5}$$

$$s.t.h\left(W^{(f)}\right) = tr\left[\left(I + \alpha W^{(f)} \circ W^{(f)}\right)^S\right] - S = 0 \tag{6}$$

where tr is the trace of the matrix $\left(I + \alpha W^{(f)} \circ W^{(f)}\right)^S$, θ is a set of parameters of the BNNs in the variational self-coder, α is a Lagrange multiplier, and ρ is a penalty parameter.

Optimization is performed using Equations (4)–(6) to update α and increase ρ, after which a stochastic optimization solver is used to obtain the adjacency matrices $W^{(f)}$ and θ . Therefore, in this paper, by using V to determine the feature representation of a node and $W^{(f)}$ to describe the causal relationship between features, not only can we obtain the node representation V_S and the causal edge representation $W_S^{(f)}$ in the source domain causal graph, but the node representation V_T and causal edge representation $W_T^{(f)}$ in the target domain causal graph can also be obtained.

3.2.2. Causal Structural Alignment

The causal structure alignment in the feature representation aligns not only the node features in the causal graph but also the similar causal relationships between the node features, allowing the model to capture the causal structure information between token features and thus learn causally invariant features that are more representative of the original semantic information.

The model uses GOT [28] as a causal alignment method at the embedding layer to obtain causally invariant knowledge in the feature representation, as shown in Figure 3. Graph optimal transmission targets two optimal transmission distances, a Wasserstein distance (denoted as WD) for node (token) matching and a Gromov-WD (GWD) for edge (causal) matching, using two forms of optimal transfer frameworks to convert cross-domain transfer into causal representations from one domain distribution to another domain distribution. The GWD is a self-normalizing alignment that improves the efficiency of reasoning about interpretable causal relationships with feature information. The model obtains the cross-domain causal similarity matrix $W_{ST}^{(f)}$ by using the causal feature graph to obtain the causality matrix $W_S^{(f)}$ and $W_T^{(f)}$, which is computed as follows:

$$W_{ST}^{(f)} = 1 - W_S^{(f)}\left(W_T^{(f)T}\right)^\top \tag{7}$$

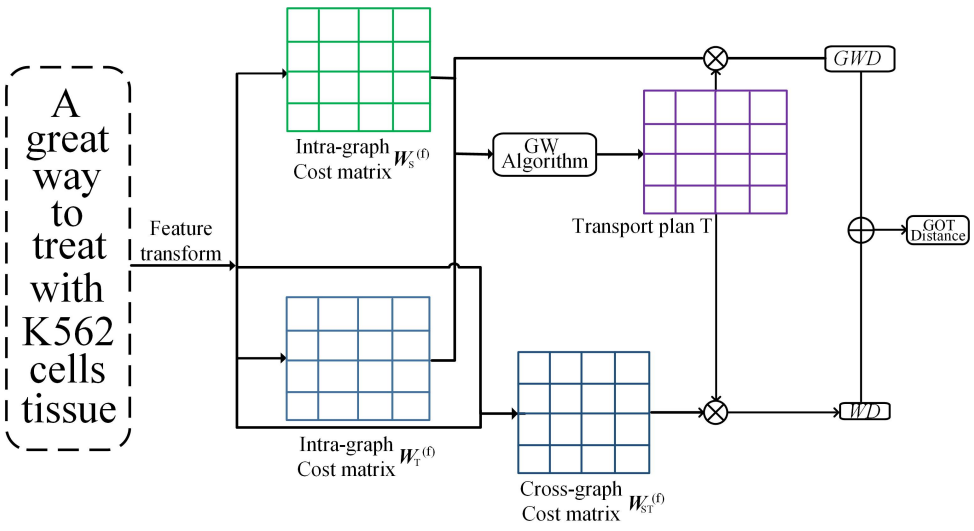

Figure 3. Causal alignment layers.

The distance between the nodes is then measured using WD, which is used for node alignment of semantic features to obtain feature knowledge that is more compatible with the target domain, as calculated below:

$$\begin{aligned} D_{\mathrm{wd}}(\mu,\nu) &= \inf_{\gamma \epsilon \Pi(\mu,\nu)} \mathbb{E}_{(x,y)\succ\gamma}[c(x,y)] \\ &= \min_{T \epsilon \prod(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} \cdot c(x_i, y_j) \end{aligned} \tag{8}$$

where μ and v are two discrete distributions, formulated as $\sum_{i=1}^{n} u_i \delta_{x_i}, \nu = \sum_{j=1}^{m} \nu_j \delta_{y_j}$, and δ_x is the Dirac function on x. $\Pi(\mu,\nu)$ denotes all the joint distributions $\gamma(x,y)$ with marginal distributions $\mu(x)$ and $\nu(y)$; μ and ν represent the weight vectors, respectively; x and y denote the semantic features of the incoming source and target domains, respectively; $\boldsymbol{x} = \boldsymbol{V}_S$ and $\boldsymbol{y} = \boldsymbol{V}_T$ denote the semantic features of the incoming source and target domains, respectively. $c(x_i, y_j)$ is the cost function for evaluating $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$, and the cosine function is chosen in this case.

