Mitigating Mention Surface Bias for Entity Disambiguation

Anonymous ACL submission

Abstract

Entity disambiguation (ED) is a foundational task in NLP for question-answering and information extraction applications. One of the main challenges of ED in real-world settings is the handling of overshadowed entities, i.e., the entities that share mention surfaces with common entities. The current approach for handling overshadowed entities relies on the coherence of entities in a given text, which is not always available and requires additional computing resources. In this paper, we formulated a causal graph for ED and found that the mention surfaces can act as a shortcut, misleading the ED models to be biased towards common entities. We propose a simple yet effective debiasing method that mitigates the effect of the mention surfaces on model predictions. Experimental results demonstrate that our method yields the best results against overshadowed entities. Source code and models will be publicly available upon acceptance.

1 Introduction

Entity disambiguation (ED) is an essential task in many natural language processing (NLP) applications, for instance, open-domain question answering (Hu et al., 2022; Saffari et al., 2021; Srivastava et al., 2021; Wang et al., 2021), fact verification (Zhou et al., 2019), and information extraction (Baldini Soares et al., 2019). The task is to identify the correct entity recorded in a knowledge base (KB), e.g., Wikidata, for each ambiguous entity mention in a given text, which is a crucial capability when performing entity linking (EL). For example, the entity mention *Michael Jordan* can refer to a machine learning researcher (Michael Irwin Jordan) or a basketball player (Michael Jeffrey Jordan), depending on contextual information. The ability to disambiguate different entities sharing the same surface form enables us to link each mention to its correct entity accurately.

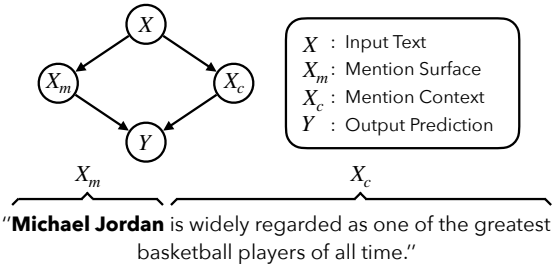

Figure 1: The causal graph of ED models.

One popular approach to mitigate the ambiguity of entity mentions is fine-tuning pre-trained language models (PLMs) to leverage contextual information. Classification-based approaches fine-tune a classification layer on top of a PLM to produce scores over entity vocabulary (Broscheit, 2019; Yamada et al., 2022) or entity types (Onoe and Durrett, 2020; Tedeschi et al., 2021). Generative-based approaches fine-tune a generative PLM to generate a unique entity name (Cao et al., 2021; De Cao et al., 2021; Du et al., 2022) and entity description (Procopio et al., 2023) for each mention. Retrieval-based approaches fine-tune a PLM to align the representations between entity mentions and entity descriptions (Wu et al., 2020; Li et al., 2020). Additionally, entity types, entity relations, and entity priors are used to improve the retrieval-based ED method (Ayoola et al., 2022a,b). While these methods effectively mitigate the ambiguity problem, they are still negatively affected by class imbalance. Provatorova et al. (2021) reveal a tendency of ED models to bias towards common entities, resulting in poor performance against overshadowed entities.

In this paper, we formulate the ED problem as a causal graph (Joshi et al., 2022) shown in Figure 1. There are two paths determining the output Y: via the mention surface X_m and via the context X_c. Ideally, we want the output Y to rely more on X_c, *"... the greatest basketball players of all time"*, than X_m, *"Michael Jordan"*. However, X_m can act as a shortcut, leading the model to produce correct answers for common entities without utilizing the

intended feature X_c. Based on this observation, we propose a *Counterfactual Training (CFT)* method to increase sensitivity to X_c and prevent the model from over-relying on X_m to determine Y.

We evaluate the effectiveness of the CFT on two ED benchmarks against three competitors, including the current state-of-the-art method, KBED (Ayoola et al., 2022a). Experimental results show that CFT is the best performer for overshadowed entities, demonstrating the robustness of CFT to mention surface bias. Regarding inference speed vs overall performance, CFT is 6.7 times faster than KBED while obtaining competitive overall scores.

2 Method: Counter Factual Training

The typical training objective of ED is to minimize the negative log-likelihood between the gold entity label Y' and the model prediction Y given a mention surface X_m and mention context X_c:

$$\begin{aligned} obj &= \min_{\theta} \mathcal{L}(Y', Y) \\ Y &= f(X_m, X_c, \theta) \end{aligned} \tag{1}$$

where $\mathcal{L}$ is any loss function (e.g., cross-entropy) and θ is parameters of the model f. However, due to the spurious correlation between mention surfaces X_m and training labels Y', training the model in this manner could mislead it to use the shortcut (X_m) when making predictions during inference.

