Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning

Mengge Xue[1,2,3] , **Weiming Cai**[1] , **Jinsong Su**[1*] , **Linfeng Song**[4] , **Yubin Ge**[4] ,
Yubao Liu[4] and **Bin Wang**[5]
[1]Xiamen University
[2]Institute of Information Engineering, Chinese Academy of Sciences
[3]School of Cyber Security, University of Chinese Academy of Sciences
[4]Rochester University
[5]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
xuemengge@iie.ac.cn, caiweiming@stu.xmu.edu.cn, jssu@xmu.edu.cn

Abstract

Benefiting from the excellent ability of neural networks on learning semantic representations, existing studies for entity linking (EL) have resorted to neural networks to exploit both the local mention-to-entity compatibility and the global interdependence between different EL decisions for target entity disambiguation. However, most neural collective EL methods depend entirely upon neural networks to automatically model the semantic dependencies between different EL decisions, which lack of the guidance from external knowledge. In this paper, we propose a novel end-to-end neural network with recurrent random-walk layers for collective EL, which introduces external knowledge to model the semantic interdependence between different EL decisions. Specifically, we first establish a model based on local context features, and then stack random-walk layers to reinforce the evidence for related EL decisions into high-probability decisions, where the semantic interdependence between candidate entities is mainly induced from an external knowledge base. Finally, a semantic regularizer that preserves the collective EL decisions consistency is incorporated into the conventional objective function, so that the external knowledge base can be fully exploited in collective EL decisions. Experimental results and in-depth analysis on various datasets show that our model achieves better performance than other state-of-the-art models. Our code and data are released at https://github.com/DeepLearnXMU/RRWEL.

1 Introduction

Entity linking (EL) is a task to link the name mentions in a document to their referent entities within a knowledge base (KB). The great significance of the research on EL can not be neglected due to the solid foundation it helps to build for multiple natural language processing tasks, such as information extraction [Ji and Nothman, 2016], semantic search [Blanco *et al.*, 2015] and so on. Nevertheless, this task is non-trivial because mentions are usually ambiguous, and the inherent disambiguation between mentions and their referent entities still maintains EL as a challenging task.

The previous studies on EL are based on the statistical models, which mainly focus on artificially defined discriminative features, which mainly focus on the utilization of artificially defined discriminative features between mentions and its target entities, such as contextual information, topic information or entities type etc. [Chisholm and Hachey, 2015; Shen *et al.*, 2015]. However, these models resolve mentions independently relying on textual context information from the surrounding words while ignoring the interdependence between different EL decisions. In order to address this problem, researches then devoted themselves to implementing collective decisions in EL, which encourages re entities of all mentions in a document to be semantically coherent [Hoffart *et al.*, 2011; Ganea *et al.*, 2015; Globerson *et al.*, 2016; Guo and Barbosa, 2018].

Recently, the studies of EL have evolved from the conventional statistical models into the neural network (NN) based models thanks to their outstanding advantages in encoding semantics and dealing with data sparsity. Similarly, the studies of NN-based EL models have also experienced the development progress from models of independent decisions [He *et al.*, 2013; Francis-Landau *et al.*, 2016; Yamada *et al.*, 2016] to those of collective decisions [Ganea and Hofmann, 2017; Cao *et al.*, 2018; Le and Titov, 2018; Kolitsas *et al.*, 2018]. For example, Ganea et al., [2017] and Le et al., [2018] solved the global training problem via a truncated fitting loopy belief propagation (LBP). Cao et al., [2018] applied Graph Convolutional Network (GCN) to integrate global coherence information for EL. However, in these works, they completely depend on NNs to automatically model the semantic dependencies between different EL decisions, while little attention has been paid on the guidance from an external KB.

In this paper, we propose a novel end-to-end neural collective model named Recurrent Random Walk based EL (RRWEL) which not only implements collective EL decisions