GWD is used for edge-structure alignment by measuring the distance between edges in the graph. For GWD the specific formula is as follows:

$$\begin{aligned} D_{\mathrm{gwd}}(\boldsymbol{\mu},\boldsymbol{\nu}) &= \inf_{\gamma \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma,(\boldsymbol{x}',\boldsymbol{y}')\sim\gamma} \left[L\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{x}',\boldsymbol{y}'\right) \right] \\ &= \min_{\hat{\boldsymbol{r}} \in \Pi(\mu,\boldsymbol{v})} \sum_{i,i',j,j'} \widehat{\boldsymbol{T}}_{ij} \widehat{\boldsymbol{T}}_{i'} L\left(\boldsymbol{x}_i, \boldsymbol{y}_j, \boldsymbol{x}_i', \boldsymbol{y}_j'\right), \end{aligned} \tag{9}$$

where (x_i, x_i') and $(y_j, y_j')'$ represent the edge structure between different nodes in the source and target domains, respectively, $L(\cdot)$ is the causal similarity depletion function that evaluates the causal similarity of node pairs between different domains for (x_i, x_i') and $(y_j, y_j')'$, for example, $L(x_i, y_j, x_i', y_j') = \|w_1(x_i, x_i') - w_2(y_i, y_i')\|$, c_1 and c_2 are the causal correlations between evaluation nodes in the same graph, and the matrices are obtained by learning to align edges in different graphs to learn the causal invariant knowledge present in the source domain. In summary, the two distances computed by graph-optimized transmission are used as a loss function for causal structure alignment to learn causal invariant knowledge in causal feature graphs. The loss function is computed as follows:

$$L_{GOT} = D_{gw} + D_{gwd} \tag{10}$$

3.3. Feature Fusion Module

To facilitate the fusion of source and target domains for the enhancement of CD-NER capabilities, this paper presents the design of a feature fusion module, which comprises an attention unit and a gating unit. The attention unit is responsible for focusing on the most pertinent information, thereby augmenting the utilization of causal invariant knowledge. In contrast, the gating unit compensates for the neglect of other relevant knowledge within the source domain. Through the integration of the attention unit and the gating unit, a more comprehensive and precise fusion feature representation is obtained. Figure 4 shows how attention units and gate units collaborate to enhance the utilization of causal invariance knowledge.

The attention unit focuses on the extent to which interactions between different locations and utilization of causally invariant knowledge between domains enhance the utilization of causally invariant knowledge in the target domain. As in the attention unit in Figure 4, causal invariant knowledge is learnt through GOT to output the relevant adaptation features, which are input to the unit, where the query matrix is the adaptive

feature V_T of the target domain, and the key matrix and the value matrix are the adaptive feature V_S of the source domain; the attention unit can be formalized as follows:

$$A_e = MultiAttention(\boldsymbol{V}_T\boldsymbol{W}_i^Q, \boldsymbol{V}_S\boldsymbol{W}_i^K, \boldsymbol{V}_S\boldsymbol{W}_i^V) \tag{11}$$

$W_i^Q \in R^{h\times d_k}, W_i^K \in R^{h\times d_k}, W_i^V \in R^{h\times d_k}$ denotes the training projection parameter $d_k = h/n$, n is the number of heads of attention, where A_e denotes the features generated by the target domain through the interaction of the attention unit with the source domain. Although the attentional mechanism can continue to learn causally invariant knowledge in the target domain that is not learned in the source domain, some useful information may be missed by low attentional weights or not fully captured due to inherent limitations in the allocation of attention. Interaction gating units are therefore introduced to compensate for the lack of attention to weakly relevant causally invariant features by the attentional mechanism and to improve the model's ability to model inter-domain interactions.