To encourage the models to rely more on the mention context X_c rather than on the mention surface X_m, we introduce CounterFactual Training (CFT). During training, we create a counterfactual view $\hat{X}$ for each training example X by performing an intervention $do_{mask\_mention}(\cdot)$ to mask all mention surface tokens with a special [MASK] token and leave the mention context tokens as original as shown in Figure 2.

$$\begin{aligned} \hat{X} &= do_{mask\_mention}(X) = \langle w_1, w_2, ..., w_n \rangle \\ \forall w_i \in X, &\begin{cases} w_i \leftarrow \text{[MASK]} & \text{if } w_i \in X_m \\ w_i \leftarrow w_i & \text{if } w_i \in X_c \end{cases} \end{aligned} \tag{2}$$

Consequently, we derive a training objective for the counterfactual view $\hat{X}$ containing $\hat{X}_m$ and $\hat{X}_c$.

$$\begin{aligned} \hat{obj} &= \min_{\theta} \mathcal{L}(Y', \hat{Y}) \\ \hat{Y} &= f(\hat{X}_m, \hat{X}_c, \theta) \end{aligned} \tag{3}$$

This training objective (Eq. 3) enforces the model to rely on contextual information. Nevertheless, training the model solely on this objective can lead to a collapse of the embedding space due to the collision of entities in the similar contexts. For example, let X_1 = "*England won the FIFA World Cup.*" and X_2 = "*Spain won the FIFA World Cup.*", the entity mentions in these examples (i.e., "*England*" and "*Spain*") refer to different entities in KB, thus, requiring different embeddings. However, the counterfactual view of these two examples is identical $\hat{X}_1 = \hat{X}_2$ = "*[MASK] won the FIFA World Cup.*", resulting in the collision of both entity mentions. To stabilize the training process, we combine the two training objectives (Eq. 1 and Eq. 3) and derive the final training objective:

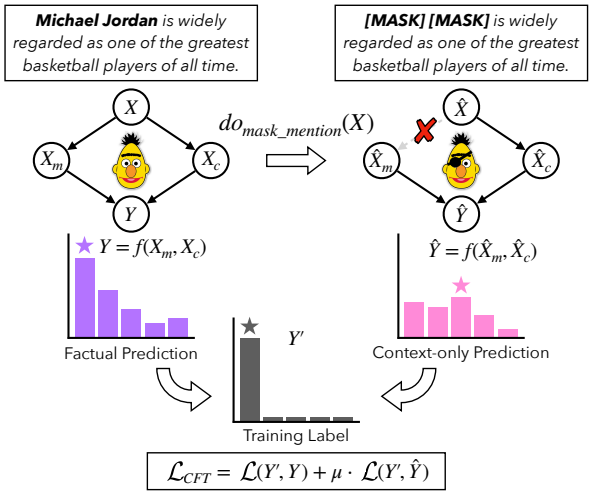

Figure 2: The system overview of the proposed method.

$$obj_{CFT} = obj + \mu \cdot \hat{obj} \tag{4}$$

where μ is a hyperparameter that controls the strength of the counterfactual training objective.

3 Experimental Settings

3.1 Training Details

We follow the training procedure from Ayoola et al. (2022b) by pretraining the models on the Wikipedia dataset and finetuning them on the training set of AIDA-CoNLL (Hoffart et al., 2011). To enable direct comparisons between models and eliminate confounding variables (i.e., entity set, candidate list, and training data), we employ the identical entity set (6.2M entities), candidate list, and training data across all models. The training datasets comprise approximately 140M mention spans, covering approximately 5.3M entities. The training takes approximately 87 hours on an A100 GPU. We collect entity priors $p(e|m)$ from the training data, which will be used to determine mention spans referring

	Shadow	Top	All
AIDA	1,286 (28.8%)	3,178 (71.2%)	4,464 (100.0%)
MSNBC	82 (12.6%)	569 (87.4%)	651 (100.0%)
AQUAINT	94 (13.1%)	625 (86.9%)	719 (100.0%)
ACE2004	46 (18.2%)	207 (81.8%)	253 (100.0%)
CWEB	3,428 (31.1%)	7,607 (68.9%)	11,035 (100.0%)
WIKI	2,261 (33.4%)	4,501 (66.6%)	6,762 (100.0%)

Table 1: Statistics of the evaluation splits report the number of In-KB mention spans for each dataset.

to overshadowed and common entities for the evaluation sets. We use the validation set of the AIDA-CoNLL dataset to tune the hyperparameter μ (see Appendix A.1). We trained each model three times and report the average performance. We explain the implementation detail in Appendix A.2.