[*] Corresponding author

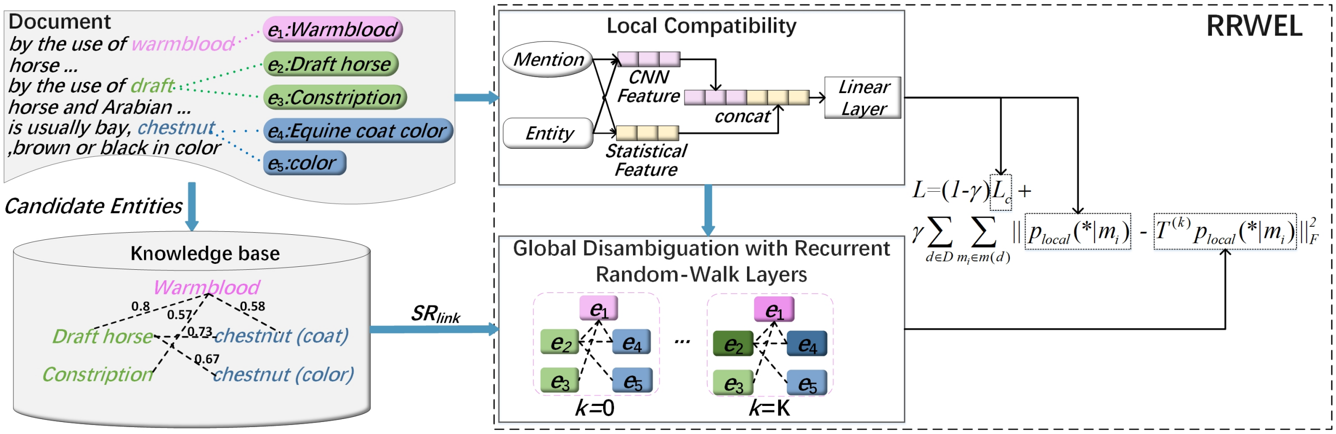

Figure 1: The architecture of our proposed RRWEL model.

but also leverages an external KB to model the semantic dependencies between candidate entities. Given a document, we first utilize some local features to measure the conditional probability of a mention referring to a specific entity, and then introduce random-walk layers to model the interdependence between different EL decisions, where the evidence for related EL decisions are computed based on the semantic relatedness between their corresponding knowledge pages, and can be fine-tuned during training. Finally, a semantic regularizer, which aims to preserve the collective EL decision consistency, is added to supplement the conventional EL loss. In that case, both the local semantic correspondence between mentions and entities and the global interdependence between different EL decisions can be fully exploited in our model to conduct collective EL decisions.

The main contributions of our work can be summarized as follows: (1) Through the in-depth analysis, we point out the drawbacks of dominant NN-based EL models and then dedicate to the study of employing KB based global interdependence between EL decisions for collective entity disambiguation; (2) We propose a novel end-to-end graph-based NN model which incorporates the external KB to collectively infer the referent entities of all mentions in the same document. To the best of our knowledge, this is the first NN-based random walk method for collective EL to explore external KB when modeling semantic interdependence between different EL decisions; (3) We evaluate the performance of our proposed model on many large-scale public datasets. The extensive experiments and results show that our model can outperform several state-of-the-art models.

2 The Proposed Model

In this section, we give a detailed description to our proposed model. As shown in Fig. 1, our model mainly includes two components: one is used to model local mention-to-entity compatibility, the other leverages external knowledge to capture global entity-to-entity semantic interdependence.

More specifically, we first employ CNNs to learn the semantic representations of each mention and its all candidate entities, which are further fed into a linear model layer with other feature vectors to measure their local semantic compatibility. Then, we introduce random-walk layers to model the global interdependence between different EL decisions, which enable the EL evidence for related mentions to be collectively reinforced into high-probability EL decisions. In particular, during model training, in order to effectively exploit both the local mention-to-entity compatibility and the global interdependence between entities for collective ELs, we augment the conventional EL loss with a semantic regularizer.

To clearly describe our model, we first introduce some annotations which are used in our model. Let d be the input document and $m(d) = \{m_1, m_2, ..., m_{N_{md}}\}$ be the set of mentions occurring in d, for each mention m_i, its candidates set $\Gamma(m_i)$ is identified. In addition, for all mentions $\{m_i\}$, we collect all their candidate entities to form the candidate entity set $e(d) = \{e_1, e_2, ..., e_{N_{ed}}\}$, where each entity e_j has a corresponding KB page p_j. Using EL models, we expect to correctly map each mention to its corresponding entity page of the referenced KB.

2.1 Local Mention-to-Entity Compatibility

Following Francis-Landau et al., [2016], we represent the mention m_i using the semantic information at three kinds of granularities: s_i, c_i, d_i, where s_i is the *surface string* of m_i, c_i is the *immediate context* within a predefined window of m_i and d_i is the *entire document* contains m_i. For the candidate entity page p_j, we use its *title* t_j and *body content* b_j to represent it, $p_j = (t_j, b_j)$.

