Context-aware Information-theoretic Causal De-biasing for Interactive Sequence Labeling

Junda Wu[1*] **Rui Wang**[2*] **Tong Yu**[3†] **Ruiyi Zhang**[3] **Handong Zhao**[3]
Shuai Li[4] **Ricardo Henao**[2] **Ani Nenkova**[3]
[1]New York University [2]Duke University [3]Adobe Research [4]Shanghai Jiao Tong University
jw6466@nyu.edu
{rui.wang16,ricardo.henao}@duke.edu
{tyu,ruizhang,hazhao,nenkova}@adobe.com
shuaili8@sjtu.edu.cn

Abstract

Supervised training of existing deep learning models for sequence labeling relies on large scale labeled datasets. Such datasets are generally created with crowd-source labeling. However, crowd-source labeling for tasks of sequence labeling can be expensive and time-consuming. Further, crowd-source labeling by external annotators may not be appropriate for data that contains user private information. Considering the above limitations of crowd-source labeling, we study interactive sequence labeling that allows training directly with the user feedback, which alleviates the annotation cost and maintains the user privacy. We identify two biases, namely, *context bias* and *feedback bias*, by formulating interactive sequence labeling via a Structural Causal Model (SCM). To alleviate the context and feedback bias based on the SCM, we identify the frequent context tokens as confounders in the backdoor adjustment and further propose an entropy-based modulation that is inspired by information theory. With extensive experiments, we validate that our approach can effectively alleviate the biases and our models can be efficiently learnt with the user feedback.

1 Introduction

Recently, deep learning models have yielded state-of-the-art performance for tasks of sequence labeling, such as POS tagging (Doostmohammadi et al., 2020; Nguyen et al., 2021a) and Named Entity Recognition (NER) (Devlin et al., 2018; Lample et al., 2016; Xu et al., 2021). Unfortunately, existing deep learning models are well-known of data-hungry (Guo et al., 2020; Hathurusinghe et al., 2021; Nguyen et al., 2021b), relying on large annotated datasets for training. These datasets are generally created with crowd-source labeling, *e.g.*, using Amazon Mechanical Turk.

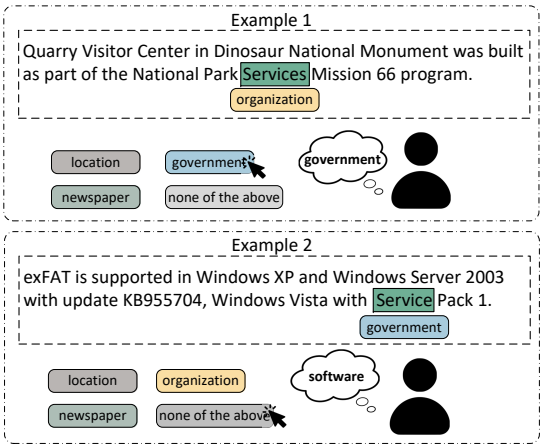

Figure 1: Example use cases in NER. In a modern NER system, there may be hundreds of entity classes. In example 1, the user gets the top 1 prediction "organization" and a list of other predicted top candidate entities and provides the feedback that "government" is correct. In example 2, the user cannot find the correct entity within the top candidate entities, and chooses "none of the above".

Crowd-source labeling can be expensive and time-consuming, especially for sequence labeling tasks. This is because training for sequence labeling generally requires token-level annotations, *i.e.*, assigning labels to each token/word within a given text sequence. Further, it may not be feasible to adopt crowd-source labeling for privacy-sensitive domains, involving private data of the users. For instance, in commercial scenarios, the user-generated text might contain Personal Identifiable Information (PII), *e.g.*, person names or home addresses. Such user data may be prohibitive to be collected or saved, not along distributed for crowd-sourcing. Therefore, we consider an alternation of crowd-sourcing, *i.e.*, *interactive sequence labeling*, where the sequence labeling model is directly trained with feedback from the users, in lieu of crowd-sourced annotators. Specifically, the sequence labeling model is trained with batches of streaming data

[*]Equal Contribution
[†]Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3436–3448
December 7-11, 2022

from the users. Each time when the model receives a batch of user data, the users will receive the top predicted token classes together with some other possible classes, and provide feedback for model training, as in Figure 1. To reduce feedback effort of users (and also sometimes with limited space in UI), each time the model only displays its top K predicted classes, after which the user can either select the correct one from the predicted classes (*positive feedback*), or provide the feedback that none of the candidates is correct (*negative feedback*). This reduces the feedback/annotation effort, compared with always asking for feedback with the ground truth label. Further, by training directly with the user feedback, we circumvent the necessity of saving and distributing the user data for crowdsource labeling, maintaining the user privacy.

Compared with requiring users to provide the exact label each time, the above feedback is simpler and alleviates the user efforts. However, the sequence labeling model trained with such feedback is likely to be biased, which we will explain in a casual perspective. In initial stages of interactive learning, model predictions may suffer from spurious correlation with certain context features (*confounders*) in the training data (Zhang et al., 2020b; Wang et al., 2020a), due to the bias from pretraining (Delobelle et al., 2021) or insufficient fine tuning. As a result, the model prediction will be confounded by such context features, such that its predictions lean towards a certain set of token classes (*i.e.*, *context bias*). Thus, when interacting with users, the model might receive negative feedback, since the ground truth label is unlikely displayed in the top K candidates. We term such a disparity of feedback as *feedback bias*. We can observe that the positive feedback of ground truth labels provides stronger supervision for training, since with negative feedback, we can only rule out K negative candidates. Consequently, the feedback bias will aggravate the context bias, with the model more sufficiently trained for data with the positive feedback, while remaining weakly supervised otherwise.

We formulate the interactive sequence labeling as a Structural Causal Model (SCM), and subsequently develop debiasing mechanisms. For the context bias, we identify the confounders as the most frequent context tokens and design a deconfounding layer based on the backdoor adjustment, reducing the effect of spurious correlation with the context features. To enable more efficient training on the negative feedback, we employ doubly robust estimation (Dudík et al., 2011) using an imputation model for weak supervision. However, with the proportion of negative feedback significantly larger than positive feedback, the feedback bias is likely to accumulate. Besides the context confounders Z, some unobserved confounders C (*e.g.*, users' prior knowledge and expertise (Wang et al., 2022, 2020b; Gao et al., 2021)) from user feedback can also form the backdoor paths and result in the feedback bias. Since the feedback bias results from both the observed confounder and some unobserved confounders, the backdoor adjustment may not be sufficient for de-biasing (Bahadori and Heckerman, 2021; Puli and Ranganath, 2020). Thus, we introduce an external random noise variable as the *instrumental variable* (Angrist and Krueger, 2001; Yue et al., 2020). To properly determine the random noise, we propose an information-theoretic causal de-biasing method, entropy-based modulation. Our approach can decrease the chances of negative feedback, when selecting the K candidates from model predictions for display.

