Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Improving entity linking with two adaptive features*

Hongbin ZHANG, Quan CHEN, Weiwen ZHANG†‡

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

†E-mail: zhangww@gdut.edu.cn

Received Oct. 18, 2021; Revision accepted Mar. 3, 2022; Crosschecked Sept. 16, 2022; Published online Sept. 24, 2022

Abstract: Entity linking (EL) is a fundamental task in natural language processing. Based on neural networks, existing systems pay more attention to the construction of the global model, but ignore latent semantic information in the local model and the acquisition of effective entity type information. In this paper, we propose two adaptive features, in which the first adaptive feature enables the local and global models to capture latent information, and the second adaptive feature describes effective information for entity type embeddings. These adaptive features can work together naturally to handle some uncertain entity type information for EL. Experimental results demonstrate that our EL system achieves the best performance on the AIDA-B and MSNBC datasets, and the best average performance on out-domain datasets. These results indicate that the proposed adaptive features, which are based on their own diverse contexts, can capture information that is conducive for EL.

Key words: Entity linking; Local model; Global model; Adaptive features; Entity type
https://doi.org/10.1631/FITEE.2100495 **CLC number:** TP391.1

1 Introduction

Entity linking (EL) is essential in obtaining unambiguous entity information in natural language processing. Entities with ambiguity should be identified for efficient information extraction (Hoffmann et al., 2011) and high-quality knowledge graph construction (Wang HF and Liu, 2019). These entities are usually referred to as entity mentions. The aim of EL is to correctly link a mention to a gold entity. A mention may have multiple ambiguities and correspond to various candidate entities. A case in point is that the mention "train" in the sentence "the goal is to make it easier to travel by train" may correspond to different candidates, such as boxing training, roller coaster, or road train. The mention "train" should be linked to the candidate "road train" as a true linked entity by EL. To accomplish this task, more effective methods should be developed for named entity recognition (NER) (Li et al., 2020).

The mainstream EL systems can be regarded as a combination of two models, i.e., local and global models. The local model concerns mainly modeling of semantic similarity of the context around a mention. In addition, entity type information about all mentions and candidates is usually added to the local model. However, the global model focuses on seeking coherence among all relevant entities in a document. Milne and Witten (2008) proposed link disambiguation to calculate entity correlations, with a knowledge base that contains detailed information about each entity to assist in EL. According to the rule developed by Le and Titov (2018), candidates for a mention can be generated first from the knowledge base, and then the mention is linked with the

‡ Corresponding author

* Project supported by the Key-Area Research and Development Program of Guangdong Province, China (No. 2019B010153002), the Program of Marine Economy Development (Six Marine Industries) Special Foundation of Department of Natural Resources of Guangdong Province, China (No. GDNRC [2020]056), the National Natural Science Foundation of China (No. 62002071), the Top Youth Talent Project of Zhujiang Talent Program, China (No. 2019QN01X516), and the Guangdong Provincial Key Laboratory of Cyber-Physical System, China (No. 2020B1212060069)

ORCID: Hongbin ZHANG, https://orcid.org/0000-0001-6568-5117; Weiwen ZHANG, https://orcid.org/0000-0002-5098-6459

most relevant and unique candidate to complete this task. On one hand, a host of novel EL systems have been proposed in recent years that focus on the extraction of global information and enhance the efficiencies in both training and inference (Ganea and Hofmann, 2017; Yang et al., 2019; Mulang et al., 2020). On the other hand, the local model should not be ignored for developing EL. A local neural attention model was provided by Ganea and Hofmann (2017); it adopts dot-product attention to gain local feature scores for candidates. However, this model is unable to capture latent semantic information in the context around the mention. This motivates us to research extraction of latent information in the local model and to explore a more comprehensive EL system.

In this study, we propose two adaptive features for EL. The first adaptive feature is applied in the local and global models, and the second adaptive feature is used for entity type modeling. These adaptive features help the EL system capture conducive information that is based on its own diverse contexts. Specifically, the first adaptive feature can explore latent relationships between individuals from a context by the self-attention mechanism used by Vaswani et al. (2017). The second adaptive feature can describe effective entity type information via a feed-forward neural network. What is more, these adaptive features are independent parts of our EL system, and can be linked naturally to work together to handle some uncertain entity type information. Therefore, we propose our EL system by integrating two adaptive features in the EL system of Yang et al. (2019) to better capture conducive information.

The contributions of this paper can be summarized as follows:

1. We propose two adaptive features that enable the EL system to capture latent information, describe effective information for entity type embeddings, and handle some uncertain entity type information.

2. We integrate these adaptive features in the local model, global model, and entity type modeling of EL systems.

3. Extensive experiments demonstrate the effectiveness of the proposed adaptive features. Significant improvement on six public datasets has been achieved; the best performance on the in-domain dataset and the highest average score on out-domain datasets have been displayed.

2 Related works

EL systems have been widely investigated, including local, global, and entity type information. Most researchers employ local and global models together for EL (Shen et al., 2015; Wu et al., 2020), while some supply entity type information in EL (Durrett and Klein, 2014; Chen et al., 2020). In the following, we review related works for EL.

2.1 Local and global information

Local models explore local information between the context of the mention and its candidates. Traditional feature extraction methods design features artificially (Bunescu and Paşca, 2006; Honnibal and Dale, 2009; Han and Sun, 2011), and introduce background knowledge and probabilities of entities in a knowledge base into EL. However, these traditional feature methods are unable to capture semantic information. With the development of neural networks, some methods (Yamada et al., 2016; Ganea and Hofmann, 2017; Hou et al., 2020; Runge and Hovy, 2020) have been beneficial in solving this issue by learning correct entity embeddings. Based on entity embeddings, a convolutional neural network model has been proposed to capture more fine-grained semantic contextual information (Francis-Landau et al., 2016). The joint model framework (Wang Z et al., 2014) and alignment technology (Zhong et al., 2015) have been used to design a feature that could present correlations between significant words around a mention and its candidates (Fang W et al., 2016). A local neural attention model has been employed to calculate local semantic similarity (Ganea and Hofmann, 2017). Although this model is simple, it is not sufficient to capture more information. Recently, Zhang et al. (2022) proposed the hidden semantic information extractor for this model to update vectors of candidates iteratively. In contrast, we propose the first adaptive feature and integrate it into this model, which helps the local model explore latent context information to construct local EL information.