To correctly implement EL decisions, we first calculate the relevance score $\phi(m_i, p_j)$ for the candidate entity page p_j, and then normalize relevance scores to obtain the conditional probability $p_{local}(p_j|m_i)$:

$$p_{local}(p_j|m_i) = \frac{\phi(m_i, p_j)}{\sum_{j' \in \Gamma(m_i)} \phi(m_i, p_{j'})}, \qquad (1)$$

where $\phi(m_i, p_j)$ is defined as follows:

$$\phi(m_i, p_j) = \sigma(W_{\text{local}}[F_{\text{sf}}(m_i, p_j); F_{\text{cnn}}(m_i, p_j)]), \qquad (2)$$

where $\sigma(*)$ is the sigmoid function, $F_{\text{sf}}(m_i, p_j)$ and $F_{\text{cnn}}(m_i, p_j)$ are two feature vectors that will be concatenated to obtain the local feature vectors, and W_{local} is the weight for local feature vectors. In the following, we give detail descriptions of $F_{\text{sf}}(m_i, p_j)$ and $F_{\text{cnn}}(m_i, p_j)$.

1). $F_{\text{sf}}(m_i, p_j)$ denotes the statistical local feature vector proposed by Cao et al. [2018]. It contains various linguistic properties and statistics, including candidates' prior probability, string based similarities, compatibility similarities and the embeddings of contexts, all of which have been proven to be effective for EL. Due to the space limitation, we omit the detail descriptions of $F_{\text{sf}}(m_i, p_j)$. Please refer to [Cao *et al.*, 2018] for the details.

2). $F_{\text{cnn}}(m_i, p_j)$ combines the cosine similarities between the representation vectors of m_i and p_j at multiple granularities. Formally, it is defined as follows:

$$\begin{aligned} F_{\text{cnn}}(m_i, p_j) = [&\cos(\bar{s}_i, \bar{t}_j), \cos(\bar{c}_i, \bar{t}_j), \cos(\bar{d}_i, \bar{t}_j), \\ &\cos(\bar{s}_i, \bar{b}_j), \cos(\bar{c}_i, \bar{b}_j), \cos(\bar{d}_i, \bar{b}_j)], \end{aligned} \quad (3)$$

where $\bar{s}_i$, $\bar{c}_i$, $\bar{d}_i$, $\bar{t}_j$, and $\bar{b}_j$ denote the distributed representations of s_i, c_i, d_i, t_j, and b_j, respectively.

In order to obtain the representations for the context word sequences of mentions and entities, we follow Francis-Landau et al., [2016] and adopt CNNs to transform a context word sequence $x \in \{s_i, c_i, d_i, t_j, b_j\}$ into the distributed representations, where each word of x is represented by a real-valued, h-dimensional vector. The convolution operation is first performed on w, where w=$[w_1; w_2; ...; w_N] \in R^{h \times N}$ is a matrix representing x. Here N is the word number of x. Then we transform the produced hidden vector sequences by a non-linear function $G(*)$ and use *Avg* function as pooling. Specifically, we employ one window size l to parameterize the convolution operation, where the window size l corresponds to a convolution matrix $M_l \in R^{v \times l \times h}$ of dimensionality v. Hence, we obtain the distributed representation $\bar{x}$ for x as $\frac{1}{N-l}\sum_{i=1}^{N-l+1} G(M_l w_{i:(i+l-1)})$. Please note that we use the same set of convolution parameters for each type of text granularity in the source document d as well as for the target entities.

At this stage, we are able to calculate $p_{local}(p_j|m_i)$ which provides abundant information of local mention-to-entity compatibility. The next step is to propagate this information based on the global interdependence between EL decisions.

2.2 Global Interdependence Between EL Decisions

Inspired by the random walk based EL [Han *et al.*, 2011] and the successful adaptation of random walk propagation to N-N [Zhao *et al.*, 2017], we introduce recurrent random-walk layers to propagate EL evidence for the purpose of effectively capturing the global interdependence between different EL decisions for collective entity predictions.