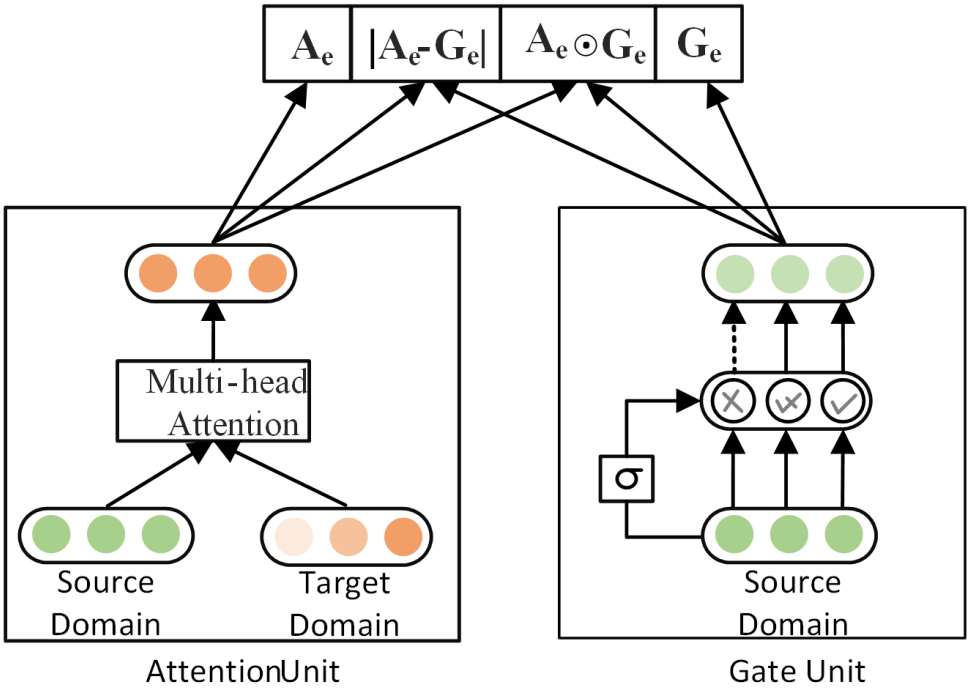

Figure 4. Domain collaboration.

As in the gate unit in Figure 4, the gating unit performs finer-grained modeling in terms of local structure and change by adjusting the input stream point by point, selectively emphasizing or attenuating the contribution of particular locations or features, filtering out confounding information in the source domain, and retaining as much causally invariant knowledge as possible. Adaptive features from the source domain are fed into interaction gates in the target domain to generate interaction features for the task.

$$\begin{aligned} g_e &= \sigma(H^{'}W_g + b_g) \\ G_e &= g_e \odot V_s \end{aligned} \tag{12}$$

where $g_e \in R^{N\times h}$ denotes the interaction gate for entity recognition in the target domain, $G_e \in R^{N\times h}$ denotes that the target domain learns the relevant features from the source domain through the gating unit, and $W_g \in R^{l\times h}$ and $b_s \in \mathbb{R}^h$ denote the training weights and biases. After the attention unit and the interaction gating unit in the feature fusion module have acted, each unit generates feature representations A_e and G_e from different perspectives that are beneficial to the task of entity recognition, and then the two features

are integrated using absolute difference and element-wise product [29] to obtain interactive feature representations for entity recognition in the target domain. Finally, the task adaptation features and interaction features are connected to obtain the task feature representation of the whole feature fusion module:

$$I'_e = MLP([A_e; |A_e - G_e|; A_e \odot G_e; G_e]) \tag{13}$$

where $MLP(\bullet)$ denotes the multilayer perceptron and $I'_e \in \mathbb{R}^{N \times h}$ denotes the interaction features of the target domain task. As shown in Figure 4, this module integrates the specificity and sharing between domain features, and enhances the causally invariant knowledge concern learnt from GOT as much as possible through the interactions between domains, so as to generate richer feature representations.

3.4. Optimization Goals

The training process uses variational Bayes in the embedding layer to calculate the variational loss L_{ELBO} to optimize the causal relationships existing between the domains, after which the alignment loss L_{GOT} is completed by graph-optimal transmission, and finally the cross-entropy loss function (CE) is introduced in the coding layer to calculate the loss L_{GAU} of the GAU, which is computed as follows:

$$L_{GAU} = CE(I'_e W_G, Y^T) \tag{14}$$

where $W_G \in R^{h \times c}$, where c denotes the number of categories of labels and $\mathbf{Y}^T$ represents the true labels in the target domain.

In summary, the overall loss function of CSA-NER can be expressed as the following:

$$L = L_{GAU} + \lambda_1 * L_{ELBO} + \lambda_2 * L_{GOT} \tag{15}$$

4. Experiments

4.1. Datasets and Settings

The source domain dataset is ConLL-2003 [30], which is a generic dataset containing the common names of persons (PER), locations (LOC), and organizations (ORG). The cross-domain datasets are CrossNER [4], BioNLP13PC (PC), and BioNLP13CG (CG) [31]. CrossNER is a specialized cross-domain NER dataset containing five domains, namely Politics (Pol.), Natural Sciences (Sci.), Music (Mus.), Literature (Lit.), and Artificial Intelligence (AI), and each domain contains not only the same entity types as ConLL-03 but also specific entity types. For example, the political field contains political parties, politicians, and elections; the science field contains scientists, disciplines, and chemical compounds; the music field contains musical genres, musical instruments, songs, and so on; the Lit. field contains authors, poems, and journals; and the AI field contains algorithms, researchers, and so on. The PC and CG datasets belong to the medical and biological domains, respectively, and the entity types mainly including simple chemical (CHEM), cellular component (CC), gene and gene product (GGP), species (SPE), and cell (CELL) are also included in BioNLP13CG. Specific dataset statistics are shown in Table 1. According to the difference in the distribution of the datasets in the relevant domains, they can be divided into two major groups of experiments. In the first group, ConLL-2003 is chosen as the source domain dataset for CD-NER experiments, and PC, CG, and Cross-NER are the target domain datasets. In the second group, given the above English dataset to perform the cross-domain transfer few-shot experiments, ConLL-2003 is chosen as the source domain dataset and Cross-NER as the target domain dataset.