Global information between entities should be considered to compensate for the limitations of local models. The global model aims to associate

all entities that have coherence of a theme in a document. This model usually has an objective function to acquire the candidate with the highest probability (Pershina et al., 2015; Globerson et al., 2016; Cao et al., 2018; Xue et al., 2019; Sevgili et al., 2020). Recently, a sequential decision method, i.e., dynamic context augmentation (DCA) (Yang et al., 2019), has emerged, which is similar to the one proposed in Fang Z et al. (2019). This method disambiguates all mentions in order for each document, but an error accumulation problem, caused by an incomplete EL system, needs to be solved. Designing a more comprehensive EL system is important because every linking result is a key for future decisions. However, based on the local model, we also introduce the first adaptive feature into the global model to enhance the acquisition of global information.

2.2 Entity type information

Entity type information is significant for EL because a mention and its predicted entity have the same type. Hence, some previous works tried to introduce entity type information into EL by completely joint learning of NER and EL (Luo et al., 2015; Nguyen et al., 2016). Based on a conditional random field (CRF) or its variant, hand-engineered features have been explored to consider the interrelations between NER and EL. Then, learnable features were proposed by Martins et al. (2019) to conduct multi-task learning. A typing system has been trained that yields 95% accuracy on the AIDA-train dataset to compensate for the lack of predicting a mention type (Xu and Barbosa, 2018). Specifically, entity types for mentions and their corresponding candidates in other datasets have been predicted by this system. Then, these entity types have been used in the EL system by training entity type embeddings.

In the mentions for the AIDA-train dataset, UNK (unknown) denotes the miscellaneous type. About 38.44% of mention entity types were identified as UNK and only about 6.58% of mention entity types were identified as geo-political entity (GPE). There is no doubt that distinctive entity type information is important, and there exists some uncertain type information, like UNK, in the EL system. To solve these issues, we propose the second adaptive feature, integrate it into this method (Yang et al., 2019), and combine it with the first adaptive feature.

3 Background

3.1 Problem definition

Given a list of mentions $m = \{\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_n\}$ in a document D, where D is from a corpus $\mathcal{D}$, the goal of the EL system is to assign each mention $\boldsymbol{m}_i$ to its corresponding gold entity $\boldsymbol{e}_i^*$. In general, the EL system is divided into two stages. First, candidate generation selects a set of potential (candidate) entities $E_i = \left\{\boldsymbol{e}_i^1, \boldsymbol{e}_i^2, ..., \boldsymbol{e}_i^R\right\}$ for $\boldsymbol{m}_i$ by applying a heuristic and deep learning model. Second, candidate ranking aims to rank all candidates; the top-ranked candidate will be linked as the predicted entity $\hat{\boldsymbol{e}}_i$ by the EL system.

In this study, we focus on the second stage, whose target is exploring high-quality features to predict the entity precisely for each mention. In particular, three kinds of information have been leveraged in the EL system in recent years—local context similarity, document-level global coherence, and entity type similarity—which are respectively scored by the local model, global model, and entity type modeling. In the following, we introduce these methods as proposed by Yang et al. (2019).

3.2 Local and global models

For a mention $\boldsymbol{m}_i$ with its local context $c_i = \left\{\boldsymbol{w}_i^1, \boldsymbol{w}_i^2, \ldots, \boldsymbol{w}_i^o\right\}$ and a candidate $\boldsymbol{e}_i^r \in E_i$ ($r \in \{1, 2, \ldots, R\}$), two feature scores are taken into account: First, local context similarity $\psi_{\mathrm{C}}\left(\boldsymbol{e}_i^r, c_i\right)$ represents the semantic similarity between $\boldsymbol{e}_i^r$ and c_i. Specifically, in this study, our first adaptive feature is incorporated into the local model (Ganea and Hofmann, 2017). Second, mention-entity prior $\hat{P}\left(\boldsymbol{e}_i^r|\boldsymbol{m}_i\right)$ is calculated using count statistics of mention-entity hyperlinks from the Web corpora, YAGO, and Wikipedia.

Unlike the local model, two different feature scores exist in the global model: global coherence $\phi_{\mathrm{D}}\left(\boldsymbol{e}_i^r, d_{i-1}\right)$ and global (neighbor) coherence $\phi_{\mathrm{N}}\left(\boldsymbol{e}_i^r, n_{i-1}\right)$. Following the DCA model, accumulating knowledge from previously linked entities as dynamic context was proposed by Yang et al. (2019). In an independent document, all mentions appear in context order and $\boldsymbol{m}_{i-1}$ is linked to a candidate as the predicted entity $\hat{\boldsymbol{e}}_{i-1}$ in order by the EL system. The mention $\boldsymbol{m}_i$ has its dynamic context $d_{i-1} = \{\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, ..., \hat{\boldsymbol{e}}_{i-1}\}$, where d_{i-1} is composed of

predicted entities before $\boldsymbol{m}_i$ in the document. Hence, for $\boldsymbol{m}_i$ with d_{i-1}, global coherence denotes the global coherence between $\boldsymbol{e}_i^r$ and d_{i-1}. Furthermore, to enhance the associative ability of the EL system, neighbor entities that have inlinks pointing to $\hat{\boldsymbol{e}}_{i-1} \in d_{i-1}$ have been collected from Wikipedia by Yang et al. (2019). For the mention $\boldsymbol{m}_i$ with its neighbor entities of dynamic context $n_{i-1} = \{\tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2, ..., \tilde{\boldsymbol{e}}_{i-1}\}$, global (neighbor) coherence represents the global (neighbor) coherence between $\boldsymbol{e}_i^r$ and n_{i-1}. In this study, we extend the local model with our first adaptive feature to the global model to calculate these global feature scores as described in Yang et al. (2019).

3.3 Entity type modeling

The feature score entity type similarity, $\psi_{\mathrm{T}}(\boldsymbol{m}_i, \boldsymbol{e}_i^r)$, means the similarity of the entity types $\boldsymbol{e}_i^r$ and $\boldsymbol{m}_i$. According to all types (person (PER), GPE, organization (ORG), and UNK), the entity type probability distribution of each entity or mention is predicted by the typing system (Yang et al., 2019), and is encoded in one-hot format. Then, based on one-hot encodings, $\psi_{\mathrm{T}}(\boldsymbol{m}_i, \boldsymbol{e}_i^r)$ is computed by dot-product attention while training entity type embeddings. Therefore, we combine this method with our second adaptive feature in this study.

