A Bidirectional Multi-paragraph Reading Model for Zero-shot Entity Linking

Hongyin Tang[1,2,*], **Xingwu Sun**[3], **Beihong Jin**[1,2,†], **Fuzheng Zhang**[3]
[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, Beijing China
[3]Meituan-Dianping Group, China
tanghongyin14@otcaix.iscas.ac.cn
beihong@iscas.ac.cn
{sunxingwu,zhangfuzheng}@meituan.com

Abstract

Recently, a zero-shot entity linking task is introduced to challenge the generalization ability of entity linking models. In this task, mentions must be linked to unseen entities and only the textual information is available. In order to make full use of the documents, previous work has proposed a BERT-based model which can only take fixed length of text as input. However, the key information for entity linking may exist in nearly everywhere of the documents thus the proposed model cannot capture them all. To leverage more textual information and enhance text understanding capability, we propose a bidirectional multi-paragraph reading model for the zero-shot entity linking task. Firstly, the model treats the mention context as a query and matches it with multiple paragraphs of the entity description documents. Then, the mention-aware entity representation obtained from the first step is used as a query to match multiple paragraphs in the document containing the mention through an entity-mention attention mechanism. In particular, a new pre-training strategy is employed to strengthen the representative ability. Experimental results show that our bidirectional model can capture long-range context dependencies and outperform the baseline model by 3-4% in terms of accuracy.

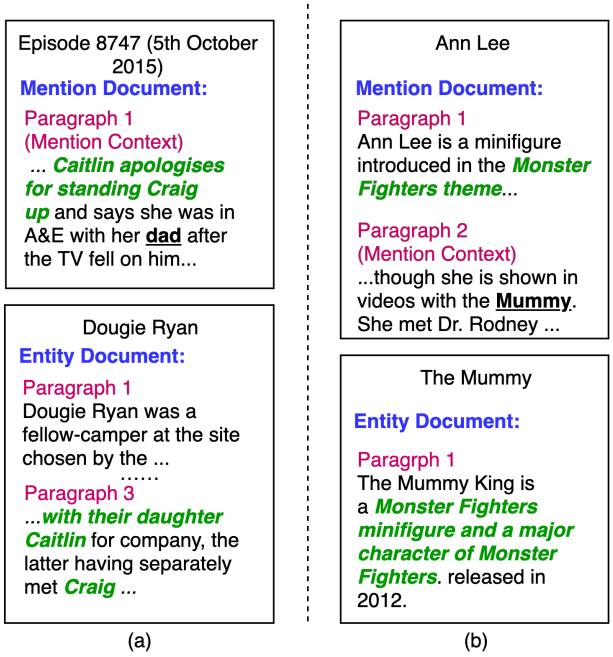

Figure 1: Left: (a) Evidence exists in other paragraphs of entity document. Right: (b) Evidence exists in other paragraphs of mention document. Underlined texts are mentions. Green italic texts are the evidences for linking.

Introduction

Entity Linking (EL) is a task of resolving ambiguous mentions to their referent entities in a knowledge base (KB). EL is a fundamental task in the area of information extraction (IE) and can benefit other NLP applications such as question answering (Chang 2016), text summarization (Amplayo, Lim, and Hwang 2018), etc.

Most previous work focuses on linking mentions to general KBs (e.g. Wikipedia). They usually train models under a setting where the entities occurring in the test set are partially or fully available for training. Moreover, they typically utilize not only textual information but also powerful resources like frequency statistics and meta-data (Ganea and Hofmann 2017; Roth et al. 2014) existing in the KBs. Compared to general domains, entity linking in specialized domains such as movies or fictions, needs to be addressed with a strong demand. However, this type of linking can be much more challenging in such domains since labeled data are not easily obtained and the resources in such KBs are not as abundant as the general ones. Therefore, models with the capability of generalizing to new domains need to be developed.