3.2 Baseline and Competitive Methods

We compare the effectiveness of CFT with the following baseline and competitive methods: (i) **ReFinED** (Ayoola et al., 2022b) is a state-of-the-art ED method that combines entity description, entity type, and entity priors to disambiguate entities. We trained the ReFinED model using the original code released by the authors. We also report the performance of the ReFinED without entity priors (ReFinED w/o PEM) as a simple baseline method. (ii) **KBED** (Ayoola et al., 2022a) is an enhancement over ReFinED by employing entity relations to improve the performance against overshadowed entities. Since the source code of KBED is unavailable, we reimplement this to the best of our knowledge. (iii) **Focal** (Lin et al., 2017) is a typical instance weighting technique for handling class imbalance problems. The method assigns high weights for hard instances and small weights for easy instances. We implement this method on the ReFinED to focus more on uncommon entities and reduce bias towards common entities. (iv) **Entity Masking (EM)** is used in an entity relation extraction task to prevent the model from memorizing the surface forms of entities and improve generalization of the model (Zhang et al., 2017; Liu et al., 2022). We implement this method on top of the ReFinED by masking the mention surface tokens with [MASK] tokens during training and inference.

3.3 Datasets and Evaluations

To assess the effectiveness of CFT, we evaluate its performance in handling overshadowed and common entities using the following ED benchmarks.

Standard ED Benchmark. This is a commonly used benchmark for evaluating ED performance. The benchmark consists of six datasets: AIDA-CoNLL (Hoffart et al., 2011), MSNBC (Cucerzan, 2007), AQUAINT (Milne and Witten, 2008), ACE2004 (Ratinov et al., 2011), WNED-CWED (CWED) (Gabrilovich et al., 2013), and WNED-WIKI (WIKI) (Alani et al., 2018). For each dataset, we extract mention spans that refer to overshadowed and common entities using the entity priors obtained from the training data. Specifically, any mention span that cannot be resolved using the priors is considered an overshadowed entity, and one that can be resolved using the priors is regarded as a common entity. We denote mention spans that refer to overshadowed entities as "Shadow" and the common entities as "Top". The statistics of Shadow and Top mention spans for each dataset are reported in Table 1.

Shadowlink Benchmark. The Shadowlink benchmark (Provatorova et al., 2021) is designed to evaluate ED performance against overshadowed and common entities. This benchmark presents a greater challenge compared to the standard benchmark due to the brief input length. It provides three subsets for overshadowed entities (Shadow), common entities (Top), and long-tail entities that are unambiguous (Tail). The benchmark contains 890, 888, and 896 *InKB* mention spans for Shadow, Top, and Tail datasets, respectively. For this benchmark, we report the results of models trained only on Wikipedia, as fine-tuning on AIDA reduces the performance of all methods.

4 Experimental Results

4.1 Standard ED Benchmark

In Table 2, the results show that the CFT can improve the performance of the ReFinED model against overshadowed entities (Shadow) in all datasets by a significant margin of 2.4 F1 on average, outperforming the best existing method (KBED) in four out of six datasets (AIDA-CoNLL, AQUAINT, CWEB, and WIKI) by an average significant margin of 0.8 F1. Our findings in Appendix A.3 indicate that KBED can assist the ReFinED model in handling overshadowed entities when related entities are present in the input texts but encounter difficulties when such entities are unavailable. Conversely, CFT can assist the ReFinED model in handling overshadowed entities even in the absence of related entities. Regarding the performance of common entities (Top), we observe