Finally, to evaluate the local model, a two-layer feed-forward neural network with 100 hidden units as depicted in Ganea and Hofmann (2017) is applied to integrate three feature scores ($\psi_{\mathrm{C}}(\boldsymbol{e}_i^r, c_i)$, $\hat{P}(\boldsymbol{e}_i^r|\boldsymbol{m}_i)$, and $\psi_{\mathrm{T}}(\boldsymbol{m}_i, \boldsymbol{e}_i^r)$) to output the conditional probability $P = (\boldsymbol{e}_i^r|\boldsymbol{m}_i)$, where $\boldsymbol{e}_i^r \in E_i$. Thus, to evaluate the EL system, two global feature scores $\phi_{\mathrm{D}}(\boldsymbol{e}_i^r, d_{i-1})$ and $\phi_{\mathrm{N}}(\boldsymbol{e}_i^r, n_{i-1})$ are added to form five feature scores based on the evaluation of the local model.

4 EL system with two adaptive features

In this section, we illustrate the EL system with two adaptive features. In Section 4.1, with the first adaptive feature, we introduce the local model and the global model. Then, with the second adaptive feature, we present details of entity type modeling in Section 4.2 to construct the comprehensive EL system.

4.1 Local and global models with the first adaptive feature

Given a list of candidates $E_i = \{\boldsymbol{e}_i^1, \boldsymbol{e}_i^2, ..., \boldsymbol{e}_i^R\}$ of mention $\boldsymbol{m}_i$ and its corresponding context $c_i = \{\boldsymbol{w}_i^1, \boldsymbol{w}_i^2, ..., \boldsymbol{w}_i^o\}$, the local model is designed mainly to calculate the local context similarity between them. The key point of the computational process is to obtain a context embedding that describes semantic information about the context, where each weight of words plays an important role in the context. Although these weights are related to the similarity measure between the context and candidates, there are potential and strong correlations between these weights according to the independent context.

The basic idea of the first adaptive feature is to capture latent information. The specific process is demonstrated in Fig. 1, where the adaptive feature consists of two steps: one computes with the embedding matrix $\boldsymbol{B}$ and the other computes with the embedding matrix $\boldsymbol{C}$. Two operations, "FNN" and "$\bigotimes$," belong to the first and second steps, respectively. We describe the calculation process of local context similarity with the first adaptive feature in four parts.

First, for each $\boldsymbol{w}_i^o \in c_i$, two maximum relevance scores are calculated. We compute the first relevance score $u_1(\boldsymbol{w}_i^o)$ by associating it with candidates as

$$u_1(\boldsymbol{w}_i^o) = \max_{\boldsymbol{e}_i^r \in E_i} \left((\boldsymbol{e}_i^r)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w}_i^o \right), \tag{1}$$

where $\boldsymbol{A}$ is a parameterized diagonal matrix. We also compute the second relevance score $u_2(\boldsymbol{w}_i^o)$ to explore latent information between the individuals from a context via the self-attention mechanism as

$$u_2(\boldsymbol{w}_i^o) = \max_{\boldsymbol{w}_i^j \in c_i} \left(\left(\boldsymbol{w}_i^j \right)^{\mathrm{T}} \boldsymbol{B} \boldsymbol{w}_i^o \right), \tag{2}$$

where $\boldsymbol{B}$ is a parameterized diagonal matrix. For these two relevance scores, we propose a two-layer feed-forward neural network with 100 hidden units to automatically train their own weights; then these scores are multiplied by their own weights. They are added to generate a new relevance score $u_{1\_2}(\boldsymbol{w}_i^o)$. The computing process is as follows:

$$\begin{aligned} u_1^*(\boldsymbol{w}_i^o) &= \boldsymbol{W}_1^2 \cdot \mathrm{ReLU}\left(\boldsymbol{W}_1^1 \left(u_1(\boldsymbol{w}_i^o) \right)^{\mathrm{T}} + \boldsymbol{B}_1^1 \right) + \boldsymbol{B}_1^2, \\ u_2^*(\boldsymbol{w}_i^o) &= \boldsymbol{W}_2^2 \cdot \mathrm{ReLU}\left(\boldsymbol{W}_2^1 \left(u_2(\boldsymbol{w}_i^o) \right)^{\mathrm{T}} + \boldsymbol{B}_2^1 \right) + \boldsymbol{B}_2^2, \\ u_{1\_2}(\boldsymbol{w}_i^o) &= \left(u_1^*(\boldsymbol{w}_i^o) \right)^{\mathrm{T}} + \left(u_2^*(\boldsymbol{w}_i^o) \right)^{\mathrm{T}}, \end{aligned} \tag{3}$$