Our contributions are summarized as follows.

- We study interactive sequence labeling that allows training directly with the user feedback, which alleviates the annotation cost and the user privacy concern by traditional crowdsource labeling. We fundamentally analyse the *context bias* and *feedback bias* involved in the interactive learning with a structural causal model (SCM).
- To alleviate the above two biases, we design a confounder layer that identifies the confounders as the most frequent context tokens and further propose an information-theoretic causal de-biasing method, entropy-based modulation, leveraging the relations between the predictions and context.

2 Interactive Sequence Labeling

In this section, we elaborate the procedures of our interactive sequence labeling on batches of streaming data of the users. Specifically, for each batch of the test sequence, the interaction between sequence labeling model and the users can be decomposed into: *display*, *feedback* and *training*.

Display: We apply the current sequence labeling model on the text sequences $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$

with length N. For each token $E \in X$, we collect the predicted probability distribution over $Y = [y_1, \ldots, y_M]$ with total M token labels, from which we select top K candidate labels based on the predictions and the entropy-based modulation detailed in Section 4.1. Then, the token E and its selected K candidates $\{d_1, \ldots, d_K\}$ are displayed to the users for feedback.

Feedback: As discussed in Section 1, our interactive sequence labeling involves two types of user feedback: *positive feedback* and *negative feedback*. Given a token and its K candidate labels, the user will provide a positive feedback via pointing out the ground truth label y if $y \in \{d_1, \ldots, d_K\}$. Otherwise, the user will feedback with "none of the above", *i.e.*, negative feedback.

Training: The sequence labeling model is periodically trained with the feedback from users over time. For a positive feedback, we train the model with the designed deconfounding layer and the cross-entropy loss with the ground truth label y. If a negative feedback is given, we train the model with the proposed entropy-based modulation and doubly robust estimation, described in Section 4.2.

3 Interactive Sequence Labeling from a Causal View

To fundamentally explain the biases and their sources in the interactive learning, we analyse interactive sequence labeling from the causal perspective and propose a Structural Causal Model (SCM).

3.1 SCM for Interactive Sequence Labeling

Figure 2 shows our SCM for our interactive sequence labeling, which contains 6 key variables in the interactive sequence labeling procedure: 1) *Token Embeddings*, embeddings of tokens from a pretrained BERT model. We reload E that refers to both the token and the token embedding; 2) *Context confounders* Z, which we identify as a set of context tokens potentially forming the *spurious correlation* with the entities; 3) *Prediction* Y, the predicted token labels. 4) *Displayed Labels* D, the K selected candidates based on Y; 5) *Unobserved confounders* C, from user feedback during interactions; 6) *Instrumental variable* ϵ, which is an external random noise variable independent from confounders C.

- *Context Bias:* In Figure 2(a), $Y \leftarrow Z \rightarrow E$ represents the backdoor paths which will introduce spurious correlation between Y and E. Based on the total probability theorem, $P(Y = y|E) = \sum_{i=1}^{K} P(Y = y|E, z_i)P(z_i|E)$. For some confounder $z \in Z$, the high value of $P(z|E)$ will make $P(Y = y|E, z)$ dominate the total probability. Thus, the prediction of the model is biased by coufounders Z.

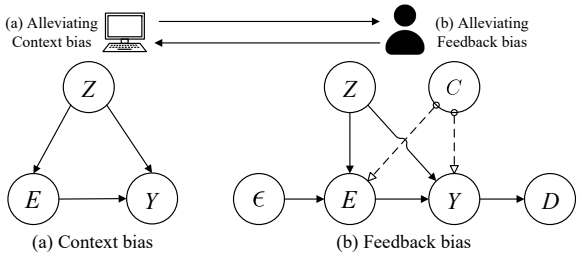

Figure 2: The proposed structural causal model for interactive sequence labeling.

- *Feedback Bias:* Our displayed candidates D are generated based on the model prediction Y. Besides the context confounders Z, some unobserved confounders C (*e.g.*, users' prior knowledge and expertise (Wang et al., 2022, 2020b; Gao et al., 2021)) from user feedback can also form the backdoor paths and result in the feedback bias. Since such confounders C are unobserved, we indicate their effects on E and Y with dashed lines. With the direct influence of the external randomness ϵ, the causal effect of confounders C are intervened, which makes ϵ an instrumental variable. In our considered interactive sequence labeling, the user can only provide a positive feedback when the ground truth label is included in the displayed D. As in Section 1, the feedback bias is caused by the disparity between such positive and negative feedbacks.

4 De-biasing Interactive Sequence Labeling

4.1 Alleviating Context Bias Based on Context Confounders

As mentioned in Section 1, predictions from sequence labeling model may suffer from spurious correlation with certain context features (confounders) in the training data. In alleviating such context bias, we add a deconfounding layer on top of a BERT encoder, leveraging the relation between model predictions and potential confounders.

Inspired by (Wang et al., 2020a) that identifies the context bounding boxes for visual tasks as

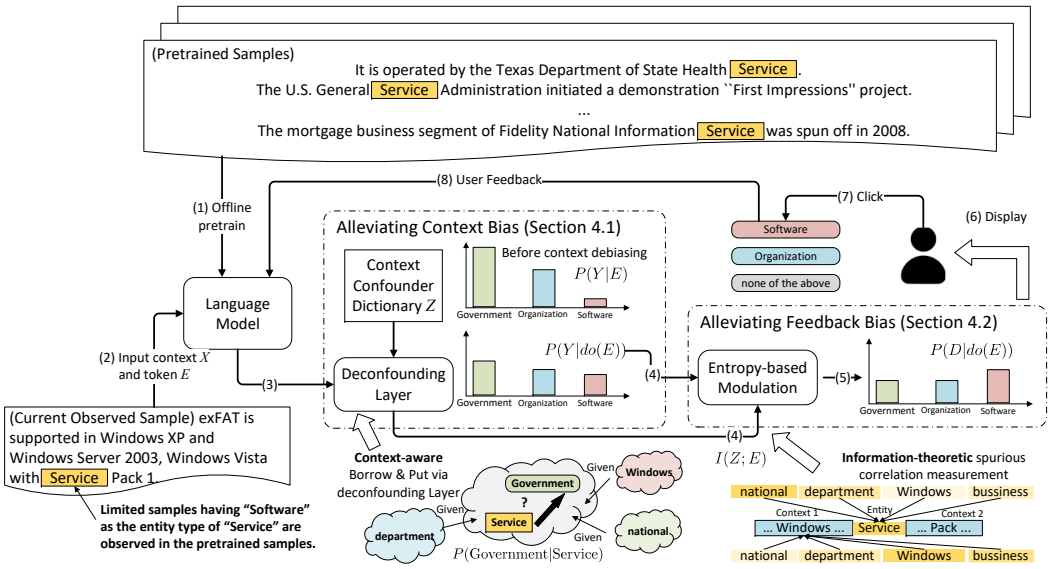

Figure 3: Our proposed context-aware information-theoretic causal de-biasing for interactive sequence labeling.