To implement the propagation of EL evidence based on random walk, we first need to define a transition matrix T between candidate entities, where T_{ij} is the evidence propagation ratio from e_j to e_i. Intuitively, the more semantically related two entities are, the more evidence should be propagated between their EL decisions. Therefore, we calculate the semantic relevance between e_i and e_j, and then normalize these relevance scores by entity to generate T:

$$T(i,j) = \frac{p(e_i \rightarrow e_j)}{\sum_{j' \in N_{e_i}} p(e_i \rightarrow e_{j'})}, \quad (4)$$

$$p(e_i \rightarrow e_j) = \text{SR}_{\text{link}}(p_i, p_j) + \text{SR}_{\text{semantic}}(p_i, p_j), \quad (5)$$

where N_{e_i} is the set of neighbor entities of entity e_i. To be specific, for two considered entity pages $p_i = (t_i, b_i)$ and $p_j = (t_j, b_j)$, we use hyperlinks to compute the semantic relevance score $\text{SR}_{\text{link}}(p_i, p_j)$:

$$\text{SR}_{\text{link}}(p_i, p_j) = 1 - \frac{log(max(|I|, |J|)) - log(|I \cap J|)}{log(|W|) - log(min(|I|, |J|))}, \quad (6)$$

where I and J are the sets of all entities that link to p_i and p_j in KB respectively, and W is the entire KB. Meanwhile, we obtain their cosine similarity $\text{SR}_{\text{semantic}}(p_i, p_j)$ based on their CNN semantic representations. Considering that the semantic relevance score between two candidate entities relies on the relative distance of their corresponding mentions, we supplement the conventional entity page $p_i = (t_i, b_i)$ with the position embedding pos_i of its entity. Formally, $\text{SR}_{\text{semantic}}(p_i, p_j)$ is defined as follows:

$$\text{SR}_{\text{semantic}}(p_i, p_j) = \cos([\bar{t}_i; \bar{e}_i] + pos_i, [\bar{t}_j; \bar{e}_j] + pos_j). \quad (7)$$

Here we follow Vaswani et al., [2017] to define the embedding pos_i of mention m_i.

With the transition matrix T, we perform the evidence propagation of EL decisions in the recurrent random-walk layers: (See the recurrent random-walk layers of Fig. 1)

$$\begin{aligned} p_{rw}^{(k+1)}(*|m_i) &= (1-\lambda)T \cdot p_{rw}^{(k)}(*|m_i) + \lambda p_{rw}^{(0)}(*|m_i) \\ &= (1-\lambda)T^{(k)} \cdot p_{rw}^{(0)}(*|m_i) + \lambda p_{rw}^{(0)}(*|m_i) \\ &= (1-\lambda)T^{(k)} \cdot p_{local}(*|m_i) + \lambda p_{local}(*|m_i), \end{aligned} \quad (8)$$

with $p_{rw}^{(k)}(*|m_i)$ being the predicted entity distribution of m_i at the k-th iteration. Please note that $p_{rw}^{(0)}(*|m_i)$ = $p_{local}(*|m_i)$ which only exploits local mention-to-entity compatibility. Obviously, by introducing K random-walk layers, we can easily propagate evidence for K times based on random walk propagation.

2.3 Model Training

Aiming at combining the global interdependence between EL decisions with the local mention-to-entity context compatibility, we propose to not only minimize the common compatibility-based EL loss but also preserve the high-order EL consistency during model training. In this way, we can significantly improve our model by embedding the global interdependence between different EL decisions into the training of CNNs for modeling mention-to-entity compatibility.

The intuition behind our implementation lies in the convergence propriety of random walk process based on Markov chain [Gilks *et al.*, 1995]. Specifically, after multiple rounds of EL evidence propagation, the predicted entity distributions for mentions will tend to converge. If the global interdependence between different EL decisions has been well

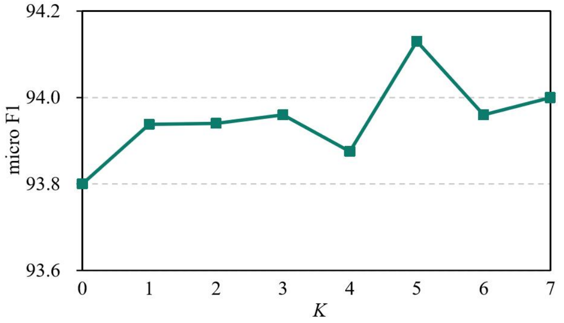

Figure 2: Experimental results on the validation set AIDA-A using different numbers of random-walk layers.

embedded into our model, we can expect that $p_{local}(*|m_i) \approx p_{rw}^{(K)}(*|m_i) = T^{(K)} \cdot p_{local}(*|m_i)$. To do this, we minimize the following error to preserve the the K-th order EL consistency of m_i, formulated as $\|p_{local}(*|m_i) - T^{(K)} \cdot p_{local}(*|m_i)\|_F^2$, where $\|\cdot\|_F^2$ is the Frobenius norm.