Table 1. Target domain datasets.

Domain	Size	Train	Dev	Test
PC	Sentence	2.5 K	0.6 K	1.7 K
	Entity	7.9 K	1.2 K	5.3 K
CG	Sentence	3.0 K	0.8 K	1.9 K
	Entity	10.8 K	1.6 K	6.9 K
Pol.	Sentence	0.2 K	0.5 k	0.6 k
	Entity	1.3 K	3.4 k	4.2 k
Sci.	Sentence	0.2 K	0.4 k	0.5 k
	Entity	1.0 K	2.5 k	3.0 k
Mus.	Sentence	0.1 K	0.3 k	0.4 k
	Entity	0.6 K	2.6 k	3.3 k
Lit.	Sentence	0.1 K	0.4 k	0.4 k
	Entity	0.5 K	2.1 k	2.2 k
AI	Sentence	0.1 K	0.3 k	0.4 k
	Entity	0.5 K	1.5 k	1.8 k

Table 2 shows the parameter settings for the experiment. The BERT-based model [32] is selected as the backbone model for the experiment, where the model is placed in the Pytorch [33] framework to complete the experiment and the version number of this pytorch is 1.8.0. The experimental model parameters are set as follows: SGD is selected as the optimizer, the learning rate is set to 1×10^{-4}, the batch size is 8, the epoch is 50, the hidden variable is 768, and dropout is set to 0.5 to prevent over-fitting.

Table 2. Hyperparameters.

Parameter	Value
Hidden variable	768
Batch size	8
Epoch	50
Dropout	0.5
Learning rate	1×10^{-4}
Optimizer	SGD

4.2. Evaluation Protocols

In the present study, we employ a metric that aligns with and is analogous to prior research endeavors. This metric evaluates the precision of predictions in terms of the accurate identification of both the entity's category and its boundaries. The key evaluative factors utilized for the computation of the ultimate score are the accuracy (Precision), the recall (Recall), and associated values (F_1). This computation follows a formulaic derivation:

$$P = \frac{TP}{TP + FP} \tag{16}$$

$$R = \frac{TP}{TP + FN} \tag{17}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{18}$$

where TP represents the number of correctly identified entities, FP represents the number of incorrectly identified entities, and FN represents the number of unidentified entities.

4.3. Baseline Models

To verify the effectiveness of this model on cross-domain NER, comparative experiments are conducted on different datasets with related models.

Coach: [7] a framework of domain adaptation, which divides the task into two stages, first detecting entities and then classifying them to solve the problem of data.

BERT-tag: [32] proposed the BERT-based baseline model which fine-tunes the BERT model with a label classifier.

LSTM: [5] proposed a multicellular LSTM structure based on Bert representation to model diverse entity types separately and perform cross-domain knowledge transfer at the entity level to solve the problem that entities have different meanings in different domains.

Tem-NER: [16] proposed a template-based approach to named entity recognition, which reduces the need for labeled data by embedding predefined templates into the pre-trained BERT model, guiding the model to generate entity labels more accurately.

LST-NER: [9] transformed the cross-domain problem into a graph matching problem to alleviate the problems of entity type mismatch and domain knowledge not being be effectively transferred to the target domain.

Cp-NER: [10] proposed a NER task that can simultaneously handle multiple domains in a single model by adding vocabulary and specific terms from different domains as prefixes to the model parameters and automatically adjusting these prefixes through collaborative training and optimization to improve cross-domain NER performance.

4.4. Result Analysis

The results of the CD-NER experiments on the CrossNER dataset, the PC dataset and the CG dataset were compared with other related methods as shown in Table 3, and the results of the FS-NER experiments on the CrossNER dataset were compared with other relevant methods as shown in Table 4. Overall the CSA-NER proposed in this paper achieves good results in different datasets.

Table 3. Experiment results table of resource-abundant English dataset (%), Bold indicates best results in this domain. * Indicates results reproduced in this domain.

Method	Pol.	Sci.	Mus.	Lit.	AI	PC	CG
Coach	61.50	52.09	51.66	48.25	45.15	-	-
BERT-tag	68.71	64.97	68.30	63.63	58.88	-	-
LSTM	70.56	66.42	70.52	66.96	58.28	86.26	80.74
LST-NER	70.44	66.83	72.08	67.12	60.32	87.14 *	82.48 *
Cp-ner	73.41	**74.65**	78.08	**70.84**	64.53	88.48 *	84.53 *
Ours	**73.58**	72.12	**78.53**	69.42	**64.58**	**88.82**	**85.45**

Table 4. Experiment results table of resource-scarce English dataset (%), Bold indicates best results in this domain.