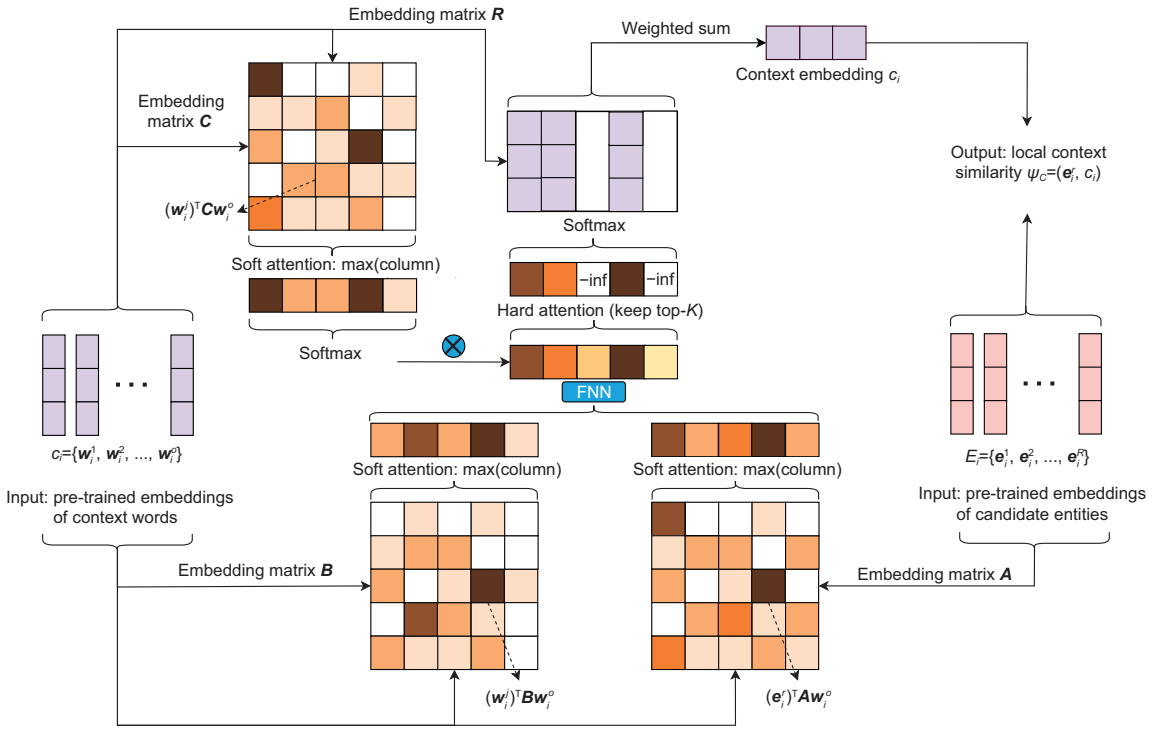

Fig. 1 Computational process of local context similarity, where the first adaptive feature is split into two parts that are respectively associated with the embedding matrices B and C. The first adaptive feature contains two operations, where "FNN" denotes the addition operation which is based on a two-layer feed-forward neural network and "⊗" denotes the multiplication operation

where ReLU is an activation function. $\boldsymbol{W}_1^1$ and $\boldsymbol{W}_1^2$ are respectively trainable weights of the first and second layer in FNN for $u_1(\boldsymbol{w}_i^o)$. $\boldsymbol{B}_1^1$ and $\boldsymbol{B}_1^2$ denote trainable biases of the first and second layer in FNN for $u_1(\boldsymbol{w}_i^o)$, respectively. $\boldsymbol{W}_2^1$ and $\boldsymbol{W}_2^2$ are respectively trainable weights of the first and second layer in FNN for $u_2(\boldsymbol{w}_i^o)$. $\boldsymbol{B}_2^1$ and $\boldsymbol{B}_2^2$ denote trainable biases of the first and second layer in FNN for $u_2(\boldsymbol{w}_i^o)$, respectively. However, some keywords with lower weights that have latent information are ignored when training the weights. Dealing with this problem is difficult in the first step with "FNN." We provide the second step with "⊗" to increase the influence of latent information to alleviate this problem in the second part.

Second, we compute the third relevance score $u_3(\boldsymbol{w}_i^o)$ between the words as

$$u_3(\boldsymbol{w}_i^o) = \max_{\boldsymbol{w}_i^j \in c_i} \left(\boldsymbol{w}_i^j \boldsymbol{C} \boldsymbol{w}_i^o \right), \tag{4}$$

where $\boldsymbol{C}$ is a parameterized diagonal matrix. Then the reinforced weight $a(u_3(\boldsymbol{w}_i^o))$ and the final relevance score $u(\boldsymbol{w}_i^o)$ are explicitly as follows:

$$\begin{cases} a(u_3(\boldsymbol{w}_i^o)) = \dfrac{\exp\left[u_3(\boldsymbol{w}_i^o)\right]}{\sum_{\boldsymbol{w}_i^j \in c_i} \exp\left[u_3(\boldsymbol{w}_i^j)\right]}, \\ u(\boldsymbol{w}_i^o) = u_{1\_2}(\boldsymbol{w}_i^o) \cdot a(u_3(\boldsymbol{w}_i^o)). \end{cases} \tag{5}$$

Third, to highlight the significant information, the top-K words are considered in the context. Weights of unselected words are set as negative infinity. The final attention weight of a word is computed as

$$a(\boldsymbol{w}_i^o) = \frac{\exp\left[u(\boldsymbol{w}_i^o)\right]}{\sum_{\boldsymbol{w}_i^j \in c_i} \exp\left[u(\boldsymbol{w}_i^j)\right]}. \tag{6}$$

Fourth, based on the final attention weight of every word, the context embedding is computed by the weighted sum. The local context similarity between $\boldsymbol{e}_i^r \in E_i$ and c_i is

$$\psi_{\mathrm{C}}\left(\boldsymbol{e}_i^r, c_i\right) = \sum_{\boldsymbol{w}_i^o \in c_i} a(\boldsymbol{w}_i^o) \left(\boldsymbol{e}_i^r\right)^{\mathrm{T}} \boldsymbol{R} \boldsymbol{w}_i^o, \tag{7}$$

where $\boldsymbol{R}$ is a parameterized diagonal matrix. Such a feature score is one part of the input for the local model or the EL system.

The first adaptive feature can also collect latent information from a dynamic context when it is incorporated in the global model. Hence, given E_i and a set of previously linked entities $d_{i-1} = \{\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, ..., \hat{\boldsymbol{e}}_{i-1}\}$ for $\boldsymbol{m}_i$, the global coherence between $\boldsymbol{e}_i^r \in E_i$ and d_{i-1} is

$$\phi_{\mathrm{D}}\left(\boldsymbol{e}_i^r, d_{i-1}\right) = \sum_{\hat{\boldsymbol{e}}_{i-1} \in d_{i-1}} a(\hat{\boldsymbol{e}}_{i-1}) \left(\boldsymbol{e}_i^r\right)^{\mathrm{T}} \boldsymbol{S} \hat{\boldsymbol{e}}_{i-1}, \quad (8)$$

where $\boldsymbol{S}$ is a parameterized diagonal matrix. Therefore, given E_i and a list of neighbor entities $n_{i-1} = \{\tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2, ..., \tilde{\boldsymbol{e}}_{i-1}\}$, the global (neighbor) coherence between $\boldsymbol{e}_i^r \in E_i$ and n_{i-1} is

$$\phi_{\mathrm{N}}\left(\boldsymbol{e}_i^r, n_{i-1}\right) = \sum_{\tilde{\boldsymbol{e}}_{i-1} \in n_{i-1}} a(\tilde{\boldsymbol{e}}_{i-1}) \left(\boldsymbol{e}_i^r\right)^{\mathrm{T}} \boldsymbol{S}' \tilde{\boldsymbol{e}}_{i-1}, \quad (9)$$

where $\boldsymbol{S}'$ is a parameterized diagonal matrix.