the confounders, we novelly identify the frequent context words as the potential confounders for sequence labeling, denoted as $Z = \{\mathbf{z}_i\}_{i=1}^{K}$, where K is the number of confounder tokens. We over load z_i for both the text token and its embedding. Following (Wang et al., 2020a; Zhang et al., 2020b), we leverage the confounders Z for the backdoor adjustment and define a deconfounding layer on top of the pretrained BERT encoder,

$$P(Y|do(E)) = \sum_{i=1}^{K} P(Y|E, \mathbf{z}_i)P(\mathbf{z}_i) \quad (1)$$

$$= \text{softmax}\left(\mathbf{W}_c \left[\mathbf{x}_i, \sum_{j=1}^{K} P(\mathbf{z}_j) \cdot \alpha_i(\mathbf{z}_j) \cdot \mathbf{z}_j\right]\right),$$

where E is the embedding of $\boldsymbol{x}_i$, the current token to be labelled in the i-th position of the sentence X, and

$$\alpha_i(\mathbf{z}_j) = \frac{\exp\left[(\mathbf{W}_x E)^\top (\mathbf{W}_z \mathbf{z}_j)\right]}{\sum_{k=1}^{K} \exp\left[(\mathbf{W}_x E)^\top (\mathbf{W}_z \mathbf{z}_k)\right]} \quad (2)$$

is the importance factor of $\mathbf{z}_j$ for $\boldsymbol{x}_i$, measured by the attention weight learned in the deconfounding layer. The attention score $\alpha_i(\mathbf{z}_j)$ measures the co-dependence between token embedding $\boldsymbol{x}_i$ and the confounder $\mathbf{z}_j$. In practice, we compute $P(\mathbf{z}_j)$ as the frequency of each confounder $\mathbf{z}_j$, $j = 1...K$ on the initial dataset (in Section 5). $P(\mathbf{z}_j)$ is normalized so that $\sum_{j=1}^{K} P(\mathbf{z}_j) = 1$.

As shown in Figure 3, when the context tokens "state" and "national" are observed, it is highly possible that "service" is predicted as "government". By also considering other confounders (*i.e.*, borrow & put (Wang et al., 2020a)) such as "windows", we are able to predict "service" as "software".

4.2 Alleviating Feedback Bias via Entropy-Based Modulation

To further alleviate the feedback bias, we leverage the doubly robust estimation (Dudík et al., 2011) to reduce the disparity between the positive and negative feedback. Since the negative feedback only contain information for weak supervision, *i.e.*, only ruling out K candidate labels that is not the ground truth, we instead train our sequence labeling model from an unbiased imputation model, denoted as σ. σ is another sequence labeling model trained offline (before interactive learning) on a small gold dataset with ground truth labels. σ is unbiased since it is not exposed to the positive and negative feedback online, *i.e.*, do not suffer from the feedback bias. For sample $\mathbf{x}_i$ that receives negative feedback from the user, our sequence label model is trained by

$$L_n = \sum_{y_i \in Y} -p_\sigma(y_i|\mathbf{x}_i) \log p(y_i|\mathbf{x}_i) \quad (3)$$

where $p_\sigma(y_i|\mathbf{x}_i)$ is the prediction from the imputation model σ. The imputation model is used when the user clicks "none of the above" (Figure 1) i.e., model receives negative feedback. In this case, instead of using "none of the above" as feedback, the model receives feedback as the predictions from the imputation model. We implement the imputation model with the pretrained BERT model (Devlin et al., 2018).

Entropy-based Modulation When the model suffers from the context bias due to spurious correlation with context confounders Z, the model prediction is likely biased. In addition, with the users involved in the sequence labeling process, there are several unobserved confounders existing in the interactions (Wang et al., 2022, 2020b; Gao et al., 2021), which can form the extra backdoor paths. For example, after the users from a specific knowledge background (*e.g.*, medical) have interacted with the system for a number of rounds, the data labels collected from the users can be biased. Thus, the model trained on these batches of data may display biased predictions for the following new user with a different knowledge background. Then, the user can only provide a negative feedback, with which the model cannot be directly supervised with the ground truth label. Consequently, the model will be biased toward data with positive feedback during interactive learning, resulting in the feedback bias.

Since the feedback bias results from both the observed confounder and some unobserved confounders, the backdoor adjustment may not be sufficient for de-biasing (Bahadori and Heckerman, 2021; Puli and Ranganath, 2020). Thus, we introduce an external random noise variable ϵ as the *instrumental variable* (Puli and Ranganath, 2020; Peysakhovich and Eckles, 2018; Angrist and Krueger, 2001; Yue et al., 2020). Under the assumption that ϵ is independent from confounders C, the token embedding E is intervened by ϵ and no longer dominated by the confounders. To properly determine the random noise, we propose an information-theoretic causal de-biasing method, entropy-based modulation. Following (Seo et al., 2022), we quantify the co-dependence between confounders Z and E via their mutual information,

$$\begin{aligned} I(Z;E) &= H(Z) - H(Z|E) \\ &\approx \left(\frac{1}{N}\sum_{i=1}^{N} H(Z|\mathbf{x}_i)\right) - H(Z|E), \end{aligned}$$

in which $H(\cdot|\cdot)$ denotes the conditional entropy and $H(Z)$ is estimated with the average effect of $H(Z|\mathbf{x}_i)$ over all the tokens in X. Formally,

$$H(Z|\mathbf{x}_i) = -\sum_{j=1}^{K} \log \alpha_i(\mathbf{z}_j) \cdot \alpha_i(\mathbf{z}_j),$$

where we approximate $P(Z|\mathbf{x}_i)$ as $\alpha_i(\mathbf{z}_j)$. Intuitively, as analysed in (1), the co-dependence can be modeled by the cross attention between the frequent text tokens Z (confounders) and the learnt token embeddings E. Thus, when the model prediction is confounded, we can expect a large attention value α_i on some context confounders, with which the learnt token embedding is correlated to a large extend. This inspires us that, by scaling the random noise ϵ with $I(Z;E)$, we can modulate the selection of the top K from model predictions via monitoring the attention value between Z and E.

Specifically, for prediction of a token $\mathbf{x}_i$ from a text sequence X, we consider selecting the top K candidates from the following modulated prediction distribution

$$\begin{aligned} &P(D|do(E)) \\ &= \text{softmax}\Big(\text{logit}P(Y|do(E)) + I\left(Z;E\right)\cdot\epsilon\Big), \end{aligned}$$

where $\epsilon \sim N(0,1)$ is the instrumental variable. Concretely, $P(D|do(E))$ is the modulated prediction distribution, with which we select the top K labels with the highest probability. The selected labels are displayed to the users for feedback.