Sample	k = 20					k = 50				
Method	**Pol.**	**Sci.**	**Mus.**	**Lit.**	**AI**	**Pol.**	**Sci.**	**Mus.**	**Lit.**	**AI**
Coach	46.15	48.71	43.37	41.64	41.55	60.97	52.03	51.56	48.73	51.15
Bert-tag	61.01	60.34	64.73	61.79	53.78	66.13	63.93	68.41	63.44	58.93
LSTM	59.58	60.55	67.12	63.92	55.39	68.21	65.78	70.47	66.85	58.67
Tem-NER	63.39	62.64	62.00	61.84	56.34	65.23	62.84	64.57	64.49	56.58
LST-NER	64.06	64.03	68.83	64.94	57.78	68.51	66.48	72.04	66.73	60.69
Ours	**65.14**	**66.27**	**70.12**	**65.83**	**58.15**	**69.32**	**68.21**	**73.26**	**67.32**	**61.02**

The full samples of the PC, CG, and CrossNER datasets are given for CD-NER experimental validation, and the results are shown in Table 1. LST-NER improves the median values of five different domains in CrossNER compared to LSTM. LSTM utilizes a multi-task learning approach to improve the cross-domain entity recognition performance in the

target domain by training both the source and target domains simultaneously, and then capturing the feature knowledge that is useful for the target domain, but CD-NER is single-task, using MULTI-CELL-LSTM is not able to better mine the causal invariant knowledge existing in the cross-domain, and the multi-task training cycle is long. LST-NER builds the network with a specific training architecture based on a single task and, after the source domain model is trained, the model migrates the causally invariant knowledge to the target domain by fine-tuning the source domain model and combines it with the target domain features, and the training period is short. Although the values of Cp-NER relative to LST-NER are also improved in different areas, the methodology of this paper mainly uses the same base framework as LST-NER; this is because Cp-NER mainly utilizes the base framework of the large model; therefore, choosing the same base framework as LST-NER is to highlight that the improvement of the effectiveness of the model in this paper is not on the basis of the framework but due to the validity of the proposed methodology.

The results show that our model consistently outperforms all the compared models in both low- and rich-resource settings. It is well illustrated that the model achieves better performance in both resource-rich and resource-poor environments by constructing causal feature graphs to establish causal relationships in token features, and then further constraining the causal feature graphs using graph matching to further learn the causal invariant knowledge present in the feature information.

The experimental results also show that there is no significant improvement in model effectiveness when comparing our model with Cp-NER and there is a difference in the selection of PLMs. We believe that the reason for this may be that Cp-NER chooses the T5 model with stronger learning capability and uses an external domain-related corpus for model pre-training, while the latter chooses the basic BERT-base model. In addition, CpNER adopts a multi-source domain learning framework, while CSA-NER is unfair to CpNER in terms of domain adaptation capability. However, CpNER does not explore the feature knowledge representation in depth and does not make full use of the causal invariant knowledge in the features. Meanwhile, CpNER fails to mitigate the negative migration problem caused by migration. In contrast, CSA-NER makes full use of such causal invariant knowledge to migrate the causally related invariant knowledge (e.g., knowledge of grammatical structures) from the source domain to the target domain, which greatly reduces the migration of spurious relational features. Therefore, even in the case of a relatively weak base model, CSA-NER has a certain performance improvement compared to Cp-NER.

4.5. Ablation Study

To rigorously evaluate the robustness and effectiveness of our model's causally invariant knowledge, we conducted an ablation study under two distinct resource settings. For the low-resource scenario, we selected the Politics and Music datasets with a modest sample size of K = 50. Conversely, in the high-resource context, we employed the more extensive PC and CG datasets, utilizing their full sample populations.

1. Removal of the causal graph construction task loss L_{ELBO}.
2. Omission of L_{GOT}, which is crucial for the graph matching task.
3. Absence of both L_{ELBO} and L_{GOT}, to assess their combined effect.
4. Exclusion of the gate mechanism in L_{GAU}.
5. Simultaneous elimination of L_{ELBO}, L_{GOT}, and the gate mechanism in L_{GAU}.

Each above setting was methodically analyzed to determine its contribution to the overall performance, offering insights into the significance of each component in achieving causally invariant knowledge under different resource constraints.

As a result of Table 5 results show that both the causal alignment module and the feature interaction module are beneficial for learning better NER models. Combining causal graph construction and graph matching can yield good results when the model utilises the causal structure alignment module. In combination with learning causally invariant knowledge in graph structures (i.e., source graphs), causally invariant knowledge becomes more effective

when attention is locked onto the structure. Furthermore, when eliminating the gating units in the causal alignment module and the feature interaction module, and utilising only Bert-base as the primary model, the approach in this paper is able to significantly improve upon Bert-base, suggesting that transferring causal invariant knowledge in the source domain is crucial and beneficial for cross-domain NERs.

Table 5. Ablation study on low and high resources (%).