4.2 Entity type modeling with the second adaptive feature

The basic idea of the second adaptive feature is to describe effective entity type information for training the four kinds of entity type embeddings. Specifically, for all mentions on the AIDA-train dataset, every entity type accounts for a different proportion: 18.40% for PER, 6.58% for GPE, 36.58% for ORG, and 38.44% for UNK. Thus, it is necessary to properly adjust each entity type's own embeddings to represent its distinctive entity type information. However, UNK denotes a miscellaneous entity type, which illustrates that careful type system classification is difficult. Hence, these two adaptive features work together naturally and can alleviate the problem that some uncertain entity type information exists for UNK. The entity type similarity calculation process with the second adaptive feature is described as follows:

First, based on random initialization, four kinds of entity type embeddings with the second adaptive feature are computed via a feed-forward neural network as

$$\begin{aligned} F_1 &= \boldsymbol{W}_3^1 \cdot \hat{\boldsymbol{t}}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right) + \boldsymbol{B}_3^1, \\ \boldsymbol{t}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right) &= \boldsymbol{W}_3^2 \cdot \mathrm{ReLU}\left(F_1\right) + \boldsymbol{B}_3^2, \end{aligned} \quad (10)$$

where $\hat{\boldsymbol{t}}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right)$ denotes a learnable embedding matrix initialized with normal distribution for four kinds of entity types.

Second, entity type similarity between $\boldsymbol{e}_i^r \in E_i$ and $\boldsymbol{m}_i$ is

$$\psi_{\mathrm{T}}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right) = \left(\hat{\boldsymbol{o}}_{\boldsymbol{m}_i} \cdot \boldsymbol{t}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right)\right)^{\mathrm{T}} \cdot \left(\hat{\boldsymbol{o}}_{\boldsymbol{e}_i^r} \cdot \boldsymbol{t}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right)\right), \quad (11)$$

where $\hat{\boldsymbol{o}}_{\boldsymbol{m}_i}$ and $\hat{\boldsymbol{o}}_{\boldsymbol{e}_i^r}$ denote the one-hot encodings of $\boldsymbol{m}_i$ and $\boldsymbol{e}_i^r$, respectively.

Based on the two-layer feed-forward neural network, five feature scores $\psi_{\mathrm{C}}\left(\boldsymbol{e}_i^r, c_i\right)$, $\hat{P}\left(\boldsymbol{e}_i^r | \boldsymbol{m}_i\right)$, $\phi_{\mathrm{D}}\left(\boldsymbol{e}_i^r, d_{i-1}\right)$, $\phi_{\mathrm{N}}\left(\boldsymbol{e}_i^r, n_{i-1}\right)$, and $\psi_{\mathrm{T}}\left(\boldsymbol{m}_i, \boldsymbol{e}_i^r\right)$ are integrated to output conditional probability $P = \left(\boldsymbol{e}_i^r | \boldsymbol{m}_i\right)$ by the EL system with two adaptive features. We then use the max-margin loss as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{D \in \mathcal{D}} \sum_{\boldsymbol{m}_i \in D} \sum_{\boldsymbol{e}_i^r \in E_i} g\left(\boldsymbol{e}_i^r, \boldsymbol{m}_i\right), \quad (12)$$

where

$$g\left(\boldsymbol{e}_i^r, \boldsymbol{m}_i\right) = \max\left(0, \gamma - P\left(\boldsymbol{e}_i^* | \boldsymbol{m}_i\right) + P\left(\boldsymbol{e}_i^r | \boldsymbol{m}_i\right)\right),$$

γ is a margin coefficient, and $\boldsymbol{m}_i$ is linked to the candidate with the highest conditional probability.

5 Experiments

In this section, we evaluate the performance of our local model and our EL system to verify the effectiveness of the two adaptive features. First, we introduce datasets and parameter settings. Then, we show performance evaluations of local models and EL systems. Finally, our two adaptive features are verified in detail using ablation analysis.

5.1 Datasets

Following Ganea and Hofmann (2017), we validate the local model and EL system with adaptive features on six public datasets. Table 1 shows the dataset statistics. Based on the entity recognition task CoNLL 2013 dataset, the AIDA-CoNLL dataset is annotated manually, and is divided into AIDA-train (training), AIDA-A (validation), and AIDA-B (test) (Hoffart et al., 2011). Updated MSNBC, AQUAINT, and ACE2004 datasets are taken from Xu and Barbosa (2018). MSNBC contains 10 news articles on various topics. AQUAINT is provided by the Xinhua News Agency, *New York Times*, and the Associated Press. ACE2004 is annotated through crowdsourcing. The other two datasets are separately extracted from Clueweb and Wikipedia, and have more noise than the others (Guo and Barbosa, 2018).

Table 1 Statistics of six datasets (reprinted from Zhang et al. (2022), Copyright 2022, with permission from ACM)

Dataset	Number of mentions	Number of docs	Mentions per doc
AIDA-train	18 448	946	19.5
AIDA-A	4791	216	22.1
AIDA-B	4485	231	19.4
MSNBC	656	20	32.8
AQUAINT	727	50	14.5
ACE2004	257	36	7.1
CWEB	11 154	320	34.8
WIKI	6821	320	21.3

AIDA-A is for validation; AIDA-B is for test. AIDA-train, AIDA-A, and AIDA-B are divided from the AIDA-CoNLL dataset

5.2 Parameter settings

We tune the hyper-parameters according to the performance of the local model and EL system on the AIDA-A dataset. The local model and EL system are trained on the AIDA-train dataset. We validate them on the AIDA-A dataset, and test them on the AIDA-B dataset and other out-domain datasets (MSNBC, AQUAINT, ACE2004, CWEB, WIKI). The pre-trained word vectors and entity vectors are taken from Pennington et al. (2014) and Ganea and Hofmann (2017), respectively, where the embedding dimensions are both 300. The dimension of entity type embeddings is 1024. The numbers of hidden units of the two-layer feed-forward neural network are 100 and 812 for the first and second adaptive features, respectively. In addition, the learning rate is set to 2×10^{-4}, and we employ the Adam optimizer during training. The learning rate is reduced by half to 1×10^{-4} when the validation accuracy reaches 92.8%. The margin coefficient γ is 0.01.

