Global Entity Disambiguation with BERT

Anonymous ACL submission

Abstract

We propose a global entity disambiguation (ED) model based on BERT (Devlin et al., 2019). To capture global contextual information for ED, our model treats not only words but also entities as input tokens, and solves the task by sequentially resolving mentions to their referent entities and using resolved entities as inputs. We train the model using a large entity-annotated corpus obtained from Wikipedia. We achieve new state-of-the-art results on five standard ED datasets: AIDA-CoNLL, MSNBC, AQUAINT, ACE2004, and WNED-WIKI.

1 Introduction

Entity disambiguation (ED) refers to the task of assigning mentions in a document to corresponding entities in a knowledge base (KB). This task is challenging because of the ambiguity between mentions (e.g., *World Cup*) and the entities they refer to (e.g., `FIFA World Cup` or `Rugby World Cup`). ED models typically rely on *local* contextual information based on words that co-occur with the mention and *global* contextual information based on the entity-based coherence of the disambiguation decisions. A key to improve the performance of ED is to effectively combine both local and global contextual information (Ganea and Hofmann, 2017; Le and Titov, 2018).

In this study, we propose a global ED model based on BERT (Devlin et al., 2019). Our model treats words and entities in the document as input tokens, and is trained by predicting randomly masked entities in a large entity-annotated corpus obtained from Wikipedia. This training enables the model to learn how to disambiguate masked entities based on words and non-masked entities. At the inference time, our model disambiguates mentions sequentially using local contextual information based on words and global contextual information based on already resolved entities (see Figure 1).

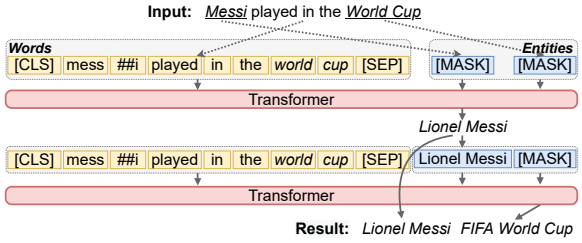

Figure 1: The inference procedure of our model with the input text "Messi played in the World Cup." Given mentions (*Messi* and *World Cup*), our model sequentially resolves them to their referent entities, and uses the resolved entities as contexts at each step.

We conduct extensive experiments using six standard ED datasets, i.e., AIDA-CoNLL, MSNBC, AQUAINT, ACE2004, WNED-WIKI, and WNED-CWEB. As a result, the global contextual information consistently improves the performance. Furthermore, we achieve new state of the art on five out of the six datasets. The source code and model checkpoint will be publicized for future research.

2 Related Work

Transformer-based ED. Several recent studies have proposed ED models based on Transformer (Vaswani et al., 2017) trained with a large entity-annotated corpus obtained from Wikipedia (Broscheit, 2019; Ling et al., 2020; Févry et al., 2020; Cao et al., 2021). Broscheit (2019) trained an ED model based on BERT by classifying each word in the document to the corresponding entity. Similarly, Févry et al. (2020) addressed ED using BERT by classifying mention spans to the corresponding entities. Ling et al. (2020) trained BERT by predicting entities using the document-level representation. Cao et al. (2021) addressed ED by training BART (Lewis et al., 2020) to generate referent entity titles of target mentions in an autoregressive manner. However, unlike our model, these models addressed the task based only on local contextual information.

Treating entities as inputs of Transformer. Recent studies (Zhang et al., 2019; Yamada et al., 2020; Sun et al., 2020) have proposed Transformer-based models that treat entities as input tokens to enrich their expressiveness using additional information contained in the entity embeddings. However, these models were designed to solve general NLP tasks and not tested on ED. We treat entities as input tokens to capture the global context that is shown to be highly effective for ED.

ED as sequential decision task. Past studies (Yang et al., 2019; Fang et al., 2019) have solved ED by casting it as a sequential decision task to capture global contextual information. We adopt a similar method with an enhanced Transformer architecture, a training task, and an inference method to implement the global ED model based on BERT.

3 Model

Given a document with N mentions, each of which has K entity candidates, our model solves ED by selecting a correct referent entity from the entity candidates for each mention.

