# Efficient Overshadowed Entity Disambiguation by Mitigating Shortcut Learning

**Panuthep Tasawong[♡], Peerat Limkonchotiwat[♠], Potsawee Manakul[♣]**
**Can Udomcharoenchaikit[♡], Ekapol Chuangsuwanich[◇], Sarana Nutanong[♡]**
[♡]School of Information Science and Technology, VISTEC, Thailand
[♠]AI Singapore, Singapore [♣]SCB 10X, Thailand
[◇]Department of Computer Engineering, Chulalongkorn University, Thailand
panuthep.t_s20@vistec.ac.th, peerat@aisingapore.org

## Abstract

Entity disambiguation (ED) is crucial in natural language processing (NLP) for tasks such as question-answering and information extraction. A major challenge in ED is handling overshadowed entities—uncommon entities sharing mention surfaces with common entities. The current approach to enhance performance on these entities involves reasoning over facts in a knowledge base (KB), increasing computational overhead during inference. We argue that the ED performance on overshadowed entities can be enhanced during training by addressing shortcut learning, which does not add computational overhead at inference. We propose a simple yet effective debiasing technique to prevent models from shortcut learning during training. Experiments on a range of ED datasets show that our method achieves state-of-the-art performance without compromising inference speed. Our findings suggest a new research direction for improving entity disambiguation via shortcut learning mitigation. The code is available at https://github.com/panuthept/EfficientOvershadowedED

## 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 KB, e.g., Wikidata, for each ambiguous entity mention in a given text, which is a crucial capability when performing entity linking (EL). In real-world ED applications, there are two important properties:

[♠]Work was conducted while Peerat Limkonchotiwat was a PhD candidate at VISTEC.

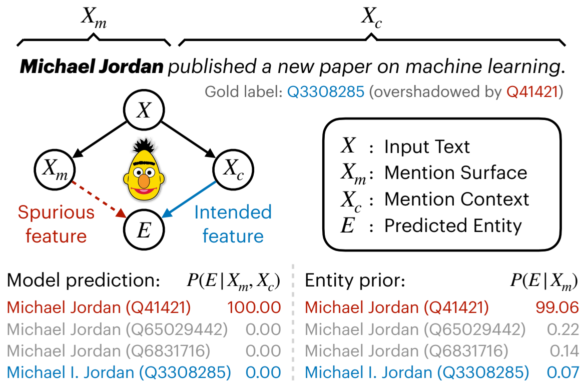


Figure 1: The causal graph of ED models. Due to the strong correlations between the spurious feature and training labels, typical ED models are prone to shortcut learning and fail to resolve overshadowed entities.

- ***Context-awareness***: The method should be able to accurately resolve entities based on the surrounding context of the entity mentions. For example, the mention of *Michael Jordan* can refer to a basketball player (Michael Jeffrey Jordan) or a machine learning researcher (Michael Irwin Jordan), depending on the context.
- ***Scalability***: The method should be capable of handling large amounts of input data efficiently. This leads to faster processing times and lower costs associated with running the ED system.

The existing ED approaches can be categorized into three main categories: (i) Classification-based approaches: These methods involve fine-tuning a classification layer on top of a pre-trained language model (PLM) to predict a score distribution over entity vocabulary (Broscheit, 2019; Yamada et al., 2022) or entity types (Onoe and Durrett, 2020; Tedeschi et al., 2021). (ii) Generative-based approaches: These methods focus on fine-tuning a generative PLM to generate a unique entity name (Cao et al., 2021; De Cao et al., 2021; Du et al., 2022) or entity description (Procopio et al., 2023). (iii) Retrieval-based approaches: These approaches consist of fine-tuning a bi-encoder (Li et al., 2020) or a cross-encoder (Wu et al., 2020) to



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15313–15321
November 12-16, 2024 ©2024 Association for Computational Linguistics

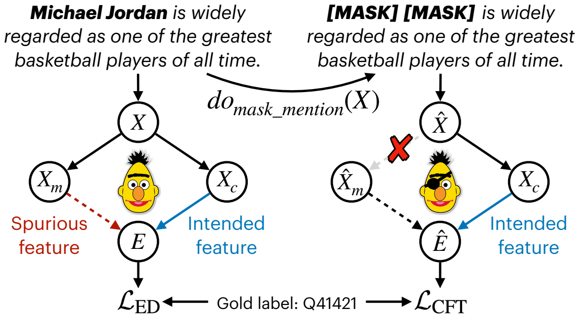


Figure 2: The system overview of the proposed method.