Finally, the recurrent random walk network learning for collective EL can be mathematically formulated as

$$L = (1-\gamma)\cdot L_c + \gamma \cdot \sum_{d\in D}\sum_{m_i\in m(d)} \|p_{local}(*|m_i) - T^{(K)} \cdot p_{local}(*|m_i)\|_F^2, \tag{9}$$

where D denotes the training corpus, K is the number of random-walk layers, and the coefficient γ is used to balance the preference between the two terms. Specifically, the first term L_c denotes the cross-entropy loss between predicted and ground truth y^g, which is defined as follows:

$$L_c = -\sum_{d\in D}\sum_{m_i\in m(d)}\sum_{e_j\in\Gamma(m_i)} y_j^g log p_{local}(e_j|m_i), \tag{10}$$

To train the proposed model, we denote all the model parameters by θ. Therefore, the objective function in our learning process is given by

$$\min L(\theta) = L + \alpha\|\theta\|^2, \tag{11}$$

where α is the trade-off parameter between the training loss L and regularizer $\|\theta\|^2$. To optimize this objective function, we employ the stochastic gradient descent (SGD) with diagonal variant of *AdaGrad* in [Duchi *et al.*, 2011]. Particularly, when training the model, we first use L_c to pre-train the model and then use L to fine-tune it.

3 Experiment

3.1 Setup

Datasets

We validated our proposed model on six different benchmark datasets used by previous studies:

- **AIDA-CONLL**: This dataset is a manually annotated EL dataset [Hoffart *et al.*, 2011]. It consists of AIDA-train for training, AIDA-A for validation and AIDA-B for testing and there are totally 946, 216, 231 documents respectively.

γ \ λ	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
0.1	93.64	93.9	**94.13**	93.92	94.00
0.3	92.64	93.49	93.47	93.22	93.3
0.5	90.94	91.51	91.51	91.13	92.58
0.7	85.35	85.67	88.15	87.43	89.03
0.9	69.72	69.48	69.03	68.57	68.91

Table 1: Experimental results on the validation set AIDA-A using different trade-off parameters.

Model	**AIDA-B**
Yamada et al., [2016]	91.5
Francis-Landau et al., [2016]	86.9
Ganea and Hofmann [2017]	92.22
Cao et al., [2018]	87.2
Guo and Barbosa [2018]	89.0
Kolitsas et al., [2018]	89.1
Le and Titov [2018]	**93.07**
RRWEL	92.36

Table 2: Micro F1 scores for AIDA-B (in-domain) test set.

- **MSNBC, AQUAINT, ACE2004**: These datasets are cleaned and updated by Guo and Barbosa [2018], which contain 20, 50 and 36 documents respectively.

- **WNED-WIKI (WIKI), WNED-CWEB (CWEB)**: These datasets are automatically extracted from ClueWeb and Wikipedia in [Guo and Barbosa, 2018; Gabrilovich *et al.*, 2013] and are relatively large with 320 documents each.

In our experiments, we investigated the system performance with AIDA-train for training and AIDA-A for validation, and then tested on AIDA-B and other datasets. Following previous works [Ganea and Hofmann, 2017; Cao *et al.*, 2018; Le and Titov, 2018], we only considered mentions that have candidate entities in the referenced KB.

Contrast Models

We compared our proposed RRWEL model to the following models:

- **[Hoffart *et al.*, 2011]** where iterative greedy method is employed to compute a subgraph with maximum density for EL.

- **[Han *et al.*, 2011]** introduces random walk algorithm to implement collective entity disambiguation.

- **[Cheng and Roth, 2013]** utilizes integer linear programming to solve global entity linking.

- **[Francis-Landau *et al.*, 2016]** applies CNN to learn the semantic representations of each mention and its all candidate entities.

- **[Yamada *et al.*, 2016]** proposes a novel embedding method specifically designed for NED which jointly maps words and entities into a same continuous vector space.

Model	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Hoffart et al., [2011]	79	56	80	58.6	63	67.32
Han et al., [2011]	88	79	73	61	78	75.8
Cheng and Roth [2013]	90	90	86	67.5	73.4	81.38
Ganea and Hofmann [2017]	93.7	88.5	88.5	77.9	77.5	85.22
Le and Titov [2018]	93.9	88.3	89.9	77.5	78.0	85.51
Guo and Barbosa [2018]	92	87	88	77	84.5	85.7
RRWEL	**94.43**	**91.94**	**90.64**	**79.65**	**85.47**	**88.43**

Table 3: Performance (**Micro F1**) of various EL models on different datasets (out-domain). Particularly, we highlight the highest score in bold for each set.