Method	AIDA			MSNBC			AQUAINT			ACE2004			CWEB			WIKI			Avg.		
	Sha	Top	All	Sha	Top	All	Sha	Top	All	Sha	Top	All	Sha	Top	All	Sha	Top	All	Sha	Top	All
ReFinED	79.4	98.3	92.9	73.4	96.4	93.6	45.8	94.2	88.6	54.1	98.1	91.4	50.5	**90.3**	78.4	63.9	97.7	86.8	61.2	95.8	88.6
w/o PEM	80.3	97.8	92.8	73.8	95.4	92.7	46.1	91.6	86.4	55.5	97.5	91.2	51.0	88.3	77.2	65.2	96.2	86.2	62.0	94.5	87.8
w/ KBED	82.2	**98.4**	93.8	**76.0**	**96.9**	**94.3**	45.8	**95.3**	**89.6**	**57.4**	**98.3**	**92.1**	50.2	90.2	78.1	65.0	97.6	87.0	62.8	**96.1**	89.1
w/ Focal	81.6	98.3	93.5	73.2	96.1	93.3	43.8	94.6	88.8	54.1	97.9	91.2	49.7	90.2	78.1	60.7	97.2	85.4	60.5	95.7	88.4
w/ EM	70.3	97.6	89.9	72.2	93.8	91.2	42.3	89.1	83.8	45.3	94.6	87.0	43.6	87.6	74.4	57.2	96.0	83.5	55.2	93.1	85.0
w/ **CFT**	**83.8**	98.2	**94.1**	74.2	96.3	93.5	**49.0**	94.7	89.4	56.8	97.9	91.7	**51.5**	**90.3**	**78.7**	**66.2**	**97.8**	**87.6**	**63.6**	95.9	**89.2**

Table 2: Experimental results (InKB micro F1-Score) of all methods on the standard ED benchmark. For each dataset, we report results for common entities (Top), overshadowed entities (Sha), and all entities (All). The best-performing results are in **bold**.

a marginal performance gain by 0.1 F1 for CFT and 0.3 F1 for KBED. The results of removing entity priors (ReFinED without PEM) suggest that entity priors play a crucial role in the performance of ReFinED on common entities (Top). Removing the entity priors can improve the performance of ReFinED on overshadowed entities. Still, the trade-off in performance for common entities will exceed the benefit, making it an impractical method. On the other hand, we observe that the Focal and EM methods, which have been used in the relation extraction task for mitigating mention surface biases, struggle to perform on the ED task. For the EM method, we suggest that it is because of the collision of entities discussed in Section 2.

4.2 Shadowlink Benchmark

Experimental results in Table 3 show that CFT can improve the ReFinED performance on overshadowed entities (Shadow) by a margin of 1.5 F1, slightly outperforming the second-best method (ReFinED without PEM) without losing performance on long-tail entities (Tail). The KBED, which relies on related entities in input texts, has difficulty performing in this benchmark due to the lack of related entities (Shadowlink only provides one entity annotation per example). We observe the trade-off in performance for all methods on the Top dataset due to some input texts containing ambiguous contexts. For instance, one of the test examples "*The family first settled in Fox Lake, later moved to nearby Wilmette.*" provides insufficient information to disambiguate the entity "*Fox Lake*".

4.3 Inference Speed

In Table 4, we report the time taken to run inference on the test set of AIDA-CoNLL using the A100 GPU, alongside the average standard ED benchmark performance. The computational overhead of the KBED for computing the relation scores resulted in a 6.7 times slower runtime for the ReFinED with KBED in exchange for the performance improvement of 0.5 F1. On the other hand, CFT provides a comparable performance gain without any computational overhead during inference.

Method	Shadow	Top	Tail	Avg.
ReFinED	48.4	**63.9**	98.3	**70.2**
w/o PEM	49.7	62.7	97.4	69.9
w/ KBED	48.1	60.7	**98.4**	69.1
w/ Focal	47.0	63.3	98.3	69.5
w/ EM	47.9	63.4	**98.4**	69.9
w/ **CFT**	**49.8**	62.1	98.3	70.1

Table 3: Experimental results (InKB micro F1-Score) of all methods on the Shadowlink benchmark.

Method	Time taken (s)	Avg. F1-Score
ReFinED	**62**	88.6
w/ KBED	413	89.1
w/ **CFT**	**62**	**89.2**

Table 4: Time taken in seconds to perform inference on AIDA-CoNLL test dataset using the A100 GPU.

5 Conclusion

This paper aims to address the challenge in *Entity Disambiguation (ED)* for handling overshadowed entities. By formulating the ED problem as a causal graph, the spurious correlation between mention surfaces and training labels can be mitigated via counterfactual training, which helps debias the model from the overreliance on surface forms for common entities. As opposed to the current approach (KBED), which depends on the extraction of entity relations, our solution does not impose additional compute requirements for inference, making it 6.7 times faster. The empirical results show that our method performs best against overshadowed entities. These results support the new research direction of modeling the entity disambiguation problem with counterfactual learning.