compute similarity scores between mentions and entity descriptions. ReFinED (Ayoola et al., 2022b) enhanced the bi-encoder's performance by incorporating entity type classification and entity priors to re-rank the bi-encoder predictions. Nonetheless, ED methods often struggle with overshadowed entities (Provatorova et al., 2021), indicating a lack of *Context-awareness* in current ED methods. KBED (Ayoola et al., 2022a) improved ReFinED's performance on overshadowed entities by leveraging KB facts. Specifically, they extract relations between every pair of mentions in input and perform reasoning over external knowledge retrieved from KB to re-rank the ReFinED's predictions. Although this method has the potential to enhance *Context-awareness* and reduce the overshadowing problem, it requires input to contain multiple mentions, and its computational burden grows as the number of mentions increases, hence compromising the *Scalability* of the ReFinED method. According to our empirical results, KBED slows down the throughput of ReFinED from 3.3 to 0.6 queries per second ($Q/s$) on standard ED datasets.

This paper tackles the overshadowing issue by addressing *shortcut learning* (Geirhos et al., 2020) during training, which does not impose a computational burden at inference. We introduce *Counterfactual Training (CFT)* as a technique to prevent the models from learning shortcut solutions and to enhance *Context-awareness*. As shown in Figure 1, each input text $X$ to ED models contains two input features: the mention surface $X_m$ (spurious feature) and the mention context $X_c$ (intended feature). The intended solution is to use the contextual feature $X_c$ to determine entity $E$. Nevertheless, the strong correlations between the spurious feature $X_m$ and training labels can induce the models to learn a shortcut (i.e., using the mention surface to determine entity $E$), obscuring the intended solution. This shortcut solution allows the models to achieve high performance on common entities but poor performance on overshadowed entities.

We assess CFT against existing methods on six standard datasets and three challenging datasets. The results show that CFT achieves the best performance on seven out of nine datasets for overshadowed entities and six out of nine datasets for overall entities without compromising the throughput at inference. We find that CFT performs surprisingly well on texts with limited contextual information (i.e., short sentences with a small number of mentions) while other methods struggle.

## 2 Counterfactual Training (CFT)

### 2.1 Counterfactual Example

For every training example $X$, we perform an intervention $do_{mask\_mention}(\cdot)$ to mask all mention surface tokens $X_m$ with special [MASK] tokens and leave the mention context tokens $X_c$ as original:

$$\hat{X} = do_{mask\_mention}(X) = \langle w_1, w_2, ..., w_n \rangle$$
$$\forall w_i \in X, \begin{cases} w_i \leftarrow \text{[MASK]} & \text{if } w_i \in X_m \\ w_i \leftarrow w_i & \text{if } w_i \in X_c \end{cases} \quad (1)$$

thereby creating a counterfactual example $\hat{X}$ that excludes the mention surface $X_m$ (spurious feature) and only contains the mention context $X_c$ (intended feature) as shown in Figure 2. We denote the masked tokens in $\hat{X}$ as $\hat{X}_m$.

### 2.2 Training Objective

The typical training objective of ED is to minimize the negative log-likelihood between the gold entity label $\tilde{E}$ and the model prediction $E$ given a mention surface $X_m$ and mention context $X_c$:

$$\mathcal{L}_{\text{ED}} = \mathcal{L}(\tilde{E}, E)$$
$$E = f(X_m, X_c, \theta) \quad (2)$$

where $\mathcal{L}$ is any loss function (e.g., cross-entropy) and $\theta$ is parameters of the model $f$. However, due to a strong correlation between mention surface $X_m$ (spurious feature) and training labels $\tilde{E}$, training the model merely on $\mathcal{L}_{\text{ED}}$ could mislead the model to use the mention surface $X_m$ (spurious feature) to resolve entities during inference.