The ideal Wrttemberger stands around high, and is usually bay, chestnut, brown, or black in color				
Ground-truth Entity	RRWEL(K=0)		RRWEL(K=5)	
chestnut (color)	*chestnut (color)*	0.996	*chestnut (color)*	0.764
Bay (horse)	*Bay (horse)* *Brown*	0.031 0.949	*Bay (horse)* *Brown*	0.513 0.447
Equine coat color	*Equine coat color* *Color*	0.273 0.483	*Equine coat color* *Color*	0.501 0.376

Table 4: An example of the predicted entity distribution using different numbers of the recurrent random-walk layers

Model	AIDA-B	MSNBC	AQUAINT	ACE2004	CWEB	WIKI
RRWEL(NN learning)	91.32	93.14	90.90	90.21	78.45	84.37
RRWEL(SR$_{\text{link}}$)	92.00	93.90	91.35	**91.06**	**80.14**	84.95
RRWEL	**92.36**	**94.43**	**91.94**	90.64	79.65	**85.47**

Table 5: Performance (**Micro F1**) under the effect of different transition matrixs on test datasets. Specifically, "RRWEL(NN learning)" indicates that the transition matrix is automatically modeled by NN, while "RRWEL(SR$_{\text{link}}$)" illustrates that we only use hyperlink information to initialize the transition matrix.

- **[Ganea and Hofmann, 2017]** solves the global training problem via truncated fitting LBP.
- **[Le and Titov, 2018]** encodes the relations between mentions as latent variables and applies LBP for global training problem.
- **[Guo and Barbosa, 2018]** proposes a greedy, global NED algorithm which utilizes the mutual information between probability distributions induced from random walk propagation on the disambiguation graph.
- **[Cao *et al.*, 2018]** applies Graph Convolutional Network to integrate both local contextual features and global coherence information for EL.

Model Details

We used the latest English Wikipedia dump[1] as our referenced KB. However, please note that our proposed model can be easily applied to other KBs. To employ CNN to learn the distributed representations of inputs, we used 64 filters with the window size 3 for the convolution operation and the non-linear transformation function *ReLU*. Meanwhile, to learn the context representations of input mentions and target entities, we directly followed Francis-Landau [2016] to utilize the window size 10 for context, and only extracted the first 100 words in the documents for mentions and entities. Besides, the standard *Word2Vec* toolkit [Mikolov *et al.*, 2013] was used with vector dimension size 300, window context size 21, negative sample number 10 and iteration number 10 to pre-train word embeddings on Wikipedia and then we fine-tuned them during model training. Particularly, following Ganea and Hofmann [2017], we kept top 7 candidates for each mention based on their prior probabilities, and while propagating the evidence, we just kept top 4 candidates for each mention m_i according to $p_{local}(*|m_i)$. Besides, we set α=1e−5. Finally, we adopted standard F1 score at Mention level (Micro) as measurement.

3.2 Effects of K and Trade-off Parameters

There are three crucial parameters in our approach, which are the number K of random-walk layers, the trade-off parameter λ for restart and the trade-off parameter γ for loss function. We tried different hyper-parameters according to the performance of our model on the validation set.

To investigate the effect of the trade-off parameter λ and γ, we kept K=5 and then tried different combinations of these two parameters with λ from 0.1 to 0.9 and γ from 0.1 to 0.9. The results are shown in Table 1. We observed that our method achieves the best performance when the trade-off parameters λ and γ are set to 0.5 and 0.1 respectively. Consequently, all the following experiments only consider the proposed model with λ=0.5 and γ=0.1.

Besides, we varied the value of K from 0 to 7 with an increment of 1 each time while keeping λ=0.5 and γ=0.1.

[1] https://dumps.wikimedia.org/enwiki/20181101/enwiki-20181101-pages-articles-multistream.xml.bz2

Please note that when K is set as 0, our model is unable to model global interdependence between different EL decisions. Fig. 2 provides the experimental results. We find that our model ($K{>}0$) obviously outperforms its variant with $K{=}0$. This results indicates that graph-based evidence propagation contributes effectively to the improvement of NN-based EL method. In particular, our model achieves the best performance when $K{=}5$, so we set $K{=}5$ for all experiments.