If $I\left(Z;E\right)$ is relatively larger, it indicates that $\mathbf{x}_i$ is highly correlated with Z, and thus the backdoor path is constructed. In order for the model to focus on the direct path, we intervene with the instrumental variable ϵ by encouraging randomly selecting the entity types with lower predicted probabilities (with $P(Y|do(E))$) to be displayed to the users. We denote our method as **C**ontext-aware **I**nformation-**T**heoretic **I**nteractive **S**equence **L**abeling (**CITISL**).

5 Experiment

5.1 Experimental Settings

Dataset and Metric. We evaluate our method on two POS tagging datasets and two NER datasets. **CoNLL-2003-POS** (Tjong Kim Sang and De Meulder, 2003) includes 34 types of tags and contains about $14.4K$ training data and $3.4K$ test data. The **UD-ENG** (Nivre et al., 2015) dataset includes 15 types of tags. This data has about $37K$ training data and $7.4K$ test data. The **Few-NERD** dataset (Ding et al., 2021) which contains 66 classes of entities. The Few-NERD dataset has $127.6K$ training data and $37.6K$ test data. The **OntoNote** dataset (Hovy et al., 2006) which contains 18 classes of entities. The OntoNote dataset has about $40.6K$ training data and $13.5K$ test data.

	CoNLL-2003-POS			UD-ENG			Few-NERD			OntoNote		
Initial Data	0.1	0.15	0.2	0.1	0.15	0.2	0.03	0.05	0.1	0.1	0.15	0.2
BERT	$66.71_{\pm 0.06}$	$66.89_{\pm 0.17}$	$66.68_{\pm 0.12}$	$48.51_{\pm 5.18}$	$54.92_{\pm 7.96}$	$46.43_{\pm 5.05}$	$50.59_{\pm 0.84}$	$51.57_{\pm 0.42}$	$51.47_{\pm 0.56}$	$75.53_{\pm 0.0}$	$75.79_{\pm 0.34}$	$75.09_{\pm 0.39}$
CBA	$66.81_{\pm 0.08}$	$66.81_{\pm 0.08}$	$66.87_{\pm 0.09}$	$53.15_{\pm 0.82}$	$50.75_{\pm 4.21}$	$50.21_{\pm 4.49}$	$51.44_{\pm 0.75}$	$50.70_{\pm 0.36}$	$51.61_{\pm 0.31}$	$78.97_{\pm 0.61}$	**82.02**$_{\pm 0.21}$	$81.51_{\pm 0.18}$
IS	$66.69_{\pm 0.04}$	$66.83_{\pm 0.06}$	$66.82_{\pm 0.13}$	$54.25_{\pm 3.81}$	$54.83_{\pm 4.71}$	$52.28_{\pm 0.96}$	$50.20_{\pm 1.42}$	$51.62_{\pm 0.56}$	$51.66_{\pm 0.48}$	$76.89_{\pm 0.33}$	$75.92_{\pm 0.96}$	$75.65_{\pm 0.82}$
CITISL	**85.90**$_{\pm 0.66}$	**91.89**$_{\pm 0.07}$	**91.39**$_{\pm 0.65}$	**80.22**$_{\pm 6.48}$	**75.06**$_{\pm 0.08}$	**79.98**$_{\pm 5.57}$	**62.48**$_{\pm 0.67}$	**62.57**$_{\pm 1.36}$	**65.14**$_{\pm 0.29}$	$78.66_{\pm 0.21}$	$81.89_{\pm 0.31}$	$81.16_{\pm 0.35}$
- EM	$85.36_{\pm 1.29}$	$85.09_{\pm 2.02}$	$83.94_{\pm 0.86}$	$54.99_{\pm 3.68}$	$54.85_{\pm 9.63}$	$53.85_{\pm 4.72}$	$61.63_{\pm 0.40}$	$60.63_{\pm 0.73}$	$62.57_{\pm 0.83}$	**81.42**$_{\pm 3.88}$	$79.64_{\pm 6.74}$	**83.35**$_{\pm 2.59}$
- EM - CBA	$85.64_{\pm 1.33}$	$83.47_{\pm 0.12}$	$83.54_{\pm 0.11}$	$53.55_{\pm 0.76}$	$58.10_{\pm 7.06}$	$61.46_{\pm 7.80}$	$61.36_{\pm 1.03}$	$61.78_{\pm 0.55}$	$62.71_{\pm 0.79}$	$76.71_{\pm 1.09}$	$75.54_{\pm 0.16}$	$74.79_{\pm 0.35}$
GroundTruth	$86.22_{\pm 0.76}$	$86.22_{\pm 0.76}$	$86.22_{\pm 0.76}$	$64.38_{\pm 11.5}$	$64.38_{\pm 11.5}$	$64.38_{\pm 11.5}$	$67.26_{\pm 0.97}$	$67.26_{\pm 0.97}$	$67.26_{\pm 0.97}$	$78.00_{\pm 0.47}$	$78.00_{\pm 0.47}$	$78.00_{\pm 0.47}$
TotalData		$91.63_{\pm 0.0006}$			$93.52_{\pm 0.0008}$			$68.59_{\pm 0.004}$			$86.78_{\pm 0.004}$	

Table 1: F1 score comparison results on CoNLL-2003, UD-ENG, Few-NERD and OntoNote.

We finetune the model on a portion of the dataset before interactive learning. In practice, we expect the data available for initial training should be limited to minimize the reliance on ground truth labels. Specifically, on CoNLL-2003-POS, UD-ENG and OntoNote5 we use 10%, 15%, 20% of the training data to train an initial NER model. On the Few-NERD dataset we use 3%, 5% and 10% to train the initial model. Afterwards, the initial model interacts with the end user on the remaining training data, receives user feedback, and learns online. During the interactions, the model expects to receive feedback from the end users. Similar to (Shen et al., 2018), we calculate the F1 score of the models on the test data after each interaction (*i.e.*, time step). The average results on the test data over 10 runs with standard errors are reported.

Baselines. We evaluate on several baselines: (i) **BERT.** We finetune BERT$_{base}$ (Devlin et al., 2018) encoder and a linear projection layer for prediction. This is a simple baseline without de-biasing. (ii) **C**ontext **B**ias **A**lleviation (**CBA**). We alleviate the pretrained correlation bias by following (Zhang et al., 2020b). Instead of considering both vision confounders and language confounders, we only leverage the language confounders which are context words. We set the size of the confounder dictionary as 100 and the embedding length as 768. (iii) **IS.** We alleviate the feedback bias by importance sampling (Kloek and van Dijk, 1978). We keep tracking the frequencies of user feedback labels and use them as importance weights to reweight the feedback loss. (iv) **GroundTruth.** During each interaction, it is assumed that the model can always receive the feedback indicating the exactly correct labels (Shen et al., 2018; Fang et al., 2017; Radmard et al., 2021). This assumption is unrealistic considering (i) there are many entity types while the space of displaying the entity types to the users is limited, as discussed in Section 1 and (ii) it is difficult for the users to provide the perfectly correct label when all entity types can be displayed but the number of types is large, as shown later in the human evaluation in Section 6.2. GroundTruth is to understand the effect of bias on the model. Specifically, it evaluates how the biased model will perform with correct labels provided during the interactions. (v) **TotalData.** Different from GroundTruth, TotalData assumes crowd-sourced training. Specifically, the whole dataset is crowd-sourced and correctly annotated before model training. TotalData is to understand the upper bound performances of the model trained on the data with ground truth labels. Since crowd-source labeling requires more human effort, training with TotalData can be expensive and is also not suitable for privacy-preserving domain. Note that our model is trained with the user feedback, among which the negative feedback may not contain ground truth labels. Differently, the crowd-sourced annotations with TotalData are usually in higher quality with ground-truth labels, i.e., from paid annotators, which may contribute to data efficiency and convergence of training, but at the expense of 1) expensive and time-consuming annotating. 2) risking the user privacy in distributing data to crowd-sourced annotators.