To enforce the model to rely on contextual information, enhancing *Context-awareness*, we incorporate the counterfactual example $\hat{X}$ in Section 2.1 to provide regularization during the training process:

$$\mathcal{L}_{\text{CFT}} = \mathcal{L}(\tilde{E}, \hat{E})$$
$$\hat{E} = f(\hat{X}_m, X_c, \theta) \quad (3)$$

We combine the $\mathcal{L}_{\text{CFT}}$ auxiliary term with the $\mathcal{L}_{\text{ED}}$ to obtain the final training objective:

$$\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{ED}} + \mu \cdot \mathcal{L}_{\text{CFT}} \quad (4)$$

where $\mu$ is a hyperparameter that controls the strength of the regularization.

# 3 Experimental Settings

## 3.1 Baselines and Competitive Methods

We report the performance of three baseline ED methods. **ReFinED** (Ayoola et al., 2022b) and **BLINK** (Wu et al., 2020) are retrieval-based ED methods that use the bi-encoder and cross-encoder architectures, respectively. **GENRE** (Cao et al., 2021) is a generative encoder-decoder ED method. We use the same candidate generation method for all baselines as previous works (Ayoola et al., 2022b; Cao et al., 2021; Le and Titov, 2018).

We compare CFT with the current state-of-the-art method for improving overshadowed entity disambiguation. **KBED** (Ayoola et al., 2022a) is a ReFinED extension with overshadowed entity disambiguation improvement. The method applies reasoning over KB facts to promote candidate entities that are coherent with entities in the context.

Since we formulate the overshadowing problem as shortcut learning, we also compare our work with existing shortcut mitigation methods. **Focal loss (Focal)** (Lin et al., 2017) and **Counterfactual inference (CFI)** (Wang et al., 2022; Qian et al., 2021) are well-known debiasing techniques for mitigating shortcut learning in computer vision and NLP. We applied these two methods to the ED problem by treating the mention surface as a spurious feature. **Entity Masking (EM)** is a technique used in Relation Extraction (RE) literature (Zhang et al., 2017; Liu et al., 2022) to prevent the model from using the mention surface feature as a shortcut for predicting relations. To the best of our knowledge, this work is the first to evaluate these three methods in entity disambiguation. See the implementation details in Appendix A.1.

## 3.2 Training Details

While CFT can be applied to any existing ED method, we employ a publicly available ED method called ReFinED (Ayoola et al., 2022b) due to its practicality in resolving entities at scales. ReFinED also forms the basis of the current state-of-the-art method, KBED, allowing for direct comparison between KBED and CFT. We trained CFT, KBED, Focal, and EM based on ReFinED by pretraining on the Wikipedia dataset and finetuning on the training set of AIDA-CoNLL (Hoffart et al., 2011). The training datasets comprise approximately 140M mention spans, covering approximately 5.3M entities. We use the validation set of the AIDA-CoNLL dataset to tune hyperparameters (Appendix A.2). We trained each method using three different seeds. We report here that we cannot reproduce the original ReFinED results using their source code. [1]

## 3.3 Datasets and Evaluations

We evaluate the effectiveness of CFT on overshadowed and common entities under two scenarios.

**Standard Set.** We employ commonly used six datasets for evaluating ED performance: 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). These datasets contain lengthy texts collected from news and web articles across several domains, such as sports, politics, and technology. The average sequence length of these datasets is 565.9, with each sequence having an average of 24.5 mention spans.

**Challenge Set.** Let us now assess the ED method with limited contextual information. We employ three test datasets: TWEEKI (Harandizadeh and Singh, 2020), MINTAKA (Sen et al., 2022), and ShadowLink (SLINK) (Provatorova et al., 2021). The datasets contain short sentences from a variety of domains, including social media, question answering, and text snippets from Wikipedia pages. The average sequence length is 17.9, with each sequence having an average of 1.3 mention spans.

For each dataset, we split mention spans into "Sha" and "Top" for overshadowed and common entities using entity prior obtained from training data. Specifically, any mention span unresolvable using the prior is considered an overshadowed entity; otherwise, it is a common entity. The statistics of each dataset are reported in Appendix A.3.