3.1 Model Architecture

Our model is based on BERT and takes words and entities (Wikipedia entities or the `[MASK]` entity). The input representation of a word or an entity is constructed by summing the token, token type, and position embeddings (see Figure 2):

Token embedding is the embedding of the corresponding token. The matrices of the word and entity token embeddings are represented as $\mathbf{A} \in \mathbb{R}^{V_w \times H}$ and $\mathbf{B} \in \mathbb{R}^{V_e \times H}$, respectively, where H is the size of the hidden states of BERT, and V_w and V_e are the number of items in the word vocabulary and that of the entity vocabulary, respectively.

Token type embedding represents the type of token, namely word ($\mathbf{C}_{word}$) or entity ($\mathbf{C}_{entity}$).

Position embedding represents the position of the token in a word sequence. A word and an entity appearing at the i-th position in the sequence are represented as $\mathbf{D}_i$ and $\mathbf{E}_i$, respectively. If an entity mention contains multiple words, its position embedding is computed by averaging the embeddings of the corresponding positions (see Figure 2).

Following Devlin et al. (2019), we tokenize the document text using the BERT's wordpiece tokenizer, and insert `[CLS]` and `[SEP]` tokens as the first and last words, respectively.

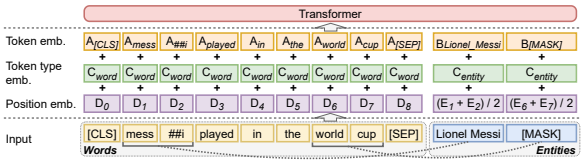

Figure 2: The input representation of our model with the text "Messi played in the World Cup" with mentions *Messi* and *World Cup*. The entity corresponding to the mention *World Cup* is replaced by the `[MASK]` token.

3.2 Training Task

Similar to the masked language model (MLM) objective adopted in BERT, our model is trained by predicting randomly masked entities. Specifically, we randomly replace some percentage of the entities with special `[MASK]` entity tokens and then trains the model to predict masked entities.

We adopt a model equivalent to the one used to predict words in MLM. Formally, we predict the original entity corresponding to a masked entity by applying softmax over all entities:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{B}\mathbf{m}_e + \mathbf{b}_o) \quad (1)$$

$$\mathbf{m}_e = \text{layernorm}\big(\text{gelu}(\mathbf{W}_f \mathbf{h}_e + \mathbf{b}_f)\big) \quad (2)$$

where $\mathbf{h}_e \in \mathbb{R}^H$ is the output embedding corresponding to the masked entity, $\mathbf{W}_f \in \mathbb{R}^{H \times H}$ is a matrix, $\mathbf{b}_o \in \mathbb{R}^{V_e}$ and $\mathbf{b}_f \in \mathbb{R}^H$ are bias vectors, $\text{gelu}(\cdot)$ is the gelu activation function (Hendrycks and Gimpel, 2016), and $\text{layernorm}(\cdot)$ is the layer normalization function (Lei Ba et al., 2016).

3.3 ED Model

Local ED Model. Our local ED model takes words and N `[MASK]` tokens corresponding to the mentions in the document. The model then computes the embedding $\mathbf{m}'_e \in \mathbb{R}^H$ for each `[MASK]` token using Eq.(2) and predicts the entity using softmax over the K entity candidates:

$$\hat{\mathbf{y}}_{ED} = \text{softmax}(\mathbf{B}^* \mathbf{m}'_e + \mathbf{b}^*_o), \quad (3)$$

where $\mathbf{B}^* \in \mathbb{R}^{K \times H}$ and $\mathbf{b}^*_o \in \mathbb{R}^K$ consist of the entity token embeddings and the bias corresponding to the entity candidates, respectively. Note that $\mathbf{B}^*$ and $\mathbf{b}^*_o$ are the subsets of $\mathbf{B}$ and $\mathbf{b}_o$, respectively.

Global ED Model. Our global ED model resolves mentions sequentially for N steps (see Algorithm 1). First, the model initializes the entity of each mention using the `[MASK]` token. Then, for each step, it predicts an entity for each `[MASK]` token, selects the prediction with the highest probability produced by the softmax function in Eq.(3), and

Algorithm 1: Algorithm of our global ED model.