Our model also have several variants. We denote our propose **E**ntropy-based **M**odulation as **EM** and **C**ontext **B**ias **A**lleviation as **CBA**. *i*) **CITISL-EM**: Our **CITISL** without **EM**. *ii)* **CITISL-EM-CBA**: Our **CITISL** without both **EM** and **CBA**. Note that both **CITISL-EM-CBA** and **CITISL-EM** are trained along with doubly robust training.

5.2 Main Results

Alleviating Context Bias To validate the effectiveness of our proposed context deconfounding method, we compare our **CITISL-EM** with its variants and the baselines, considering alleviating context bias. The results are shown in Table 1. There are several observations. Firstly, our approach can successfully alleviate the context correlation bias, compared with the baselines. On

	CoNLL-2003-POS			UD-ENG			Few-NERD			OntoNote		
Initial Data	0.1	0.15	0.2	0.1	0.15	0.2	0.03	0.05	0.1	0.1	0.15	0.2
CITISL	**74.05**$_{\pm 2.98}$	**97.05**$_{\pm 0.13}$	**96.50**$_{\pm 0.43}$	**77.68**$_{\pm 8.83}$	**70.48**$_{\pm 0.29}$	**77.55**$_{\pm 7.72}$	**54.86**$_{\pm 3.78}$	**54.56**$_{\pm 4.37}$	**58.65**$_{\pm 0.83}$	76.25$_{\pm 0.48}$	75.86$_{\pm 0.73}$	75.96$_{\pm 0.63}$
- EM	72.54$_{\pm 4.96}$	71.06$_{\pm 7.72}$	66.43$_{\pm 3.37}$	42.97$_{\pm 5.21}$	43.05$_{\pm 13.1}$	41.86$_{\pm 6.37}$	47.57$_{\pm 2.89}$	49.06$_{\pm 1.37}$	49.39$_{\pm 1.16}$	77.76$_{\pm 2.43}$	69.57$_{\pm 15.2}$	**79.52**$_{\pm 4.19}$
- EM - CBA	73.90$_{\pm 5.55}$	64.81$_{\pm 0.07}$	64.72$_{\pm 0.07}$	41.24$_{\pm 0.89}$	47.54$_{\pm 9.72}$	52.00$_{\pm 10.6}$	48.21$_{\pm 1.67}$	49.41$_{\pm 1.91}$	49.53$_{\pm 0.55}$	**78.88**$_{\pm 1.77}$	**76.07**$_{\pm 0.46}$	76.02$_{\pm 0.50}$
GroundTruth	74.98$_{\pm 3.02}$	74.98$_{\pm 3.02}$	74.98$_{\pm 3.02}$	55.65$_{\pm 15.8}$	55.65$_{\pm 15.8}$	55.65$_{\pm 15.8}$	73.07$_{\pm 1.20}$	73.07$_{\pm 1.20}$	73.07$_{\pm 1.20}$	80.90$_{\pm 0.20}$	80.90$_{\pm 0.20}$	80.90$_{\pm 0.20}$
TotalData		97.06$_{\pm 0.0006}$			95.97$_{\pm 0.003}$			74.43$_{\pm 0.006}$			86.16$_{\pm 0.008}$	

Table 2: Analysis of our method and its ablations (**CITISL-EM** and **CITISL-EM-CBA**). We experiment on interactive learning with classes that are unseen in initial training.

ConLL-2003-POS, the improvements of **CITISL-EM** over **BERT** are 27.95%, 27.21% and 25.89% on 10%, 15% and 20%, respectively. On Few-NERD, the improvements of **CITISL-EM** over **BERT** are 21.8%, 17.6% and 21.2% when 3%, 5% and 10% initial data are used. On UD-ENG and OntoNote, we observe the similar improvements. Secondly, compared with only using doubly robust training, context deconfounding can further alleviate the context bias. On the ConLL-2003-POS dataset, **CITISL-EM** achieves 1.94% final improvement over **CITISL-EM-CBA**, when the initial dataset size is 15%. On OntoNote5 with 20% initial data, **CITISL-EM** gains 11.45% improvement over **CITISL-EM-CBA**. When the initial data sizes are 15% and 10%, the improvements of **CITISL-EM** over **CITISL-EM-CBA** are 5.43% and 6.14%.

Alleviating Feedback Bias To validate the effectiveness of our proposed entropy-based modulation, we compare our method **CITISL** with its variants and the baselines. By the entropy-based modulation, our approach can learn on the entities with unobserved labels more sample-efficiently. There are some observations showing that our proposed entropy-based modulation can successfully alleviate feedback bias. On Few-NERD, when there are 5% and 10% initial data, **CITISL** outperforms **CITISL-EM** by 3.2% and 4.1% respectively. On ConLL-2003-POS, when 10% initial data are used, **CITISL** and **CITISL-EM** are comparable. When the initial data increase to 15% and 20%, the improvements are increased to 8.0% and 8.9%. On UD-ENG, when 10%, 15% and 20% data are used, **CITISL** gains 45.9.4%, 36.8% and 48.5% improvements over **CITISL-EM**.

We further compare our proposed method, **CITISL**, to GroundTruth, which has access to the perfectly correct labels at each time. The results are in Table 1. Our approach achieves very similar performance, compared to GroundTruth. In some cases, our approach outperforms GroundTruth, because GroundTruth learns on datasets with imbalanced entity types which is actually the context bias, while our de-biased approach encourages the model to also learn on the minor entity types (*i.e.*, entities with lower probabilities to display) and alleviates the feedback bias. All the baseline models are trained in an interactive way as our method with the training dataset split being exactly the same as in our method. Please refer to Appendix A.2 for the learning curves, showing the behaviors of different algorithms during the interactive learning.