To evaluate the domain generalization of entity linking systems, Lajanugen et al.(Logeswaran et al. 2019) presented a zero-shot entity linking task which has two key properties:

[*]The work was done when the first author was an intern at Meituan Group.
[†]Corresponding Author

(1) no mentions or entities in the testing have been observed during training, (2) no other information is available other than texts. In this task, each mention is extracted from a document and each entity corresponds to a document that describes the details of it. Meanwhile, they proposed a baseline model based on BERT (Devlin et al. 2019) which links the mention with its referent entity by reading two pieces of text, i.e., the mention context and the entity description. Specifically, the baseline model truncates a fixed length of text around the mention as the mention context and the beginning of entity document as the entity description. Then, it concatenates the two pieces of texts and feeds it into BERT which is able to perform deep cross-attention between their tokens, producing a matching score indicating how much the candidate entity is related to the mention. However, this model cannot fully capture the coherence evidence between a mention and its golden entity due to the limitation of its input length.

To illustrate why it is insufficient, we give two examples in Figure 1. The mention contexts and entity descriptions are extracted as the baseline model does. Then, we split the rest texts of the documents into paragraphs whose lengths are same as them. In the left example, the mention is "dad" whose daughter "Caitlin apologies for standing Craig up". The golden entity of the mention is "Doubie Ryan" since the paragraph 3 of the entity document shows "Caitlin" is "Doubie Ryan"'s daughter who met "Craig". Obviously, the BERT-based baseline model may not link the entity correctly since the paragraph 3 is not fed in. In the right example, the golden entity of the mention "Mummy" is "in the Monster Fighters theme". However, the evidence exists in paragraph 1 of the mention document which is not presented in the mention context, thus the baseline model would confuse which theme the "Mummy" belongs to.

Such two examples show that in the entity linking task, the evidence may scatter in different paragraphs. Expanding the length of the mention context and entity description with more paragraphs may benefit the model's performance. However, since BERT is a deep cross-attention encoder, directly expanding the input length in the BERT-based baseline model is infeasible since both the time and space complexity grows with the square of tokens.

This problem is rarely discussed in the entity linking literature. In contrast, some existing work in the machine reading comprehension area has already studied on finding answers from multiple paragraphs of documents. For example, (Clark and Gardner 2018) proposes a typical model which feeds the query and each paragraph into a reading comprehension module and then select a final answer from the extracted candidate answers of each paragraph. This paradigm is sufficient in reading comprehension task since there is an explicit query. However, in the entity linking task, both the mention context and entity description are en-wrapped with several different paragraphs in corresponding documents respectively. No explicit queries and answers are available. Therefore, a new multi-paragraph reading model should be built to adapt for the zero-shot entity linking task.

In this paper, we propose a multi-paragraph reading model for zero-shot entity linking which can make use of more textual information. The key idea of our model is to take more paragraphs into consideration. In particular, during the encoding of each entity paragraph, we send both the mention context and one entity document paragraph into a BERT encoder to perform deep cross-attention between each other. Then, the obtained encodings are aggregated by an inter-paragraph attention mechanism. To deal with the problem that information about mention context is insufficient, we add an additional backward multi-paragraph reading step. Specifically, the paragraphs of the mention document are encoded separately. Then, the encoding obtained in the first step and the encoded mention paragraphs are matched by another attention module. Such a model can be summarized as a Bidirectional Multi-Paragraph Reading (Bi-MPR) model for entity linking which exploits more textual information in both mention and entity documents. The main contributions of our work are summarized as follows.

- We propose a two-step forward-backward matching process to deal with the zero-shot entity linking.
- We present an inter-paragraph attention mechanism to capture rich semantics in the forward matching step and an entity-mention attention mechanism to fully comprehend the mention context and entity description in the backward matching step.
- We evaluate the performance of Bi-MPR model on the zero-shot entity linking dataset. The experimental results show that our model achieves significant improvements over the BERT-based baseline. An extensive analysis of the length parameters shows our model achieves good accuracy using a relatively short inference time.

Models

In this section, we firstly introduce the BERT-based baseline proposed in (Logeswaran et al. 2019). Then, we describe a Uni-MPR model which leverages more textual information in the entity document and adopts Whole Entity Masking pre-training strategy. Finally, we extend the former model to a bidirectional model (Bi-MPR model) which incorporates more textual information in the mention documents. Figure 2 shows the architectures of the baseline, Uni-MPR model, and Bi-MPR model, respectively.