3.3 Overall Results

Experimental results are shown in Tables 2 and 3. We directly cited the experimental results of the state-of-the-art models reported in [Francis-Landau *et al.*, 2016; Ganea and Hofmann, 2017; Le and Titov, 2018; Guo and Barbosa, 2018]. On average, our model achieves the highest Micro F1 scores on all out-domain datasets, expect AIDA-B test set (in-domain). It is worth noting that although our model is slightly inferior to [Le and Titov, 2018] on AIDA-B, however, its latent relations modeling can be adapted to refine our model.

3.4 Qualitative Analysis

Effects of Recurrent Random-Walk Layers

In table 4, we show a hard case (where entities of two mentions are incorrectly predicted because of their low evidence scores), which is from WNED-WIKI dataset, can be correctly solved by our random-walk layers. We can notice that RRWEL($K{=}0$) is unable to find enough disambiguation clues from context words for the two mentions, specifically, *"brown"*, *"color"*. By contrast, RRWEL can effectively identify the correct entities with the help of the semantic interdependence based on KB: *"Bay (horse)"* and *"Equine coat color"* receive higher evidence scores than others.

Effects of the External KB

In order to better understand how our model benefits from the graph-based evidence propagation based on external knowledge, we consider different kinds of transition matrix T and the performances are represented in Table 5. We can see that by leveraging external knowledge to model the semantic dependencies between candidate entities, the global interdependence between different EL decisions can be more accurately captured by our model to conduct collective EL decisions.

4 Related Work

In early studies, the dominant EL models are mostly statistical ones exploring manually defined discriminative features to model the local context compatibility [Ji and Grishman, 2011; Mendes *et al.*, 2011; Shen *et al.*, 2015] and the global interdependence between EL decisions [Hoffart *et al.*, 2011; Cheng and Roth, 2013; Ganea *et al.*, 2015; Globerson *et al.*, 2016; Guo and Barbosa, 2018]. With the extensive applications of deep learning in NLP, the studies of EL have marched into NN related research from resorting to conventional statistical models. Similar to the precious statistical models, the early NN-based models mainly focused on how to apply NNs to measure the local context compatibility. For example, He et al., [2013] employed Stacked Denoising Auto-encodes to learn entity representations, Francis-Landau et al., [2016] combined CNN-based representations with sparse features to model mentions and entities, while Yamada et al., [2016] proposed a NN model to learn distributed representations for texts and KB entities. Obviously, these works only focus on individual EL tasks with little attention paid on the interdependence between these target decisions. To tackle this problem, Ganea et al., [2017] utilized unrolled deep LBP network to model global information. Le and Titov [2018] treated relations between textual mentions in a document as latent variables in our neural EL model. Moreover, GCN was applied by Cao et al., [2018] for the sake of integrating global coherence information for EL. In these works, notwithstanding semantic dependencies between various entities are able to be automatically modeled by constructing a neural network, the guidance from an external KB has always been neglected.

To address the above-mentioned issue, in this work, we exploit a neural network with recurrent random-walk layers leveraging external knowledge for collective EL. The most related works to ours include [Han *et al.*, 2011; Huu *et al.*, 2016; Ganea and Hofmann, 2017; Guo and Barbosa, 2018; Cao *et al.*, 2018]. Significantly different from [Han *et al.*, 2011] and [Guo and Barbosa, 2018], we resort to NN rather than statistical models for collective EL. Compared with [Huu *et al.*, 2016], we further adopt the mutual instead of unidirectional effects between EL decisions. Also, different from [Ganea and Hofmann, 2017; Cao *et al.*, 2018], we explore external KB to model global semantic interdependence between different entities, which has been proved to be more effective in our experiments.

5 Conclusions and Future Work

This paper has presented a novel end-to-end neural network with recurrent random-walk layers for collective EL, which reinforce the evidence for related EL decisions into high-probability decisions with the help of external KB. Experimental results and in-depth analysis strongly demonstrate the effectiveness of our proposed model.

Our model is generally applicable to other tasks similar to EL, such as word sense disambiguation, cross-lingual entity disambiguation, and lexical selection [Su *et al.*, 2015]. Therefore, we will investigate the effectiveness of our approach on these tasks. Besides, we would like to fully exploit other resources of Wikipedia beyond hyperlinks to refine our model. Finally, inspired by the recent success of graph neural network in NLP [Zhang *et al.*, 2018; Song *et al.*, 2019], we plan to explore graph neural network based EL model in the future.