5.3 Results and discussion

We demonstrate the effectiveness of the two adaptive features comprehensively, compare our local model with other different local models, contrast our EL system with other diverse EL systems, and analyze our two adaptive features by ablation analysis.

5.3.1 Results for local models and disambiguation systems on the in-domain dataset

Table 2 shows the performance evaluation of the local model on the AIDA-B test dataset, where the micro F1 score is used to evaluate the effectiveness. There are six local models for comparison.

Table 2 Micro F1 score for performance evaluation of local models and EL systems with adaptive features on the AIDA-B test set (adapted from Zhang et al. (2022), Copyright 2022, with permission from ACM)

Local model	Micro F1 score (%)
Prior (Medelyan et al., 2009)	71.90
Plato+star (Globerson et al., 2016)	87.90
Skip-gram (Yamada et al., 2016)	87.20
ETHZ-Attn (Yang et al., 2019)	90.88
BERT-Entity-Sim (Chen et al., 2020)	90.06
Knowledge-aware (Deng et al., 2020)	90.38
Our local model	**90.99**
EL system	**Micro F1 score (%)**
MulFocal-Att (Globerson et al., 2016)	91.00
Two-step (Yamada et al., 2016)	91.50
Deep-ED (Ganea and Hofmann, 2017)	92.22
Ment-norm (Le and Titov, 2018)	93.07
DCA (Yang et al., 2019)	93.73
BERT-Entity-Sim (Chen et al., 2020)	93.60
Knowledge-aware (Deng et al., 2020)	93.60
Our system	**94.20**

The best results are in bold

1. Prior (Medelyan et al., 2009) is a screening method that was employed to screen out the candidate entity with the highest conditional probability.

2. Plato+star (Globerson et al., 2016) is based on an attention model in which inference is tractable.

3. Skip-gram (Yamada et al., 2016) explores proper word representations according to word-entity statistics for EL.

4. ETHZ-Attn (Yang et al., 2019) is based on the local neural attention model (Ganea and Hofmann, 2017) with entity type information from the type system (Xu and Barbosa, 2018).

5. BERT-Entity-Sim (Chen et al., 2020) integrates the BERT-based entity similarity score into the local model (Ganea and Hofmann, 2017) to capture latent entity type information.

6. Knowledge-aware (Deng et al., 2020) designs a graph convolutional network for entity embeddings and uses the multi-hop attention mechanism to balance the context and external knowledge.

To accelerate the training of the local model, only the first adaptive feature is applied in the local model. We can observe that our local model is better than other local models on this dataset, which demonstrates the effectiveness of the first adaptive feature. We also integrate these two adaptive features in the EL system to evaluate the proposed

adaptive features comprehensively.

We compare our EL system with other well-known EL systems. We choose seven kinds of methods that are based on neural networks.

1. MulFocal-Att (Globerson et al., 2016) puts forward multi-focal attention and a star model to resolve global disambiguation. The system can easily integrate attention into learning and inference.

2. Two-step (Yamada et al., 2016) and Deep-ED (Ganea and Hofmann, 2017) are proposed to map words and entities to the same continuous vector space.

3. Ment-norm (Le and Titov, 2018) enhances a disambiguation system by extracting relationships between mentions as potential variables.

4. DCA (Yang et al., 2019) improves later decisions by accumulating knowledge from previously linked entities as dynamic context.

5. BERT-Entity-Sim (Chen et al., 2020) deals with errors of entity type information by combining BERT (Devlin et al., 2019) with a local model.

6. Knowledge-aware (Deng et al., 2020) proposes the multi-hop attention mechanism for local models and the graph-based search algorithm for global models.

In Table 2, our EL system with the two adaptive features outperforms the other EL systems and achieves the highest score. This demonstrates that our two adaptive features are beneficial for EL.

5.3.2 Results for disambiguation systems on out-domain datasets

Seven EL systems are compared to assess our EL system on other out-domain datasets. In Table 3, the micro F1 score is used to evaluate the effectiveness. EL systems from Ganea and Hofmann (2017), Le and Titov (2018), Yang et al. (2019), and Deng et al. (2020) are also compared with our EL system. In addition, three other EL systems are compared:

1. WNED (Guo and Barbosa, 2018) designs an effective graph-based approach that includes a random walk algorithm and a global disambiguation algorithm based on information theory.

2. CoSimTC (Xin et al., 2019) proposes extracting global features locally to combine the superiority of local and global models.

3. FGS2EE (Hou et al., 2020) provides fine-grained semantic information for training better entity embeddings to reduce distinctiveness.

Compared with other EL systems (Table 3), our EL system achieves the highest score on the MSNBC dataset and ranks second on the AQUAINT dataset. Our EL system achieves the best average performance on out-domain datasets, indicating that the proposed two adaptive features are beneficial for EL. CoSimTC (Xin et al., 2019) has the highest scores on the AQUAINT and ACE2004 datasets, and its tree connection method, based on the syntactic distance of a dependency parse tree, is good for clearer and shorter sentences on the MSNBC, AQUAINT, and ACE2004 datasets. Deep-ED (Ganea and Hofmann, 2017) has the highest score on the CWEB dataset, but is slightly worse than ours on other datasets. On the WIKI dataset, WNED (Guo and Barbosa, 2018) performs especially well, but its performance is relatively bad on the MSNBC, AQUAINT, and ACE2004 datasets.