Input: Words and mentions $m_1, \ldots m_N$.
Initialize: $e_i \leftarrow$ `[MASK]`, $i = 1 \ldots N$
repeat N **times**
  For all `[MASK]`s, obtain predictions using Eq.(3) with words and entities e_1, ..., e_N as inputs
  Select a mention m_j and its prediction $\hat{e}_j$ with the highest probability
  $e_j \leftarrow \hat{e}_j$
end
return $\{e_1, \ldots, e_N\}$

resolves the corresponding mention by assigning the predicted entity to it. This model is denoted as **confidence-order**. We also test a model that selects mentions according to their order of appearance in the document and denote it by **natural-order**.

3.4 Modeling Details

Our model is based on $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019). The parameters shared with BERT are initialized using BERT, and the other parameters are initialized randomly. We treat the hyperlinks in Wikipedia as entity annotations and randomly mask 30% of all entities. We train the model by maximizing the log likelihood of entity predictions. Further details are described in Appendix A.

4 Experiments

Our experimental setup follows Le and Titov (2018). In particular, we test the proposed ED models using six standard datasets: AIDA-CoNLL (CoNLL) (Hoffart et al., 2011), MSNBC, AQUAINT, ACE2004, WNED-CWEB (CWEB), and WNED-WIKI (WIKI) (Guo and Barbosa, 2018). We consider only the mentions that refer to valid entities in Wikipedia. For all datasets, we use the *KB+YAGO* entity candidates and their associated $\hat{p}(e|m)$ (Ganea and Hofmann, 2017), and use the top 30 candidates based on $\hat{p}(e|m)$. For the CoNLL dataset, we also test the performance using *PPRforNED* entity candidates (Pershina et al., 2015). We report the in-KB accuracy for the CoNLL dataset and the micro F1 score (averaged per mention) for the other datasets. Further details of the datasets are provided in Appendix C.

Furthermore, we optionally fine-tune the model by maximizing the log likelihood of the ED predictions ($\hat{\mathbf{y}}_{ED}$) using the training set of the CoNLL dataset with the KB+YAGO candidates. We mask 90% of the mentions and fix the entity token embeddings ($\mathbf{B}$ and $\mathbf{B}^*$) and the bias ($\mathbf{b}_o$ and $\mathbf{b}_o^*$). The model is trained for two epochs using AdamW.

Name	Accuracy (KB+YAGO)	Accuracy (PPRforNED)
Baselines:		
Yamada et al. (2016)	91.5	93.1
Ganea and Hofmann (2017)	92.2	-
Yang et al. (2018)	93.0	95.9
Le and Titov (2018)	93.1	-
Fang et al. (2019)	94.3	-
Yang et al. (2019)	94.6	
Broscheit (2019)	87.9	-
Ling et al. (2020)	-	94.9
Févry et al. (2020)	92.5	96.7
Cao et al. (2021)	93.3	-
Our model w/o fine-tuning:		
confidence-order	92.4	94.6
natural-order	91.7	94.0
local	90.8	94.0
Our model w/ fine-tuning:		
confidence-order	**95.0**	**97.1**
natural-order	94.8	97.0
local	94.5	96.8

Table 1: In-KB accuracy on the CoNLL dataset.

Additional details are provided in Appendix B.

4.1 Results

Table 1 and Table 2 present our experimental results. We achieve new state of the art on all datasets except the CWEB dataset by outperforming strong Transformer-based ED models, i.e, Broscheit (2019), Ling et al. (2020), Févry et al. (2020), and Cao et al. (2021). Furthermore, on the CoNLL dataset, our confidence-order model trained only on our Wikipedia-based corpus outperforms Yamada et al. (2016) and Ganea and Hofmann (2017) trained on its in-domain training set.

Our global models consistently perform better than the local model, demonstrating the effectiveness of using global contextual information even if local contextual information is captured using expressive BERT model. Moreover, the confidence-order model performs better than the natural-order model on most datasets. An analysis investigating why the confidence-order model outperforms the natural-order model is provided in the next section.