Acknowledgments

The authors were supported by Beijing Advanced Innovation Center for Language Resources, National Natural Science Foundation of China (No. 61672440), the Fundamental Research Funds for the Central Universities (Grant No. ZK1024), Scientific Research Project of National Language Committee of China (Grant No. YB135-49), National Key Research and Development Program of China (grant No. 2016QY03D0505) and Project S201910384407 supported by XMU Training Program of Innovation and Enterpreneurship for Undergraduates.

References

[Blanco *et al.*, 2015] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. Fast and space-efficient entity linking for queries. In *WSDM*, pages 179–188, 2015.

[Cao *et al.*, 2018] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. Neural collective entity linking. In *COLING*, pages 675–686, 2018.

[Cheng and Roth, 2013] X. Cheng and D. Roth. Relational inference for wikification. In *EMNLP*, pages 1787–1796, 2013.

[Chisholm and Hachey, 2015] Andrew Chisholm and Ben Hachey. Entity disambiguation with web links. *TACL*, pages 145–156, 2015.

[Duchi *et al.*, 2011] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, pages 2121–2159, 2011.

[Francis-Landau *et al.*, 2016] Matthew Francis-Landau, Greg Durrett, and Klein Dan. Capturing semantic similarity for entity linking with convolutional neural networks. In *NAACL*, pages 1256–1261, 2016.

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

[Ganea and Hofmann, 2017] Octavian Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In *EMNLP*, pages 2619–2629, 2017.

[Ganea *et al.*, 2015] Octavian Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. Probabilistic bag-of-hyperlinks model for entity linking. In *WWW*, pages 927–938, 2015.

[Gilks *et al.*, 1995] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC press, 1995.

[Globerson *et al.*, 2016] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. Collective entity resolution with multi-focal attention. In *ACL*, pages 621–631, 2016.

[Guo and Barbosa, 2018] Zhaochen Guo and Denilson Barbosa. Robust named entity disambiguation with random walks. *Semantic Web*, pages 459–479, 2018.

[Han *et al.*, 2011] Xianpei Han, Le Sun, and Jun Zhao. Collective entity linking in web text:a graph-based method. In *SIGIR*, pages 765–774, 2011.

[He *et al.*, 2013] Z. He, S. Liu, Y. Song, M. Li, M. Zhou, and H. Wang. Efficient collective entity linking with stacking. In *EMNLP*, pages 30–34, 2013.

[Hoffart *et al.*, 2011] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen , Furstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792, 2011.

[Huu *et al.*, 2016] Thien Huu, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi. Joint learning of local and global features for entity linking via neural networks. In *COLING*, pages 2310–2320, 2016.

[Ji and Grishman, 2011] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *ACL*, pages 1148–1158, 2011.

[Ji and Nothman, 2016] Heng Ji and Joel Nothman. Overview of TAC-KBP2016 tri-lingual EDL and its impact on end-to-end KBP. In *TAC*, 2016.

[Kolitsas *et al.*, 2018] Nikolaos Kolitsas, Octavianeugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *CoNLL*, pages 519–529, 2018.

[Le and Titov, 2018] Phong Le and Ivan Titov. Improving entity linking by modeling latent relations between mentions. In *ACL*, pages 1595–1604, 2018.

[Mendes *et al.*, 2011] Pablo N. Mendes, Max Jakob, and Christian Bizer. Dbpedia spotlight:shedding light on the web of documents. In *I-Semantics*, pages 1–8, 2011.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[Shen *et al.*, 2015] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans Knowl Data Eng*, pages 443–460, 2015.

[Song *et al.*, 2019] Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. Semantic neural machine translation using amr. *arXiv preprint arXiv:1902.07282*, 2019.

[Su *et al.*, 2015] Jinsong Su, Deyi Xiong, Shujian Huang, Xianpei Han, and Junfeng Yao. Graph-based collective lexical selection for statistical machine translation. In *EMNLP*, pages 1238–1247, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Yamada *et al.*, 2016] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *CoNLL*, pages 250–259, 2016.

[Zhang *et al.*, 2018] Yue Zhang, Qi Liu, and Linfeng Song. Sentence-state LSTM for text representation. In *ACL*, pages 317–327, 2018.

[Zhao *et al.*, 2017] Zhou Zhao, Hanqing Lu, Deng Cai, Xiaofei He, and Yueting Zhuang. Microblog sentiment classication via recurrent random walk network learning. In *IJCAI*, pages 3532–3538, 2017.