6 Discussion

6.1 Analysis

In the main experiments, some classes are made unseen to the model before interactions, to simulate a biased sequence labeling setting. To further analyze our proposed causal de-biasing methods, *i.e.*, **CITISL** and its ablations (**CITISL-EM** and **CITISL-EM-CBA**), we report the F1 score comparison results only on unseen classes (entity types). The results are shown in Table 2. On ConLL-2003-POS, there are 10.3K tokens with unseen classes and 29.6K tokens with seen classes. On UD-ENG, classes of 40.7K tokens are unseen and classes of 15.2K tokens are seen. On Few-NERD, there are 43.8K tokens with unseen classes and 147.6K tokens with seen classes. On OntoNote, classes of 13.5K tokens are unseen and classes of 19.0K tokens are seen. From Table 2, our proposed methods result in superior performance that can be close to TotalData and GroundTruth. Additionally, **CITISL** has much better performance than its ablations (**CITISL-EM** and **CITISL-EM-CBA**). This shows that our proposed entropy-based modulation and context debasing can actively explore and display the classes which are uncertain to the current trained model. Thus, collected user feedback is less biased and more efficient for interactive learning with our model.

6.2 Human Evaluations

We further conduct human evaluations, to understand the (i) quality of the feedback provided by

the human and (ii) required efforts (*i.e.*, labeling time) when users provide feedback to the model update using our approach. By our approach, the users are required to select the correct entity from the top-K predictions by our model or 'none of the above'. For comparison, the users are also required to select the correct entity from all candidate entities, as in the traditional annotation mechanism (Shen et al., 2018; Fang et al., 2017; Radmard et al., 2021). The human evaluations are conducted on the FewNERD dataset. We collect human feedback in 1600 sessions by 5 users. In each session, the system shows a sentence with predicted candidates to the user, and the user provides feedback.

In the evaluation of the quality of the feedback provided by the human, by traditional annotation mechanism, 41.5% sentences receive groundtruth feedback from the users (*i.e.*, users label the correct type). By our feedback mechanism, 44.5%, 43.8% and 43.0% sentences receive groundtruth feedback from the users, when $K = 3, 5$ and 7, respectively. If we only consider the (68.3%, 70.0% and 70.5%) sentences where the groundtruth entity type is within the top-K predictions by our model, 65.2%, 62.5% and 61.0% sentences receive groundtruth feedback, when $K = 3, 5$ and 7, respectively. In the evaluation of labeling efforts, by traditional annotation mechanism, in average 23.96 seconds are needed to label a sentence. By using the feedback studied in our paper, in average 16.99 seconds are needed. The above results show that by using the studied feedback, we can collects feedback of higher quality with less human labeling efforts, compared to the traditional annotation mechanism.

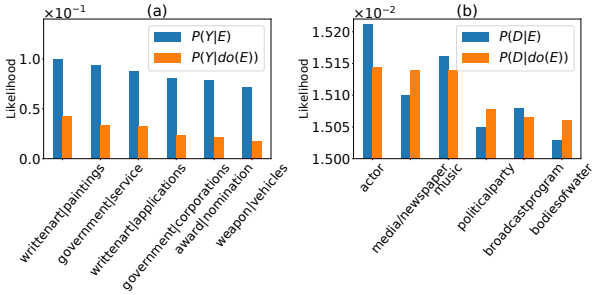

Figure 4: Following (Devlin et al., 2018; Wang et al., 2020a; Rao et al., 2021), we show differences between the likelihood before and after de-biasing. (a) shows the difference by alleviating the context bias and (b) shows the difference by alleviating the feedback bias.

6.3 Case Study

To analyze the effect of our proposed causal intervention methods, we show some examples of the likelihood before and after the intervention on the context bias and feedback bias respectively. In Figure 4(a), before intervention on the context bias, the likelihood of $P(Y|E)$ can be significantly high due to spurious correlations. For example, because the token "service" and the label "government" frequently co-occur in the imbalanced dataset, $P(government|service)$ can dominate the prediction probabilities. With the backdoor adjustment intervention (Wang et al., 2020a; Zhang et al., 2020b), $P(government|do(service))$ is lower and enables the model to predict other possible labels (e.g. organization, software). In Figure 4(b), we show the average display probabilities of six token labels. Due to the context bias, some frequently occurred labels (*e.g.*, actor, music) have relatively much higher probabilities to be displayed. The introduced random noise instrumental variable ϵ modulates the display probabilities. Scaled by the measurement of spurious correlations, the modulation can promote labels with low prediction probabilities to be displayed.

7 Related Work

Causal Inference for Sequence Labeling Previous works (Zhang et al., 2021; Nan et al., 2021; Zeng et al., 2020) identified several causal effects in sequence labeling problems including named-entity recognition and part-of-speech tagging. To identify the confounders and analyze spurious correlations in their models, they fundamentally formulate their problems via a Structural Causal Model (SCM). Further, they employ the backdoor adjustment (Zhang et al., 2021; Nan et al., 2021; Zeng et al., 2020) to remove the spurious correlation introduced by the backdoor paths in the SCM.

Active Learning Given a dataset of unlabeled samples, active learning aims at selecting the most task-informative subset of samples (queries) for labeling, so as to maximize the model performance trained with the acquired labels, while minimizing the annotation cost (Shen et al., 2018; Fang et al., 2017; Radmard et al., 2021; Zhang et al., 2020a; Siddhant and Lipton, 2018; Yao et al., 2019; Shelmanov et al., 2019). Our setting of interactive sequence labeling is different from active learning in the following perspectives: *i)* Unlike active learning that select samples for annotation from an

unlabeled dataset, our sequence labeling model is trained with batches of streaming data. *ii)* We send all the data of each batch to the users, *i.e.*, we do not perform sample selection, but instead focusing on alleviating the context and feedback biases in interactive learning.

De-biasing for Interactive Learning (Qian et al., 2020) studies the task of interactive learning for named entity normalization. However, different from our work, they ignore the bias problems involved with the top K display. Previous works (Lakkaraju et al., 2017; Swaminathan and Joachims, 2015a; Yuan et al., 2019) realized the importance of handling non-displayed events. They regarded non-displayed events as unlabeled instances, and model the CTR prediction as a learning problem with labeled and unlabeled instances, which aims to learn under covariate shift (sample bias corrections). Several counterfactual estimators have been developed. Importance sampling (IS) is a simple way to tackle this issue, but suffers from high variances. Classic variance reduction techniques (Bottou et al., 2013; Li et al., 2015; Swaminathan and Joachims, 2015b) for IS are useful for counterfactual evaluation and learning. Inverse propensity score (IPS) (Horvitz and Thompson, 1952) weights each labeled event with the inverse of its propensity score, which is determined by the likelihood of the logged data. Doubly robust for counterfactual learning (Dudík et al., 2011) takes advantage of the IPS (Horvitz and Thompson, 1952) and direct method (Yuan et al., 2019) to increase the chances of accurate ratio estimations.