The fine-tuning on the CoNLL dataset significantly improves the performance on this dataset (Table 1). However, it generally degrades the performance on the other datasets (Table 2). This suggests that Wikipedia entity annotations are more suitable than the CoNLL dataset to train general-purpose ED models.

Additionally, our models perform worse than Yang et al. (2018) on the CWEB dataset. This is because this dataset is significantly longer on aver-

Name	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Average
Baselines:						
Ganea and Hofmann (2017)	93.7	88.5	88.5	77.9	77.5	85.2
Yang et al. (2018)	92.6	89.9	88.5	**81.8**	79.2	86.4
Le and Titov (2018)	93.9	88.3	89.9	77.5	78.0	85.5
Fang et al. (2019)	92.8	87.5	91.2	78.5	82.8	86.6
Yang et al. (2019)	93.8	88.3	90.1	75.6	78.8	85.3
Cao et al. (2021)	94.3	89.9	90.1	77.3	87.4	87.8
Our model w/o fine-tuning:						
confidence-order	**96.3**	**93.5**	**91.9**	78.9	89.1	**89.9**
natural-order	96.1	92.9	**91.9**	78.4	**89.2**	89.7
local	96.1	91.9	**91.9**	78.4	88.8	89.4
Our model w/ fine-tuning:						
confidence-order	94.1	91.5	90.7	78.3	87.6	88.4
natural-order	94.1	90.9	90.7	78.3	87.4	88.3
local	94.1	90.8	90.7	78.2	87.2	88.2

Table 2: Micro F1 score on the MSNBC, AQUAINT, ACE2004, CWEB, and WIKI datasets.

#annotations	confidence-order	natural-order	local	G&H2017
0	1.0	1.0	1.0	0.8
1–10	95.55	95.55	95.55	91.93
11–50	96.98	96.70	96.43	92.44
≥51	96.64	96.38	95.80	94.21

Table 3: Accuracy on the CoNLL dataset split by the frequency of entity annotations. Our models were fine-tuned using the CoNLL dataset. **G&H2017**: The results of Ganea and Hofmann (2017).

age than other datasets, i.e., approximately 1,700 words per document on average, which is more than three times longer than the 512-word limit that can be handled by BERT-based models including ours. Yang et al. (2018) achieved excellent performance on this dataset because their model uses various hand-engineered features capturing document-level contextual information.

4.2 Analysis

To investigate how global contextual information helps our model to improve performance, we manually analyze the difference between the predictions of the local, natural-order, and confidence-order models. We use the fine-tuned model using the CoNLL dataset with the YAGO+KB candidates. Although all models perform well on most mentions, the local model often fails to resolve mentions of common names referring to specific entities (e.g., "New York" referring to *New York Knicks*). Global models are generally better to resolve such difficult cases because of the presence of strong global contextual information (e.g., mentions referring to basketball teams).

Furthermore, we find that the confidence-order model works especially well for mentions that require a highly detailed context to resolve. For example, a mention of "Matthew Burke" can refer to two different former Australian rugby players. Although the local and natural-order models incorrectly resolve this mention to the player who has the larger number of occurrences in our Wikipedia-based corpus, the confidence-order model successfully resolves this by disambiguating its contextual mentions, including his teammates, in advance. We provide detailed inference sequence of the corresponding document in Appendix D.

4.3 Performance for Rare Entities

We examine whether our model learns effective embeddings for rare entities using the CoNLL dataset. Following Ganea and Hofmann (2017), we use the mentions of which entity candidates contain their gold entities and measure the performance by dividing the mentions based on the frequency of their entities in the Wikipedia annotations used to train the embeddings.

As presented in Table 3, our models achieve enhanced performance for rare entities. Furthermore, the global models consistently outperform the local model both for rare and frequent entities.

5 Conclusion and Future Work

We propose a new global ED model based on BERT. Our extensive experiments on a wide range of ED datasets demonstrate its effectiveness.

One limitation of our model is that, similar to existing ED models, our model cannot handle entities that are not included in the vocabulary. In our future work, we will investigate the method to compute the embeddings of such entities using a post-hoc training with an extended vocabulary (Tai et al., 2020).