**Evaluation.** We report average *InKB* micro-F1 over three different seeds for each method. We measure the inference rate ($Q/s$) on one V100 32GB GPU. We exclude "Sha" and "Top" results from BLINK and GENRE because each baseline

[1] https://github.com/amazon-science/ReFinED. We noticed that the original ReFinED model is trained using a different implementation from the source code provided, as the number of parameters is inconsistent with the model in the code.

| Method | AIDA Sha | AIDA Top | AIDA All | MSNBC* Sha | MSNBC* Top | MSNBC* All | AQUAINT* Sha | AQUAINT* Top | AQUAINT* All | ACE2004* Sha | ACE2004* Top | ACE2004* All | CWEB* Sha | CWEB* Top | CWEB* All | WIKI* Sha | WIKI* Top | WIKI* All | Avg. Sha | Avg. Top | Avg. All | Rate ($Q/s$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLINK | - | - | 86.7 | - | - | 90.3 | - | - | 88.9 | - | - | 88.7 | - | - | **82.6** | - | - | 86.1 | - | - | 87.2 | 0.1 |
| GENRE | - | - | 93.3 | - | - | **94.3** | - | - | **89.9** | - | - | 90.1 | - | - | 77.3 | - | - | 87.4 | - | - | 88.7 | 0.4 |
| 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 | **3.3** |
| 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 | **3.3** |
| w/ EM | 70.2 | 97.7 | 89.9 | 72.6 | 95.1 | 92.3 | 42.7 | 90.8 | 85.3 | 47.3 | 95.9 | 88.5 | 43.5 | 88.3 | 74.7 | 57.5 | 96.5 | 83.9 | 55.6 | 94.0 | 85.8 | **3.3** |
| w/ CFI | 80.5 | 98.1 | 93.1 | 72.7 | 96.6 | 93.6 | 46.3 | 93.7 | 88.3 | 56.1 | 98.1 | 91.7 | 50.3 | 90.1 | 78.1 | 65.3 | 97.5 | 87.1 | 61.9 | 95.7 | 88.6 | 3.1 |
| 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 | 0.6 |
| 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** | **3.3** |

Table 1: Experimental (InKB micro F1-Score) results on standard datasets with abundant contextual information. We report results for overshadowed entities (Sha), common entities (Top), and all entities (All). **Bold** and underline represent the best and second-performing, respectively. (†) indicates a statistically significant improvement measured using the Almost Stochastic Dominance test (Ulmer et al., 2022) with a significant level of alpha = 0.05. (*) denotes out-of-domain datasets. We used the original parameters for BLINK and GENRE.

is trained on a different dataset and possesses a different entity prior, making results incomparable to those of ReFinED-based.

## 4 Experimental Results

**Standard Set.** The results in Table 1 demonstrate the effectiveness and efficiency of our method (CFT) on texts with abundant context. CFT outperforms the state-of-the-art method (KBED) on overshadowed entity disambiguation by a significant margin. CFT also performs the best compared to other debiasing methods. Focal performs well only on the in-domain dataset (AIDA) but struggles to perform on out-of-domain datasets. Although EM and CFI are widely used in RE to mitigate shortcut learning, it is ineffective in ED. For the $Q/s$ rate, Focal, EM, and CFT achieve the same throughput as ReFinED, while CFI and KBED show a drop in throughput. The case study and analysis of CFT and KBED are discussed in Section 5.

**Challenge Set.** Table 2 shows that CFT is the most effective method for disambiguating entities on out-of-domain datasets with limited contextual information (TWEEKI and MINTAKA). BLINK performs well only on the Wikipedia domain dataset (SLINK). Although KBED performs well on input texts with abundant context, it struggles when context is limited. The results of the $Q/s$ rates conform with those of the standard set.