BERT-based Baseline

Due to the constraints of zero-shot entity linking, the only information we can obtain is the texts of the document containing the mention and the texts of the candidate entities documents. Therefore, the key of the zero-shot entity linking task is to select the target entity through a reading comprehension like process between the two kinds of texts. (Logeswaran et al. 2019) proposed a baseline model based on BERT which achieves the state-of-the-art performance on such tasks. Since the entities in the test domain cannot be seen during training, the model adopts domain-adaptive pre-training (DAP) before fine-tuning. Specifically, the model employs a BERT model which has been pre-trained on open corpora (i.e., Wikipedia and BookCorpus). Then, it keeps on pre-training using the texts in all domains that require linking, through the MLM (Masked Language Model) task. For

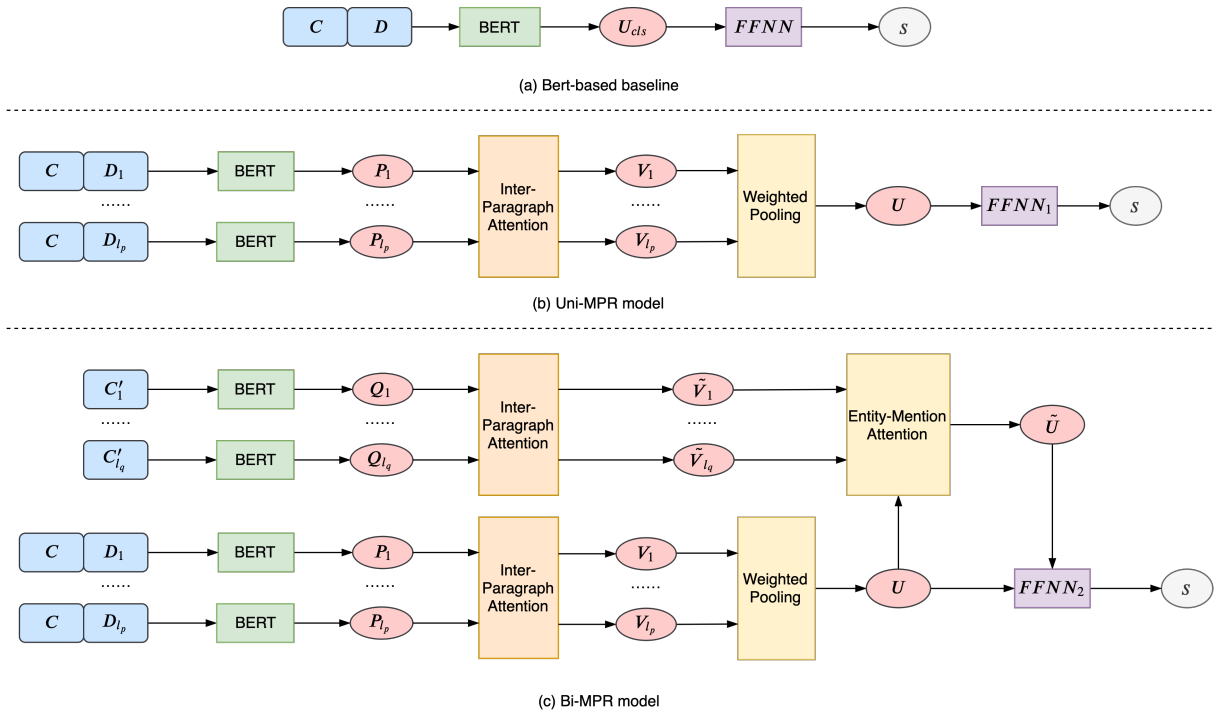

Figure 2: Architecture of the baseline model, the Uni-MPR model and the Bi-MPR model.

each target domain that the model is going to be applied on, an extra pre-training stage is added before fine-tuning only using the data in the target domain. After that, the model is fine-tuned by labeled data.

Taking the t-th candidate entity as an example, the baseline model concatenates n tokens $C = \{C_i\}_{i=1}^{n}$ surrounding the mention (i.e., mention context), first m tokens $D^t = \{D_i^t\}_{i=1}^{m}$ from the entity document (i.e., entity description) and some delimiters as input in the form of

$$\texttt{[CLS]}\, C\, \texttt{[SEP]}\, D^t\, \texttt{[SEP]} \tag{1}$$

where `[CLS]` and `[SEP]` are special placeholders from the vocabulary of BERT. At last, a coherence score s is calculated by sending the output of the last hidden layer corresponding to the position of the `[CLS]` token to a feed-forward neural network (FFNN). Figure 2(a) shows the overall structure of the baseline model.