References

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*, pages 677–685.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint Entity Linking with Deep Reinforcement Learning. In *The World Wide Web Conference*, pages 438–447.

Thibault Févry, Nicholas FitzGerald, Livio Baldini Soares, and Tom Kwiatkowski. 2020. Empirical Evaluation of Pretraining Strategies for Supervised Entity Linking. In *Automated Knowledge Base Construction*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629.

Zhaochen Guo and Denilson Barbosa. 2018. Robust Named Entity Disambiguation with Random Walks. *Semantic Web*, 9(4):459–479.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415v3*.

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 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

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.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450v1*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning Cross-Context Entity Representations from Text. *arXiv preprint arXiv:2001.03765v1*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized Page Rank for Named Entity Disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 238–243.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized Language and Knowledge Embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670.

Wen Tai, H T Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259.

Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning Dynamic Context Augmentation for Global Entity Linking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 271–281.

Yi Yang, Ozan Irsoy, and Kazi Shefaet Rahman. 2018. Collective Entity Disambiguation with Structured Gradient Tree Boosting. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 777–786.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Appendix for "Global Entity Disambiguation with BERT"

A Details of Proposed Model

As the input corpus for training our model, we use the December 2018 version of Wikipedia, comprising approximately 3.5 billion words and 11 million entity annotations. We generate input sequences by splitting the content of each page into sequences comprising ≤ 512 words and their entity annotations (i.e., hyperlinks). The input text is tokenized using BERT's tokenizer with its vocabulary consisting of $V_w = 30,000$ words. Similar to Ganea and Hofmann (2017), we create an entity vocabulary consisting of $V_e = 128,040$ entities, which are contained in the entity candidates in the datasets used in our experiments.

Our model consists of approximately 440 million parameters. To reduce the training time, the parameters that are shared with BERT are initialized using BERT. The other parameters are initialized randomly. The model is trained via iterations over Wikipedia pages in a random order for seven epochs. To stabilize the training, we update only those parameters that are randomly initialized (i.e., fixed the parameters initialized using BERT) at the first epoch, and update all parameters in the remaining six epochs. We implement the model using PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020), and the training takes approximately ten days using eight Tesla V100 GPUs. We optimize the model using AdamW. The hyper-parameters used in the training are detailed in Table 4.

B Details of Fine-tuning on CoNLL Dataset

The hyper-parameters used in the fine-tuning on the CoNLL dataset are detailed in Table 5. We select these hyper-parameters from the search space described in Devlin et al. (2019) based on the accuracy on the development set of the CoNLL dataset. A document is split if it is longer than 512 words, which is the maximum word length of the BERT model.

C Details of ED Datasets

The statistics of the ED datasets used in our experiments are provided in Table 6.

Name	Value
number of hidden layers	24
hidden size	1024
attention heads	16
attention head size	64
activation function	gelu
maximum word length	512
batch size	2048
learning rate (1st epoch)	5e-4
learning rate decay (1st epoch)	none
warmup steps (1st epoch)	1000
learning rate	5e-5
learning rate decay	linear
warmup steps	1000
dropout	0.1
weight decay	0.01
gradient clipping	1.0
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-6

Table 4: Hyper-parameters used for training on Wikipedia entity annotations.

Name	Value
maximum word length	512
number of epochs	2
batch size	16
learning rate	2e-5
learning rate decay	linear
warmup proportion	0.1
dropout	0.1
weight decay	0.01
gradient clipping	1.0
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-6

Table 5: Hyper-parameters during fine-tuning on the CoNLL dataset.

D Example of Inference by Confidence-order Model

Figure 3 shows an example of the inference performed by our confidence-order model fine-tuned on the CoNLL dataset. The document is obtained from the test set of the CoNLL dataset. As shown in the figure, the model starts with unambiguous player names to recognize the topic of the document, and subsequently resolves the mentions that are challenging to resolve.