**Scalability.** Figure 3 displays a bar chart with the average inference time per query on the y-axis. The x-axis organizes the queries into eight octiles ranked according to the number of mentions per query, where queries in the eighth octile have the highest number of mentions. We can see that the performance gap between CFT and KBED widens as we move from the first to the eighth octile. This finding shows that not only is CFT faster, but it can also scale better than KBED as the number of mentions per query grows. The statistics of each octile are reported in Appendix A.4

| Method | TWEEKI* Sha | TWEEKI* Top | TWEEKI* All | MINTAKA* Sha | MINTAKA* Top | MINTAKA* All | SLINK Sha | SLINK Top | SLINK All | Rate ($Q/s$) |
|---|---|---|---|---|---|---|---|---|---|---|
| BLINK | - | - | 80.5 | - | - | 85.1 | - | - | **74.6** | 0.4 |
| GENRE | - | - | 79.8 | - | - | 84.2 | - | - | 56.5 | 15.7 |
| ReFinED | 42.1 | **93.5** | 82.1 | 37.3 | 95.9 | 87.1 | 43.0 | 93.0 | 69.2 | **39.0** |
| w/ Focal | 42.0 | 93.1 | 81.8 | 35.7 | 95.7 | 86.7 | 41.8 | 93.0 | 68.8 | **39.0** |
| w/ EM | 32.3 | 90.1 | 77.3 | 27.9 | 91.9 | 82.3 | 43.1 | 91.7 | 68.0 | **39.0** |
| w/ CFI | 42.6 | 93.3 | 81.9 | 38.3 | 95.8 | 87.1 | 43.5 | **93.1** | 69.2 | 24.3 |
| w/ KBED | 40.9 | 92.8 | 81.2 | 37.1 | 95.5 | 86.6 | 41.5 | 93.0 | 68.1 | 27.5 |
| w/ **CFT** | **44.6†** | **93.5** | **82.6†** | **38.7** | **96.0** | **87.3** | **44.1†** | 92.8 | 69.5 | **39.0** |

Table 2: Results on challenge datasets with limited contextual information. (*) denotes out-of-domain datasets.

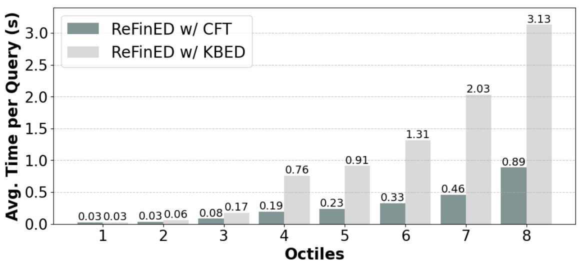


Figure 3: Time taken to process queries with different numbers of mentions. The queries are organized into eight octiles ranked by the number of mentions.

## 5 Qualitative Analysis

In Table 3, we examine the success and failure cases of CFT in comparison to KBED.

Success cases 1 and 2 demonstrate scenarios in which overshadowed entities appear in texts, both with and without relevant entities (entities

| |
|---|
| ***Success Case 1:*** *... An Air Afrique Boeing-727 jet was the third passenger liner looted in the past month by armed robbers while awaiting takeoff robbers while awaiting takeoff at Nigeria's largest international airport, the* <u>*Lagos*</u> *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 ...*<br>**Gold label** → Q7738431 (Nigerian independent daily newspaper), **Entity Prior** → Q11148 (British national daily newspaper) ×<br>**KBED** → Q7738431 (Nigerian independent daily newspaper) ✓, **CFT** → Q7738431 (Nigerian independent daily newspaper) ✓ |
| ***Success Case 2:*** *When the flame is lit that smoke is being burned. The smoke is vaporized wax. When you blow it out, the wick is still hot enough to vaporize wax but not ignite it. If you cool the wick like lick your finger or put in water, the wick is no longer hot enough to vaporize wax.*<br>**Gold label** → Q6452502 (Vaporization), **Entity Prior** → Q132814 (Evaporation) ×<br>**KBED** → Q132814 (Evaporation) ×, **CFT** → Q6452502 (Vaporization) ✓ |
| ***Failure Case 1:*** *I absolutely love the MCU movies, but Spider-Man said it best in Civil War when he saw* <u>*Cap*</u> *throwing his shield and said, "That thing doesn't obey the laws of physics at all."*<br>**Gold label** → Q131559 (Shield), **Entity Prior** → Q131559 (Shield) ✓<br>**KBED** → Q131559 (Shield) ✓, **CFT** → Q690141 (Captain America's shield) × |
| ***Failure Case 2:*** *... also began broadcasts directed to Iraq on Friday. In a trial period of several weeks, the station will broadcast one 30 minute program a day to Iran and Iraq. The Farsi language service to Iran was approved by the Czech government in August. Radio Free Europe began transmitting from Munich, Germany, in 1951, spreading uncensored news to Soviet-controlled countries behind the ...*<br>**Gold label** → Q1155216 (politics of the Czech Republic), **Entity Prior** → Q5015587 (Government of the Czech Republic) ×<br>**KBED** → Q213 (Czech Republic) ×, **CFT** → Q213 (Czech Republic) × |