6 Limitations

The limitations of our work are as follows.

- The scope of experiments in this paper does not cover the performance of downstream tasks. Further studies are needed to assess the effect of our method on tasks that rely on ED, e.g., knowledge-graph question answering (KGQA).
- Although our approach does not incur any computational overhead during inference, it incurs a computational overhead during training which is equivalent to performing two forward passes per input. Consequently, this approach might not be appropriate for larger models such as LLMs.

References

Harith Alani, Zhaochen Guo, and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semant. Web*, 9(4):459–479.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. 2022a. Improving entity disambiguation by reasoning over a knowledge base. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912, Seattle, United States. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022b. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Highly parallel autoregressive entity linking with discriminative correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7662–7669, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christina Du, Kashyap Popat, Louis Martin, and Fabio Petroni. 2022. Entity tagging: Extracting entities in text without mention supervision.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. 2022. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9804–9817, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Lihui Liu, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. 2022. Joint knowledge graph completion and question answering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1098–1108, New York, NY, USA. Association for Computing Machinery.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 509–518, New York, NY, USA. Association for Computing Machinery.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8576–8583.

Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. Entity disambiguation with entity definitions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1297–1303, Dubrovnik, Croatia. Association for Computational Linguistics.

Vera Provatorova, Samarth Bhargav, Svitlana Vakulenko, and Evangelos Kanoulas. 2021. Robustness evaluation of entity disambiguation using prior probes: the case of entity overshadowing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10501–10510, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Amir Saffari, Armin Oliya, Priyanka Sen, and Tom Ayoola. 2021. End-to-end entity resolution and question answering using differentiable knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4193–4200, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saurabh Srivastava, Mayur Patidar, Sudip Chowdhury, Puneet Agarwal, Indrajit Bhattacharya, and Gautam Shroff. 2021. Complex question answering on knowledge graphs using machine translation and multi-task learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3428–3439, Online. Association for Computational Linguistics.

Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. Named Entity Recognition for Entity Linking: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2021. Retrieval, re-ranking and multi-task learning for knowledge-base question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 347–357, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with BERT. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3264–3271, Seattle, United States. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameter details

To train our model (ReFinED with CFT), we trained the model using the hyperparameters setting in Table 5 following the original ReFinED setting.

Hyperparameter	Value
learning rate	3e-5
batch size	56
max sequence length	300
dropout	0.05
description embeddings dim.	300
# training steps	1M
# candidates	30
# entity types	1400
mention transformer init.	roberta-base
# mention encoder layers	12
description transformer init.	roberta-base
# description encoder layers	2
# description tokens	32
mention mask prob.	0.0
$(\lambda_2, \lambda_3, \lambda_4)$	(1, 0.01, 1)
μ	0.1

Table 5: Our model hyperparameters.

We performed a hyperparameter search for μ in a range of [0.1, 0.2, 0.3, 0.4] on the validation set of AIDA-CoNLL, we get the best value of 0.1. We reduced the *batch size* from 64 to 56 due to the additional memory requirement of CFT during the training. Since this paper focuses on entity disambiguation, we omit the mention detection module. The model has approximately 154M parameters.

A.2 Implementation Details of Our Method

In this subsection, we explain how we implement our method over the state-of-the-art instance-based ED method, ReFinED. The ReFinED model predicts entities's scores based on the descriptions, types, and priors of the entities. The model comprises three sub-modules:

- **Entity description module** calculates the description score for each entity by computing the dot product between the two embeddings of mention and description of the entity obtained from the knowledge base. The module is trained using a cross-entropy loss $\mathcal{L}_d$.
- **Entity typing module** predicts types probability distribution for each mention and then calculates the typing score by computing the Euclidean distance between the predicted types and entity types obtained from the knowledge base. The module is trained using a binary cross-entropy loss $\mathcal{L}_t$.
- **Combined score module** uses a linear layer to aggregate the description score, typing score, and entity prior to a final prediction score. The module is trained using a cross-entropy loss $\mathcal{L}_c$. Note that the inputs to this module, description score and typing score, are detached. Thus, the update gradients from $\mathcal{L}_c$ will not affect other parts of the model.

During training, we employ CFT on the Entity description module. Specifically, we replace the training objective of the Entity description module with obj_{CFT} (Eq. 4) where $\mathcal{L} = \mathcal{L}_d$.