T	Low Resources		High Resources	
Datasets	Politics	Music	PC	CG
Ours	69.32	73.26	88.82	85.45
w/o L_{ELBO}	68.80	72.13	87.82	83.95
w/o L_{GOT}	68.65	71.10	86.55	83.68
w/o *gate*	68.12	72.56	86.84	84.77
w/o L_{ELBO} + L_{GOT}	67.78	70.89	85.16	83.13
w/o L_{ELBO} + L_{GOT} + *gate*	68.12	69.13	85.02	82.21

4.6. Performance with Different Data

We evaluate the performance of our model with different amounts of target domain labeled data on the CG domain and make comparisons with baselines LST-NER,the specific results are shown in Figure 5. We use the same few-shot sampling strategy as in the low-resource setting. We find that even when in a highly low-resource scenario (K = 5, 10), the proposed model shows competitive performance with the few-shot NER model LST-NER.When more data are available,our model consistently outperforms LST-NER.In contrast, the performance of the LST-NER model flattens out when there is relatively enough data. We suggest that the reason for this may be that while LST-NER improves domain adaptation, it lacks the emphasis on causal invariant knowledge. The results suggest that the method in this paper improves the model's emphasis on causal invariant knowledge compared to the small-sample approach.

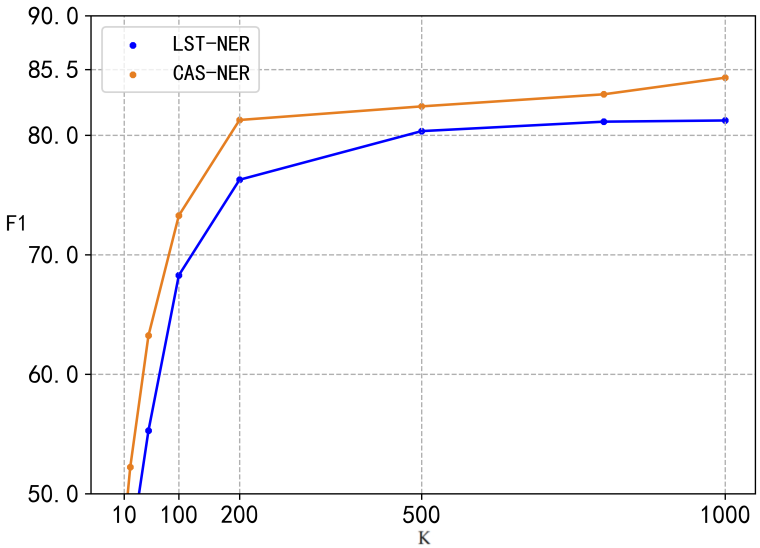

Figure 5. Comparisons when utilizing different amounts of data for training in CG domain.

4.7. Hyperparameter Discussion

In our study, we conducted a detailed exploration of the impact of two weight parameters, λ_1 and λ_2, on the model's performance. Both parameters are integral in modulating different aspects of the learning process:

- λ_1 is designed to control the uncertainty distribution within the causal feature graph. This is achieved by adjusting L_{ELBO}, the loss associated with the causal graph construction task. By tuning λ_1, we effectively manage how the model accounts for uncertainty in the causal relationships it identifies and represents.
- λ_2, on the other hand, is pivotal in aligning causally invariant knowledge within the causal features. It does this by adjusting L_{GOT}, which is instrumental in the graph matching task, ensuring that the causally relevant features are accurately aligned across different domains.

The experiment's results, as depicted in Figure 6, provide insightful observations regarding these parameters, especially in the context of political data. When we analyzed the effect of these parameters on the F1 score, a metric commonly used to evaluate the accuracy of a model, we found that these parameters, surprisingly, did not play a major role in enhancing the model's performance. This outcome is particularly significant as it reinforces the validity of the method proposed in our paper.

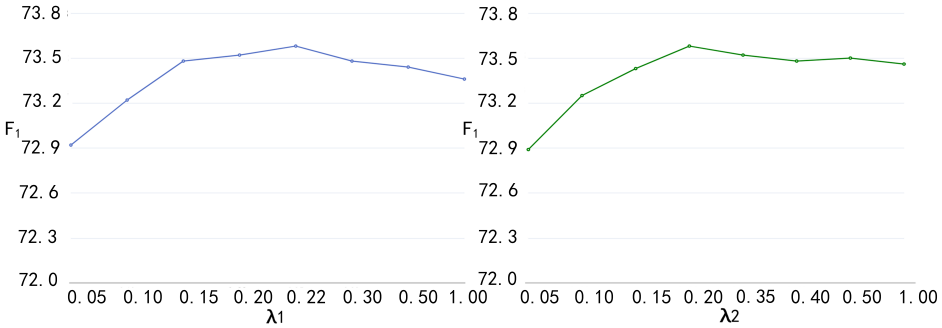

Figure 6. The impact of weight parameters λ_1 and λ_2 on the performance in the politics domain.

4.8. Fine-Grained Analysis

In order to further analyze the prediction effect of CAS-NER on different entity types, a fine-grained analysis was performed on the PC dataset for the CD-NER task, and the main entity types in the PC dataset were CCP, GCP, CHEM, and Complex; Figure 7 shows the prediction results in the CD-NER task. In the entity identification task, CAS-NER is more effective than CP-NER in almost all entity types, although the results are similar to those of Cp-NER in the CCP type, but this better highlights the effectiveness of this paper's method because the data samples of the CCP type are small, only 6. 1% of the total number, but this also better illustrates that this paper's method is also comparable to the large model-based Cp-NER approach, and that this paper is able to extract causal invariant knowledge hidden in entities and contexts for cross-domain migration regardless of the amount of data.