Our EL system is analyzed from two perspectives to explain drops in performance on out-domain datasets. First, entity types are divided into only

Table 3 Micro F1 score for performance evaluation of EL systems on out-domain datasets (adapted from Zhang et al. (2022), Copyright 2022, with permission from ACM)

EL system	Micro F1 score (%)					Average score (%)
	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	
WNED (Guo and Barbosa, 2018)	92.00	87.00	88.00	77.00	**84.50**	85.70
Deep-ED (Ganea and Hofmann, 2017)	93.70	88.50	88.50	**77.90**	77.50	85.22
Ment-norm (Le and Titov, 2018)	93.90	88.30	89.90	77.50	78.00	85.52
DCA (Yang et al., 2019)	93.80	88.25	90.14	75.59	78.84	85.32
CoSimTC (Xin et al., 2019)	94.16	**90.90**	**92.92**	76.96	75.02	85.99
FGS2EE (Hou et al., 2020)	94.26	88.47	90.70	77.41	77.66	85.70
Graph-based-EL (Deng et al., 2020)	**94.41**	89.23	90.54	76.64	78.20	85.80
Our system	**94.41**	90.21	90.54	76.97	78.16	**86.06**

The best results are in bold. If a result has a confidence interval, we determine the median from it

four kinds by the type system, but more fine-grained entity types should be found. Although our two adaptive features can train the entity type embeddings adaptively to describe entity type information and alleviate the problem as presented in Section 4.2, it is difficult for them to capture latent fine-grained entity types from the four kinds of entity types. Second, the two adaptive features aim to achieve the best average performance on out-domain datasets. For one thing, the last two datasets (CWEB and WIKI) have more noise. For another, the AQUAINT dataset has less context information than MSNBC and ACE2004 datasets. Although the two adaptive features can capture latent context and entity type information, our EL system requires more additional knowledge to capture more latent information to improve system performance, such as entity descriptions.

5.3.3 Ablation analysis

The two proposed adaptive features are divided into three independent parts to guarantee the fairness of ablation analysis in this study. First, only_AP_first_model, without the second adaptive feature, shows that only the first step with "FNN" is retained in the first adaptive feature. Second, only_AP_model, without the second adaptive feature, indicates the first adaptive feature with two steps that have "FNN" and "$\bigotimes$," respectively. Third, only_AP_type, without the first adaptive feature, denotes the second adaptive feature. As shown in Table 4, only_AP_first_model reached the highest score on the MSNBC dataset, which demonstrates the effectiveness of the first step of the first adaptive feature for automatically training weights. Only_AP_model achieved the highest scores on the ACE2004 and WIKI datasets by capturing latent information that is hidden in contexts, which also verifies that the second step can compensate for the first step in the first adaptive feature and can increase the influence of latent information. In addition, only_AP_type performs better than the other two systems on the AQUAINT and CWEB datasets, because it can describe effective entity type information for training proper entity type embeddings. To summarize, the above three systems achieved better results on different datasets. Thus, the flexibility of our adaptive features enables them to be merged naturally, to play their own roles in alleviating the problem of some uncertain entity type information existing in UNK. In particular, three systems from Table 4 were compared to our system. Our system acquired obvious improvement on the AQUAINT and CWEB datasets that are more challenging.

6 Conclusions and future work

In this paper, we have proposed two adaptive features for EL. The first adaptive feature has been applied in the local and global models and the second adaptive feature has been employed in entity type modeling. We have verified the effectiveness of these two adaptive features via comparative experiments and ablation analysis. Extensive experiments demonstrated that the first adaptive feature is beneficial for exploring the latent relationships between individuals from a context. We have further proposed the second adaptive feature that describes effective entity type information for training four kinds of entity type embeddings, to properly adjust their own embeddings to represent their distinctive entity type information. We also found that these two adaptive features work together naturally to handle some uncertain entity type information for UNK in EL. Although our EL system ranked second on the AQUAINT dataset, it achieved the best performance on the AIDA-B and MSNBC datasets. It also achieved the best average performance on out-domain datasets.

Table 4 Micro F1 score for EL systems of ablation analysis on out-domain datasets

EL system	Micro F1 score (%)					Average score (%)
	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	
only_AP_first_model	**94.72**	88.39	90.14	75.64	77.46	85.27
only_AP_model	94.41	88.25	**91.35**	75.70	**78.72**	**85.69**
only_AP_type	93.96	**89.51**	89.34	**76.23**	78.41	85.49

The best results are in bold

In the future, we will investigate a dependency parse tree that is suitable for our adaptive features to compensate for the deficiency in our EL system. We will also study a better entity type system or a proper fine-grained entity type system to improve the EL system performance.

Contributors

Hongbin ZHANG designed the research. Hongbin ZHANG, Quan CHEN, and Weiwen ZHANG processed the data. Weiwen ZHANG validated the research. Hongbin ZHANG drafted the paper. Quan CHEN helped organize the paper. Hongbin ZHANG and Weiwen ZHANG revised and finalized the paper.

Compliance with ethics guidelines

Hongbin ZHANG, Quan CHEN, and Weiwen ZHANG declare that they have no conflict of interest.

References

Bunescu R, Paşca M, 2006. Using encyclopedic knowledge for named entity disambiguation. Proc 11th Conf of the European Chapter of the Association for Computational Linguistics, p.9-16.

Cao YX, Hou L, Li JZ, et al., 2018. Neural collective entity linking. Proc 27th Int Conf on Computational Linguistics, p.675-686.

Chen S, Wang JP, Jiang F, et al., 2020. Improving entity linking by modeling latent entity type information. Proc 34th AAAI Conf on Artificial Intelligence, p.7529-7537. https://doi.org/10.1609/aaai.v34i05.6251

Deng ZH, Li ZX, Yang Q, et al., 2020. Improving entity linking with graph networks. Proc 21st Int Conf on Web Information Systems Engineering, p.343-354. https://doi.org/10.1007/978-3-030-62005-9_25

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Durrett G, Klein D, 2014. A joint model for entity analysis: coreference, typing, and linking. *Trans Assoc Comput Linguist*, 2:477-490. https://doi.org/10.1162/tacl_a_00197

Fang W, Zhang JW, Wang DL, et al., 2016. Entity disambiguation by knowledge and text jointly embedding. Proc 20th SIGNLL Conf on Computational Natural Language Learning, p.260-269. https://doi.org/10.18653/v1/K16-1026

Fang Z, Cao YN, Li Q, et al., 2019. Joint entity linking with deep reinforcement learning. Proc World Wide Web Conf, p.438-447. https://doi.org/10.1145/3308558.3313517

Francis-Landau M, Durrett G, Klein D, 2016. Capturing semantic similarity for entity linking with convolutional neural networks. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.1256-1261. https://doi.org/10.18653/v1/N16-1150

Ganea OE, Hofmann T, 2017. Deep joint entity disambiguation with local neural attention. Proc Conf on Empirical Methods in Natural Language Processing, p.2619-2629. https://doi.org/10.18653/v1/D17-1277

Globerson A, Lazic N, Chakrabarti S, et al., 2016. Collective entity resolution with multi-focal attention. Proc 54th Annual Meeting of the Association for Computational Linguistics, p.621-631. https://doi.org/10.18653/v1/P16-1059

Guo ZC, Barbosa D, 2018. Robust named entity disambiguation with random walks. *Semant Web*, 9(4):459-479. https://doi.org/10.3233/SW-170273

Han XP, Sun L, 2011. A generative entity-mention model for linking entities with knowledge base. Proc 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, p.945-954.