Table 3: Examples of success and failure cases of CFT. Highlight indicates the target entity. <u>Underline</u> indicates the relevant entity in the context that allows KBED to perform reasoning to resolve the target entity.

related to the target entity in the knowledge base). In both cases, CFT can accurately resolve the two overshadowed entities, regardless of the availability of the relevant entity, while KBED struggles when the relevant entity is unavailable. These examples highlight the advantages of the debiasing method and the limitations of the reasoning method for dealing with overshadowed entities.

Failure Case 1 reveals the limitations of current ED benchmarking. As shown in Table 3, both CFT and KBED make technically correct predictions (i.e., Captain America's shield and Shield). However, existing ED datasets only provide a single gold entity for each mention, leading to correct predictions that do not align with the dataset's annotation bias being classified as incorrect. Lastly, Failure Case 2 shows that both CFT and KBED are still prone to make simple mistakes, e.g., confusion between the governing body and the country. These failure cases underscore the need for continued improvement in ED datasets and models.

## 6 Conclusion

This paper addresses the challenge of handling overshadowed entities in *Entity Disambiguation (ED)*. By formulating the ED problem as shortcut learning mitigation, the spurious correlation between mention surfaces and training labels can be mitigated via CFT, which reduces the model's reliance on surface forms for common entities. As opposed to the current SOTA (KBED), our solution *does not* impose additional inference time, making it 5 times faster than KBED. The empirical results show that CFT achieves the best performance on overshadowed entities. These results support the new research direction of modeling the entity disambiguation problem with counterfactual learning.

## 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).

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780.

Bahareh Harandizadeh and Sameer Singh. 2020. Tweeki: Linking named entities on Twitter to a knowledge graph. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 222–231, Online. Association for Computational Linguistics.

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.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. 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.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. 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 5434–5445, Online. 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.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on 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.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations, ICLR 2022 ; Conference date: 25-04-2022 Through 29-04-2022.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. 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 Implementation Details

### A.1.1 Counterfactual Training

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.1.2 Counterfactual Inference

This section explains how we implement counterfactual inference (Wang et al., 2022; Qian et al., 2021) for ED. For every test example $X$, we perform an intervention $do_{\textit{mask\_context}}(\cdot)$ to mask all context tokens $X_c$ with special [MASK] tokens and leave the mention surface tokens $X_m$ as original:

$$X' = do_{\textit{mask\_context}}(X) = \langle w_1, w_2, ..., w_n \rangle$$
$$\forall w_i \in X, \begin{cases} w_i \leftarrow \text{[MASK]} & \text{if } w_i \in X_c \\ w_i \leftarrow w_i & \text{if } w_i \in X_m \end{cases} \tag{5}$$

thereby creating a counterfactual example $X'$ that excludes the mention context $X_c$ (intended

| 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) |
| $\mu$ | 0.1 |

Table 4: ReFinED with CFT hyperparameters.

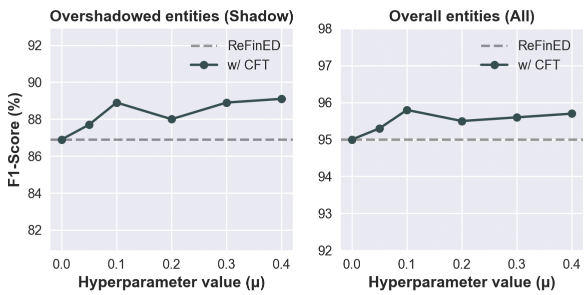


Figure 4: Results of ReFinED with CFT with different $\mu$ values on the validation set of AIDA dataset.

feature) and only contains the mention surface $X_m$ (spurious feature). We denote the masked tokens in $X'$ as $X'_c$. This counterfactual example $X'$ is then used to estimate the effect of mention surfaces $X_m$ on output predictions:

$$E' = f(X_m, X'_c, \theta) \tag{6}$$

To mitigate the effect of mention surfaces $X_m$ on output predictions, we subtract the original model prediction $E$ with the estimated effect $E'$:

$$E_{\text{final}} = E - \lambda \cdot E' \tag{7}$$

where $\lambda$ is a hyperparameter that controls the effect of the mention surfaces that we want to reduce.