A.3 Case Study and Analysis

To comprehend how our debiasing method (CFT) can improve the base model (ReFinED) and outperform the best current method (KBED) on overshadowed entities, we analyze the predictions of our method on various scenarios of overshadowed entities compared with other methods. In Table 7, example 1 illustrates the situation when an overshadowed entity appears in a text with related entities, e.g., "Guardian newspaper" (Nigerian independent daily newspaper) is related to "Lagos" (Largest city in Nigeria). In contrast, example 2 demonstrates the situation when an overshadowed entity appears in a text without related entities. The findings demonstrate that the ReFinED model has trouble handling overshadowed entities because it favors predicting popular entities with respect to entity priors (PEM). The KBED can assist the ReFinED model in handling overshadowed entities when related entities are present in the input text (example 1) but struggle when related entities are absent (example 2). The experimental results in Table 6 confirm that our debiasing method exhibits superior performance compared to the KBED in enhancing the ReFinED model against overshadowed entities, regardless of the presence (Coherence) and absence (Incoherence) of the related entities.

Method	Coherence	Incoherence
ReFinED	63.7	49.7
w/ KBED	<u>64.5</u>	<u>49.9</u>
w/ **CFT**	**66.0**	**50.8**

Table 6: Experimental results (InKB micro F1-Score) of our debiasing method on overshadowed entities with and without related entities (Coherence and Incoherence).

No.	Example	Prediction
1	*... An Air Afrique Boeing-727 jet was the third passenger liner looted in the past month by armed robbers while awaiting takeoff at Nigeria's largest international airport, the Lagos Guardian newspaper reported on Thursday. The thieves broke into the aircraft's luggage compartment and escaped with a large quantity of baggage as the plane was awaiting permission to take off ...*	PEM → Q11148 × ReFinED → Q11148 × w/ KBED → Q7738431 ✓ w/ **CFT** → Q7738431 ✓
Remark: *Q7738431 (Nigerian independent daily newspaper), Q11148 (British national daily newspaper)		
2	*... Word of the agreement leaked out when former captain Courtney Walsh, head of the West Indies players association, told the Caribbean News Agency that Lara and Hooper had been reinstated and the tour was going ahead. The crisis came to a head last Wednesday when the West Indies Cricket Board fired superstar batsman Lara as captain and Hooper as vice-captain. The two ...*	PEM → Q669037 × ReFinED → Q920396 × w/ KBED → Q920396 × w/ **CFT** → Q912881 ✓
Remark: *Q912881 (West Indies cricket team), Q669037 (West Indies) Q920396 (British West Indies)		
3	*... Saban was introduced as Alabama's coach on Thursday, touting his championship aspirations and citing his love of college football as a reason for taking a pay cut to leave the Miami Dolphins. Alabama has had four losing seasons since '97. "His teams always play with confidence and pride and I know that in order to win a national championship, a team has to be mentally as well as ...*	PEM → Q173 × ReFinED → Q4705216 × w/ KBED → Q4705216 × w/ **CFT** → Q4705216 ×
Remark: *Q492318 (University of Alabama), Q4705216 (Alabama Crimson Tide football), Q173 (State of the United States of America)		
4	*... Iran will protest to the International Court of Justice at the Hague and other global bodies about the U.S.-funded Radio Free Europe, the Iran Daily reported Monday. It quoted Foreign Minister Kamal Kharrazi as saying the radio "was set up to interfere in Iran's internal affairs" It did not say when the complaints will be filed. The English-language daily also did not say ...*	PEM → Q7330070 ✓ ReFinED → Q7330070 ✓ w/ KBED → Q2565708 × w/ **CFT** → Q2565708 ×
Remark: *Q7330070 (Foreign minister), Q2565708 (Ministry of Foreign Affairs of Iran)		

Table 7: Case studies for our debiasing method on ED datasets. We highlight the target entity and related entities with underline and double underline respectively. * indicates the gold entity label.

Finally, we examine fail cases of our method for overshadowed and non-overshadowed entities compared with other methods. In Table 7, example 3 demonstrates the case when all methods fail to resolve an overshadowed entity. Interestingly, they predict entities that suit the context well and are semantically similar to the gold entity label. Example 4 demonstrates a fail case when our method and KBED fail to resolve a non-overshadowed entity in a context containing entities that related to incorrect entities, e.g., "Foreign minister" (Ministry of Foreign Affairs of Iran) is related to "Iran" (Country in Western Asia). Our method and KBED predict an incorrect entity that suits the context.