Hoffart J, Yosef MA, Bordino I, et al., 2011. Robust disambiguation of named entities in text. Proc Conf on Empirical Methods in Natural Language Processing, p.782-792.

Hoffmann R, Zhang CL, Ling X, et al., 2011. Knowledge-based weak supervision for information extraction of overlapping relations. Proc 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, p.541-550.

Honnibal M, Dale R, 2009. DAMSEL: the DSTO/Macquarie system for entity-linking. Proc 2nd Text Analysis Conf.

Hou F, Wang RL, He J, et al., 2020. Improving entity linking through semantic reinforced entity embeddings. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.6843-6848. https://doi.org/10.18653/v1/2020.acl-main.612

Le P, Titov I, 2018. Improving entity linking by modeling latent relations between mentions. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.1595-1604. https://doi.org/10.18653/v1/P18-1148

Li ZZ, Feng DW, Li DS, et al., 2020. Learning to select pseudo labels: a semi-supervised method for named entity recognition. *Front Inform Technol Electron Eng*, 21(6):903-916. https://doi.org/10.1631/FITEE.1800743

Luo G, Huang XJ, Lin CY, et al., 2015. Joint entity recognition and disambiguation. Proc Conf on Empirical Methods in Natural Language Processing, p.879-888. https://doi.org/10.18653/v1/D15-1104

Martins PH, Marinho Z, Martins AFT, 2019. Joint learning of named entity recognition and entity linking. Proc 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, p.190-196. https://doi.org/10.18653/v1/P19-2026

Medelyan O, Milne D, Legg C, et al., 2009. Mining meaning from Wikipedia. *Int J Hum-Comput Stud*, 67(9):716-754. https://doi.org/10.1016/j.ijhcs.2009.05.004

Milne D, Witten IH, 2008. Learning to link with Wikipedia. Proc 17th ACM Conf on Information and Knowledge Management, p.509-518. https://doi.org/10.1145/1458082.1458150

Mulang IO, Singh K, Prabhu C, et al., 2020. Evaluating the impact of knowledge graph context on entity disambiguation models. Proc 29th ACM Int Conf on Information & Knowledge Management, p.2157-2160. https://doi.org/10.1145/3340531.3412159

Nguyen DB, Theobald M, Weikum G, 2016. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. *Trans Assoc Comput Linguist*, 4:215-229. https://doi.org/10.1162/tacl_a_00094

Pennington J, Socher R, Manning C, 2014. GloVe: global vectors for word representation. Proc Conf on Empirical Methods in Natural Language Processing, p.1532-1543. https://doi.org/10.3115/v1/D14-1162

Pershina M, He YF, Grishman R, 2015. Personalized page rank for named entity disambiguation. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.238-243. https://doi.org/10.3115/v1/N15-1026

Runge A, Hovy E, 2020. Exploring neural entity representations for semantic information. Proc 3rd BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, p.204-216. https://doi.org/10.18653/v1/2020.blackboxnlp-1.20

Sevgili O, Shelmanov A, Arkhipov M, et al., 2020. Neural entity linking: a survey of models based on deep learning. https://arxiv.org/abs/2006.00575

Shen W, Wang JY, Han JW, 2015. Entity linking with a knowledge base: issues, techniques, and solutions. *IEEE Trans Knowl Data Eng*, 27(2):443-460. https://doi.org/10.1109/TKDE.2014.2327028

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Wang HF, Liu ZQ, 2019. An error recognition method for power equipment defect records based on knowledge graph technology. *Front Inform Technol Electron Eng*, 20(11):1564-1577. https://doi.org/10.1631/FITEE.1800260

Wang Z, Zhang JW, Feng JL, et al., 2014. Knowledge graph and text jointly embedding. Proc Conf on Empirical Methods in Natural Language Processing, p.1591-1601. https://doi.org/10.3115/v1/D14-1167

Wu JS, Zhang RC, Mao YY, et al., 2020. Dynamic graph convolutional networks for entity linking. Proc Web Conf, p.1149-1159. https://doi.org/10.1145/3366423.3380192

Xin KX, Hua W, Liu Y, et al., 2019. Entity disambiguation based on parse tree neighbours on graph attention network. Proc 20th Int Conf on Web Information Systems Engineering, p.523-537. https://doi.org/10.1007/978-3-030-34223-4_33

Xu P, Barbosa D, 2018. Neural fine-grained entity type classification with hierarchy-aware loss. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.16-25. https://doi.org/10.18653/v1/N18-1002

Xue MG, Cai WM, Su JS, et al., 2019. Neural collective entity linking based on recurrent random walk network learning. Proc 28th Int Joint Conf on Artificial Intelligence, p.5327-5333. https://doi.org/10.24963/ijcai.2019/740

Yamada I, Shindo H, Takeda H, et al., 2016. Joint learning of the embedding of words and entities for named entity disambiguation. Proc 20th SIGNLL Conf on Computational Natural Language Learning, p.250-259. https://doi.org/10.18653/v1/K16-1025

Yang XY, Gu XT, Lin S, et al., 2019. Learning dynamic context augmentation for global entity linking. Proc Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing, p.271-281. https://doi.org/10.18653/v1/D19-1026

Zhang HB, Chen Q, Zhang WW, et al., 2022. HSIE: improving named entity disambiguation with hidden semantic information extractor. Proc 14th Int Conf on Machine Learning and Computing, p.251-257. https://doi.org/10.1145/3529836.3529920

Zhong HP, Zhang JW, Wang Z, et al., 2015. Aligning knowledge and text embeddings by entity descriptions. Proc Conf on Empirical Methods in Natural Language Processing, p.267-272. https://doi.org/10.18653/v1/D15-1031