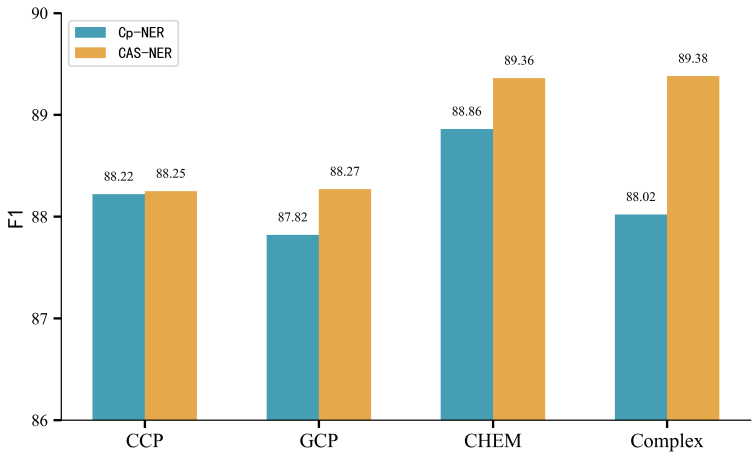

Figure 7. Comparison of F_1 values for different entity types.

4.9. Case Study

As shown in Figure 8, the CSA-NER method enhances the performance of prediction in the target domain by extracting the causal invariant knowledge embedded in the features. When compared with the LST-NER model in a low-resource environment, the method is able to more accurately mark entities in text as correct, and the target domain improves the generalization ability of the model by learning causal invariant knowledge from the source domain and better enriching the causal invariant knowledge by combining with the attention gating mechanism. For Cp-NER, the semantic features of multiple source domains are used and the source domains are trained under a large language model, which is more of a generalization of multiple source domains, whereas the model in this paper is trained in a single source domain and mainly extracts causally invariant knowledge to enhance the effect of entity recognition, so the model of this paper can achieve a similar effect to that of Cp-NER in comparison with the model of Cp-NER. For example, for Joseph, our method can also accurately label him as a researcher instead of a person. In summary, CSA-NER shows good performance improvement in cross-domain named entity recognition.

Input Sentence:	Treatment of K562 cells with is a good way
LST-NER:	Treatment of K562(O) cells(O) with is a good way
CSA-NER :	Treatment of K562(Cell) cells(Cell) with is a good way
Input Sentence:	Woman 1984, Cary Joji Fukunaga's No Time to Die and Joseph Kosinski's Top Gun: Maverick
LST-NER:	Woman 1984, Cary Joji Fukunaga's No Time to Die and Joseph(People) Kosinski's Top Gun: Maverick
Cp-NER:	Woman 1984, Cary Joji Fukunaga's No Time to Die and Joseph(Researcher) Kosinski's Top Gun: Maverick
CSA-NER:	Woman 1984, Cary Joji Fukunaga's No Time to Die and Joseph (Researcher) Kosinski's Top Gun: Maverick

Figure 8. Prediction comparison of different models in different fields.

5. Conclusions

In this paper, we propose a cross-domain named entity recognition model based on causal structure alignment (CAS-NER) by investigating the enhanced capability of causal invariant knowledge in cross-domain named entity recognition. By aligning similar causal structures, the model effectively improves the entity recognition ability in the target domain and achieves better cross-domain knowledge transfer. Experimental results show that CAS-NER performs better than current cross-domain approaches, which further demonstrates that the use of causal invariant knowledge can also facilitate cross-domain knowledge transfer. In the future, we will further optimize the model and investigate the effect of the causal learning mechanism on the target domain migration mechanism under multiple source domains based on other cross-domain sequence annotation tasks.

Author Contributions: Responsible for literature research, research methodology, experimental design, thesis writing, and full text revision: M.C.; Responsible for proposing research ideas, modeling frameworks, content planning, guidance, and full-text revisions: X.L.; Responsible for lab instruction, guidelines, and full text revisions: G.Y. and H.W.; Responsible for providing guidance, revising, and reviewing full texts: Y.L. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Key Research and Development Program of China (2020AAA0109703); National Natural Science Foundation of China (No.62376207 and No.62076167); the Open Research Fund from the Guangdong Provincial Key Laboratory of Big Data Computing,