8 Conclusion

The state-of-the-art sequence labeling models rely on an adequate amount of labeled data. Crowdsource labeling for sequence labeling can be expensive, time-consuming, and not be appropriate for data containing user private information. We study how to efficiently and accurately train sequence labeling models directly with the user feedback, with a simple feedback mechanism. Moreover, we fundamentally analyze and explain the biases involved in interactive sequence labeling, formulating interactive sequence labeling from a causal view and propose a structural causal model. Based on the structural causal model, we learn debiased interactive sequence labeling, via identify the confounders as the most frequent context tokens for the backdoor adjustment and further propose an information-theoretic causal de-biasing method, *i.e.*, entropy-based modulation, by leveraging the relations between the contexts and entity tokens. With extensive evaluations, we validate the effectiveness of our proposed de-biasing approaches.

9 Limitations

One of the limitations is that our interactive learning is based on user feedback on the token-level prediction of the NER model. This may not be convenient for NER models that generates sequence-level prediction, *e.g.*, those predicts with a Conditional Random Field (CRF) (Sutton et al., 2012) module, with which we need additional forward and backward operations to extract the token-level predictions. It would be interesting for the future work to also consider user feedback on the sequence level prediction with CRF-based NER models.

References

Joshua D Angrist and Alan B Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85.

Mohammad Taha Bahadori and David Heckerman. 2021. Debiasing concept-based explanations with causal analysis. In *ICLR*.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR*.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Hai-Tao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. *arXiv preprint arXiv:2105.07464*.

Ehsan Doostmohammadi, Minoo Nassajian, and Adel Rahimi. 2020. Persian ezafe recognition using transformers and its role in part-of-speech tagging. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 961–971.

Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1097–1104, Madison, WI, USA. Omnipress.

Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

Junruo Gao, Mengyue Yang, Yuyang Liu, and Jun Li. 2021. Deconfounding representation learning based on user interactions in recommendation systems. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 588–599. Springer.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. 2020. Multi-scale self-attention for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7847–7854.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45.

Daniel G Horvitz and Donovan J Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

T. Kloek and H. K. van Dijk. 1978. Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica*, 46(1):1–19.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lihong Li, Rémi Munos, and Csaba Szepesvári. 2015. Toward minimax off-policy value estimation. In *AISTATS*. PMLR.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.

Dat Quoc Nguyen et al. 2021a. Phonlp: A joint multi-task learning model for vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 1–7.

Minh-Tien Nguyen, Guido Zuccon, Gianluca Demartini, et al. 2021b. Loss-based active learning for named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, et al. 2015. Universal dependencies 1.2.

Alexander Peysakhovich and Dean Eckles. 2018. Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference*, pages 699–707.

Aahlad Puli and Rajesh Ranganath. 2020. General control functions for causal effect estimation from instrumental variables. *Advances in Neural Information Processing Systems*, 2020-December.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Partner: Human-in-the-loop entity name understanding with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13634–13635.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence based deep active learning for named entity recognition. In *ACL/IJCNLP (1)*, volume 1, pages 4310–4321. Association for Computational Linguistics.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2022. Information-theoretic bias reduction via causal view of spurious correlation.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V Dylov. 2019. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489. IEEE.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

Aditya Siddhant and Zachary C Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Adith Swaminathan and Thorsten Joachims. 2015a. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755.

Adith Swaminathan and Thorsten Joachims. 2015b. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020a. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020b. Causal inference for recommender systems. In *Fourteenth ACM Conference on Recommender Systems*, pages 426–431.

Zhenlei Wang, Shiqi Shen, Zhipeng Wang, Bo Chen, Xu Chen, and Ji-Rong Wen. 2022. Unbiased sequential recommendation with latent confounders. In *Proceedings of the ACM Web Conference 2022*, pages 2195–2204.

Lei Xu, Shuang Li, Yuchen Wang, and Lizhen Xu. 2021. Named entity recognition of bert-bilstm-crf combined with self-attention. In *International Conference on Web Information Systems and Applications*, pages 556–564. Springer.

Ziyu Yao, Yu Su, Huan Sun, and Wen-tau Yih. 2019. Model-based interactive semantic parsing: A unified framework and a text-to-sql case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5447–5458.

Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 329–338.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020a. Seqmix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020b. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4803–4813.

A Appendix

A.1 User Feedback

We show some illustrations of how users interact with the system and provide feedback in our human evaluations (Section 6.2). Figure 5 shows the main page with introductions.

Figure 6 shows how the user interacts with the system in the GroundTruth approach. In this case, the user is provided all the candidate types and selects the most suitable one. It can be very time-consuming and requires more human effort, since the user needs to scroll down, compare all the 66 candidates and select the most suitable one.

Figure 7 shows how the user interacts with the system in our proposed approach CITISL. In this case, the user gets the most likely prediction "person-athlete", as well as two more candidates "person-other" and "other-livingthing". Among the top 3 candidates, it is easier for the user to decide the most suitable one, compared to that in the GroundTruth approach. If the user cannot find the most suitable one, the user will select "none of the above". Due to the context bias and feedback bias, it is possible that the correct entity type is not within the top 3 candidates. By our causal de-biasing approach, we alleviate the bias and the model can efficiently learns to identify the correct entity type , including it in its top 3 predictions.

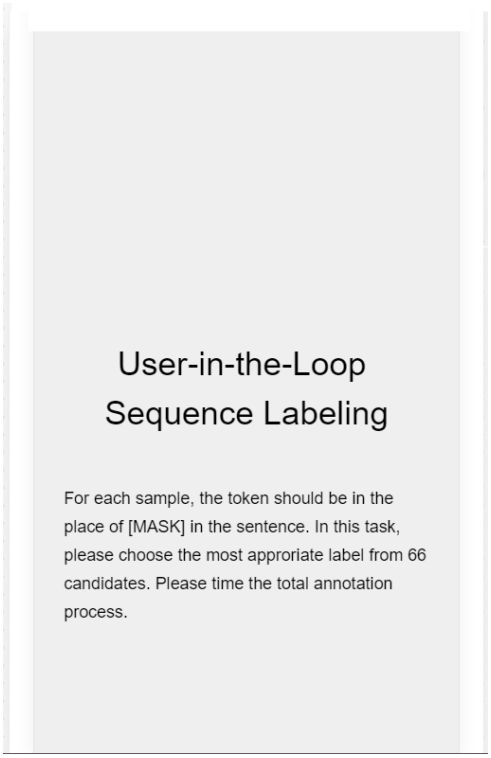

Figure 5: Main page.