## A.2 Hyperparameter details

To train our model (ReFinED with CFT), we trained the model using the hyperparameters setting in Table 4 following the original ReFinED setting.

We performed a hyperparameter search for $\mu$ in a range of [0.05, 0.1, 0.2, 0.3, 0.4] on the validation set of AIDA-CoNLL, we got the best value of 0.1 as shown in Figure 4. 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. The training took approximately 87 hours on an A100 GPU.

### A.3 Datasets statistics

Table 5 shows the InKB statistics of each test dataset. The overshadowed entities are determined using entity prior collected from the training dataset of ReFinED. The standard set contains long article ED datasets that have approximately 24.5 mentions and 564.9 words per query. The challenge set contains short sentence ED datasets that have approximately 1.3 mentions and 17.9 words per query. The standard and challenge sets have similar proportion of overshadowed entities, 30.1% and 27.4%, respectively.

| | Mentions | | Seq. Length | Shadow |
|---|---|---|---|---|
| **Dataset** | Count | Mean | Mean | % |
| **Standard Set** | | | | |
| AIDA | 4,464 | 19.4 | 177.2 | 28.8% |
| MSNBC | 651 | 32.6 | 565.9 | 12.6% |
| AQUAINT | 719 | 14.4 | 220.5 | 13.1% |
| ACE2004 | 253 | 7.2 | 375.5 | 18.2% |
| CWEB | 11,035 | 34.5 | 1,212.3 | 31.1% |
| WIKI | 6,734 | 21.1 | 269.8 | 33.5% |
| Avg. | 23,856 | 24.5 | 564.9 | 30.1% |
| **Challenge Set** | | | | |
| TWEEKI | 860 | 1.8 | 16.4 | 24.1% |
| MINTAKA | 5,703 | 1.5 | 10.1 | 17.1% |
| SLINK | 2,674 | 1.0 | 29.7 | 50.5% |
| Avg. | 9,237 | 1.3 | 17.9 | 27.4% |

Table 5: Statistics of test datasets.

### A.4 Scalability Study

Table 6 shows the statistics of each octile in Figure 3. The octiles are created by ranking queries from seven datasets: AIDA, MSNBC, AQUAINT, ACE2004, CWEB, WIKI, and TWEEKI, in ascending order according to the number of mentions in queries, then divided into eight equal-sized octiles.

| | Queries | Number of Mentions | | |
|---|---|---|---|---|
| **Octile** | Count | Min | Max | Mean ± Std. |
| 1 | 549 | 1 | 1 | 1.0 ± 0.0 |
| 2 | 549 | 1 | 2 | 1.7 ± 0.5 |
| 3 | 549 | 2 | 5 | 3.2 ± 1.0 |
| 4 | 549 | 5 | 16 | 11.4 ± 3.2 |
| 5 | 549 | 16 | 21 | 18.7 ± 1.6 |
| 6 | 549 | 21 | 27 | 23.7 ± 1.8 |
| 7 | 549 | 27 | 36 | 31.2 ± 2.6 |
| 8 | 537 | 36 | 114 | 45.1 ± 10.4 |

Table 6: Statistics of octiles.

### A.5 Masking Mentions During Inference

The proposed method (CFT) has demonstrated substantial improvement by masking mentions during training, raising questions about the impact of masking mentions during inference. To investigate this, we conducted experiments on six standard datasets using CFT with and without masked mentions during inference.

| **Method** | Sha | Top | All |
|---|---|---|---|
| CFT w/o masked mentions during inference | **63.6** | **95.9** | **89.2** |
| CFT w/ masked mentions during inference | 54.8 | 91.1 | 83.7 |

Table 7: Results of CFT on six standard datasets with and without masked mentions during inference.

As shown in Table 7, masking mentions during inference notably diminishes model performance. This finding suggests that masking mentions during inference for ED is not beneficial.