The Chinese University of Hong Kong, Shenzhen, under Grant No.B10120210117-OF06; Guangxi Key Laboratory of Machine Vision and Intelligent Control (No. 2022B10); State Key Lab. for Novel Software Technology, Nanjing University (No. KFKT2022B25).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Ehrmann, M.; Hamdi, A.; Pontes, E.L. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–47. [CrossRef]
2. Ahmad, P.N.; Shah, A.M.; Lee, K. A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain. *Healthcare* **2023**, *11*, 1268. [CrossRef]
3. Tsai, C.-M. Stylometric Fake News Detection Based on Natural Language Processing Using Named Entity Recognition: In-Domain and Cross-Domain Analysis. *Electronics* **2023**, *12*, 3676. [CrossRef]
4. Liu, Z.; Xu, Y.; Yu, T. Crossner: Evaluating cross-domain named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 13452–13460.
5. Chen, J.; Zhang, Y. Multi-cell compositional LSTM for NER domain adaptation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5906–5917.
6. Tan, Z.; Chen, Y.; Liang, Z. Named Entity Recognition for Few-Shot Power Dispatch Based on Multi-Task. *Electronics* **2023**, *12*, 3476. [CrossRef]
7. Liu, Z.; Winata, G.I.; Xu, P. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 19–25.
8. Hu, J.; Zhao, H.; Guo, D.; Wan, X.; Chang, T. A label-aware autoregressive framework for cross-domain NER. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 July 2022; pp. 2222–2232.
9. Zheng, J.; Chen, H.; Ma, Q. Cross-domain named entity recognition via graph matching. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 2670–2680.
10. Chen, X.; Li, L.; Fei, Q.; Zhang, N.; Tan, C.; Jiang, Y.; Chen, H. One Model for All Domains: Collaborative Domain-Prefix Tuning for Cross-Domain NER. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (JCAL-23), Macao, China, 19–25 August 2023; Volume 2301, p. 10410.
11. Chevalley, M.; Bunne, C.; Krause, A.; Bauer, S. Invariant causal mechanisms through distribution matching. *arXiv* **2022**, arXiv:2206.11646.
12. Chen, Y.; Zhang, Y.; Bian, Y.; Yang, H.; Ma, K.; Xie, B.; Liu, T.; Han, B.; Cheng, J. Learning causally invariant representations for out-of-distribution generalization on graphs. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22131–22148.
13. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant Risk Minimization. *arXiv* **2019**, arXiv:1907.02893.
14. Fritzler, A.; Logacheva, V.; Kretov, M. Few-shot classification in named entity recognition task. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; pp. 993–1000.
15. Tong, M.; Wang, S.; Xu, B.; Cao, Y.; Liu, M.; Hou, L.; Li, J. *Learning from Miscellaneous Other-Class Words for Few-Shot Named Entity Recognition*; Association for Computational Linguistics (ACL): Cedarville, OH, USA, 2021; pp. 6236–6247.
16. Cui, L.; Wu, Y.; Liu, J.; Yang, S.; Zhang, Y. Template-Based Named Entity Recognition Using BART. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 1835–1845.
17. Ma, R.; Zhou, X.; Gui, T.; Tan, Y.; Li, L.; Zhang, Q. Template-free Prompt Tuning for Few-shot NER. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 10–15 July 2022; Volume 2109, p. 13532.
18. Lu, W.; Wang, J.; Li, H.; Chen, Y.; Xie, X. Domain-invariant Feature Exploration for Domain Generalization. *Trans. Mach. Learn. Res.* **2022**, 2835 –8856.
19. Li, X.; Li, B.; Jin, X.; Lan, C.; Chen, Z. Learning Distortion Invariant Representation for Image Restoration from A Causality Perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1714–1724.
20. Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; Peters, J. Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **2018**, *19*, 1309–1342.
21. Yang, S.; Yu, K.; Cao, F.; Liu, L.; Wang, H.; Li, J. Learning causal representations for robust domain adaptation. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 2750–2764. [CrossRef]
22. Kocaoglu, M.; Snyder, C.; Dimakis, A.G.; Vishwanath, S. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. *Int. Conf. Learn. Represent.* **2018**, *1709*, 02023.
23. Wei, D.; Gao, T.; Yu, Y. DAGs with No Fears: A closer look at continuous optimization for learning Bayesian networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3895–3906.
24. Zheng, X.; Aragam, B.; Ravikumar, P.K.; Xing, E.P. Dags with no tears: Continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9472–9483.

25. Zhai, P.; Yang, Y.; Zhang, C. Causality-based CTR prediction using graph neural networks. *Inf. Process. Manag.* **2023**, *60*, 103137 [CrossRef]
26. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114.
27. Ng, I.; Zhu, S.; Chen, Z.; Fang, Z. A Graph Autoencoder Approach to Causal Structure Learning. *arXiv* **2019**, arXiv:1911.07420.
28. Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; Liu, J. Graph optimal transport for cross-domain alignment. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1542–1553.
29. Van Lint, J.H.; Wilson, R.M. *A Course in Combinatorics*; Cambridge University Press: Cambridge, UK, 2001.
30. Mou, L.; Men, R.; Li, G.; Xu, Y.; Zhang, L.; Yan, R.; Jin, Z. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 130–136.
31. Nédellec, C.; Bossy, R.; Kim, J.D.; Kim, J.J.; Ohta, T.; Pyysalo, S.; Zweigenbaum, P. Overview of BioNLP shared task 2013. In Proceedings of the BioNLP Shared Task 2013 Workshop, Sophia, Bulgaria, 9 August 2013; pp. 1–7.
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
33. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.P.; Chanan, G.; Chintala, S.P. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.