Notably, the model correctly resolves the mention "Nigel Walker" to the corresponding former rugby player instead of a football player, and the mention "Matthew Burke" to the correct former

Document:

"Campo has a massive following in this country and has had the public with him ever since he first played here in 1984," said Andrew, also likely to be making his final **20:** Twickenham appearance. On tour, **17:** Australia have won all four tests against **46:** Italy, **47:** Scotland, **48:** Ireland and **45:** Wales, and scored 414 points at an average of almost 35 points a game. League duties restricted the **28:** Barbarians' selectorial options but they still boast 13 internationals including **44:** England full-back **16:** Tim Stimpson and recalled wing **22:** Tony Underwood, plus **12:** All Black forwards **25:** Ian Jones and **14:** Norm Hewitt.
Teams: **27:** Barbarians - 15 - **7:** Tim Stimpson (**31:** England); 14 - **50:** Nigel Walker (**36:** Wales), 13 - **1:** Allan Bateman (**32:** Wales), 12 - **10:** Gregor Townsend (**39:** Scotland), 11 - **4:** Tony Underwood (**34:** England); 10 - **17:** Rob Andrew (**33:** England), 9 - **2:** Rob Howley (**35:** Wales); 8 - **15:** Scott Quinnell (**37:** Wales), 7 - **8:** Neil Back (**38:** England), 6 - **19:** Dale McIntosh (**41:** Pontypridd), 5 - **24:** Ian Jones (**51:** New Zealand), 4 - **11:** Craig Quinnell (**40:** Wales), 3 - **5:** Darren Garforth (**42:** Leicester), 2 - **18:** Norm Hewitt (**52:** New Zealand), 1 - **3:** Nick Popplewell (**49:** Ireland). **43:** Australia - 15 - **53:** Matthew Burke; 14 - **9:** Joe Roff, 13 - **26:** Daniel Herbert, 12 - **20:** Tim Horan (captain), 11 - **23:** David Campese; 10 - **29:** Pat Howard, 9 - Sam Payne; 8 - Michael Brial, 7 - **30:** David Wilson, 6 - **13:** Owen Finegan, 5 - **21:** David Giffin, 4 - Tim Gavin, 3 - Andrew Blades, 2 - Marco Caputo, 1 - **6:** Dan Crowley.

Order of Inference by Confidence-order Model:

Allan Bateman ➔ Rob Howley ➔ Nick Popplewell ➔ Tony Underwood ➔ Darren Garforth ➔ Dan Crowley ➔ Tim Stimpson ➔ Neil Back ➔ Joe Roff ➔ Gregor Townsend ➔ Craig Quinnell ➔ All Black ➔ Owen Finegan ➔ Norm Hewitt ➔ Scott Quinnell ➔ Tim Stimpson ➔ Australia ➔ Norm Hewitt ➔ Dale McIntosh ➔ Tim Horan ➔ David Giffin ➔ Tony Underwood ➔ David Campese ➔ Ian Jones ➔ Ian Jones ➔ Daniel Herbert ➔ Barbarians ➔ Barbarians ➔ Pat Howard ➔ David Wilson ➔ England ➔ Wales ➔ England ➔ England ➔ Wales ➔ Wales ➔ Wales ➔ England ➔ Scotland ➔ Wales ➔ Pontypridd ➔ Leicester ➔ Australia ➔ England ➔ Wales ➔ Italy ➔ Scotland ➔ Ireland ➔ Ireland ➔ Nigel Walker ➔ New Zealand ➔ New Zealand ➔ Matthew Burke

Figure 3: An illustrative example showing the inference performed by our fine-tuned confidence-order model on a document in the CoNLL dataset. Mentions are shown as underlined. Numbers in boldface represent the selection order of the confidence-order model.

Name	#mentions	#documents
CoNLL (training)	18,448	946
CoNLL (development)	4,791	216
CoNLL (test)	4,485	231
MSNBC	656	20
AQUAINT	727	50
ACE2004	257	36
CWEB	11,154	320
WIKI	6,821	320

Table 6: Statistics of ED datasets.

Australian rugby player born in 1973 instead of the former Australian rugby player born in 1964. This is accomplished by resolving other contextual mentions, including their colleague players, in advance. These two mentions are denoted in red in the figure. Note that our local model fails to resolve both mentions, and our natural-order model fails to resolve "Matthew Burke."