A.2 Learning Curve

To further understand the behaviors of different algorithms during the interactive learning and validate the efficiency of our algorithm, we also show the learning curves by different algorithms, in addition to the comparison results shown in Table 1. The curves are shown in Figure 8, 9, 10 and 11. We can observe that our approach improves quickly and outperforms most of the baselines even in the early stages. On CoNLL-2003 and UD-ENG, our approach outperforms GroundTruth, because GroundTruth learns on datasets with imbalanced

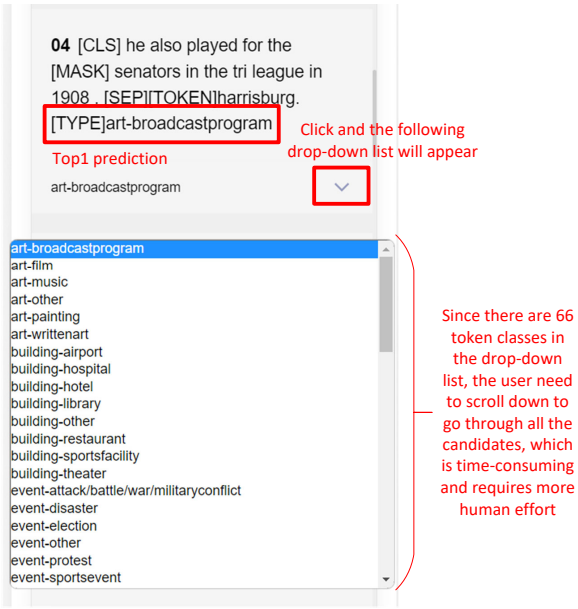

Figure 6: How the user interacts with the system in the GroundTruth approach.

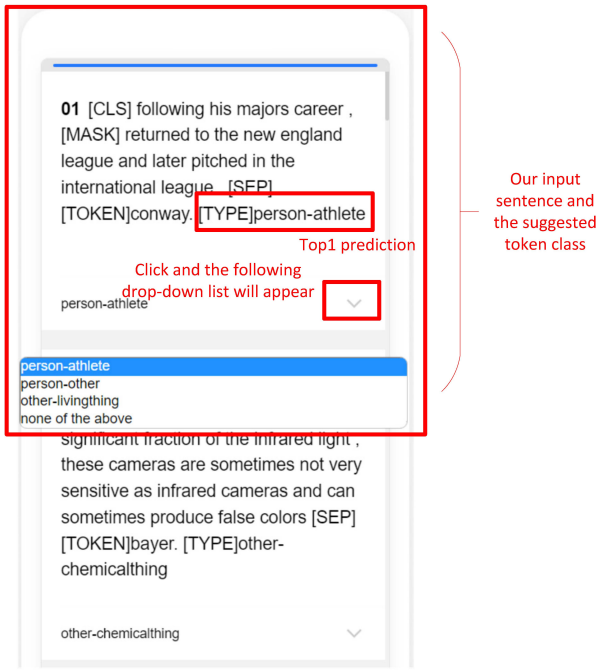

Figure 7: How the user interacts with the system in our proposed approach, CITISL. In this example, the top 3 candidates are suggested to the user.

entity types which reflects context bias, while our de-biased approach encourages the model to also learn on the minor entity types, alleviating the feedback bias.

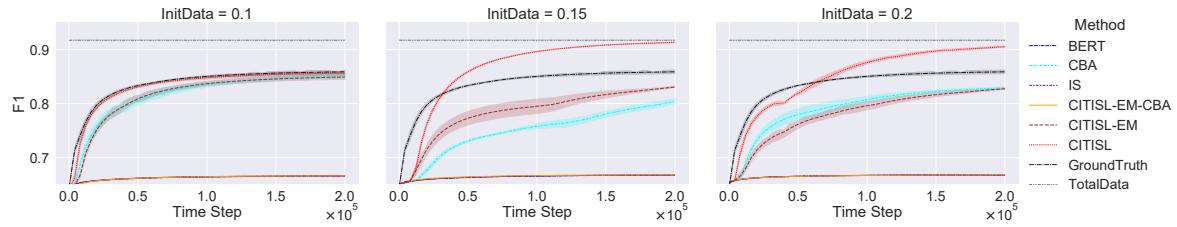

Figure 8: Comparisons between different approaches on CoNLL-2003.

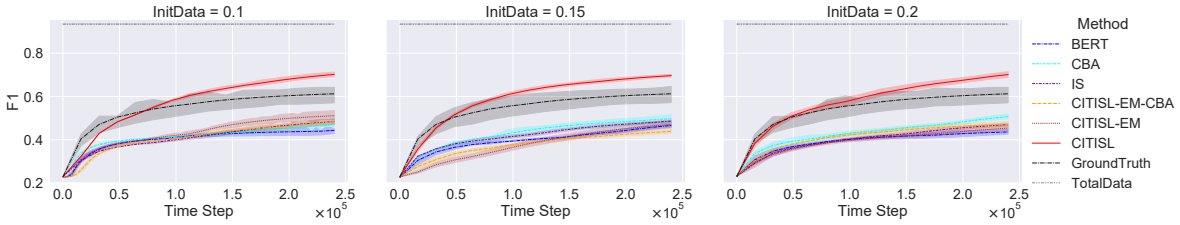

Figure 9: Comparisons between different approaches on UD-ENG.

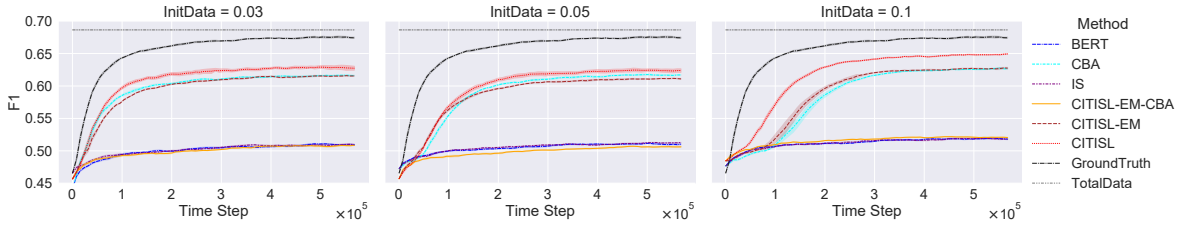

Figure 10: Comparisons between different approaches on Few-NERD.

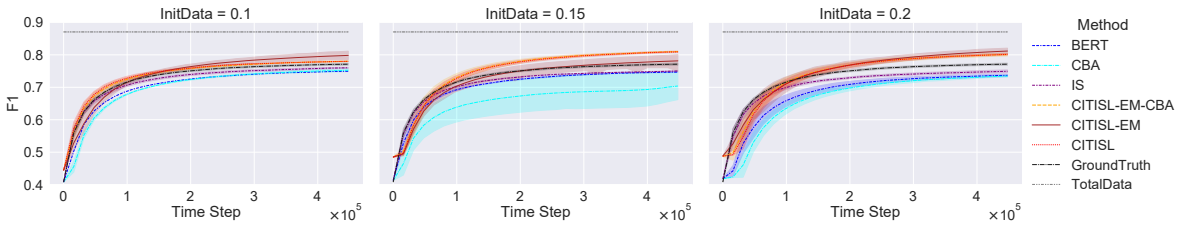

Figure 11: Comparisons between different approaches on OntoNote.