Leveraging the attention mechanism proposed in the deep transformer (Vaswani et al. 2017), the baseline model performs a precise matching between the mention context and entity description. However, by referring to the examples shown in Figure 1, the limited input length of baseline definitely affects its performance directly. In order to read and detect the evidence in different paragraphs, we propose our models as follows.

Unidirectional Multi-Paragraph Reading

We note that in the baseline, a vanilla MLM task (i.e., replacing some tokens randomly with "`[MASK]`" and then predicting them by context tokens) is adopted in pre-training. However, masking all of the tokens randomly causes insufficient training. For example, as shown in Figure 3, the vanilla random masking strategy masks "heroes" in "super heroes minifigure". It is easy for the model to predict "heroes" in this case since "super heroes minifigure" is a named entity so that the three words always occur together. Therefore, inspired by (Zhang et al. 2019) and (Cui et al. 2019), we randomly select some entity names in the entity collection and mask all the words within them. In this way, the model is forced to predict the entities by understanding their contexts thus learn better entity-sensitive representations. We name this strategy "Whole Entity Masking"(WEM). In practice, we both randomly select some entity names and other ordinary words. Such method enhances the entity-sensitive representation ability in comparison to the vanilla MLM strategy. In the Experiments, we will show the proposed pre-training method is more effective.

As we have mentioned above, the problem of the baseline model partially comes from the limited length of entity description. To tackle this problem, we propose a Uni-MPR model which leverages more entity description by matching multiple paragraphs in the candidate entity document with the mention context. Specifically, same as the baseline model, the mention context consists of n tokens around the mention. For t-th entity, we extend the entity description by collecting l_p paragraphs in the entity document, each of which has m tokens, that is, $D^t = \{D_i^t\}_{i=1}^{l_p}$, $D_i^t = \{D_{ij}^t\}_{j=1}^{m}$. The input is in a form of

$$\texttt{[CLS]}\, C\, \texttt{[SEP]}\, D_i^t\, \texttt{[SEP]} \tag{2}$$

, where `[SEP]` is the separator and `[CLS]` indicates the po-

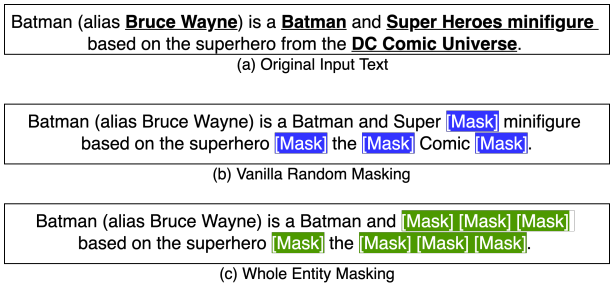

Figure 3: Vanilla Random Masking vs. Whole Entity Masking

sition of mention-entity representation. By BERT, we obtain a representation of each paragraph

$$P_i^t = \text{BERT}\left(C, D_i^t\right) \tag{3}$$

where $P_i^t \in \mathbb{R}^{d_{md}}$ is the mention-entity representation and d_{md} is the size of hidden states of BERT.

Next, we build an inter-paragraph attention module to gather the semantic dependence information among the paragraphs. We implement it by the multi-head attention proposed in the transformer(Vaswani et al. 2017) which enables the model to attend in different sub-spaces. Let $P^t = \{P_i^t\}_{i=1}^{l_p}$, we have the following steps to get the vector representation of the entity document V^t.

$$H_h^t = \text{Attention}(P^t W_{Qh}^1, P^t W_{Kh}^1, P^t W_{Vh}^1) \tag{4}$$

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \text{Softmax}\left(\frac{\hat{Q}\hat{K}^T}{\sqrt{d_k}}\right)\hat{V} \tag{5}$$

$$V^t = \text{Concat}(H_1^t, ..., H_{d_{head}}^t) W_O \tag{6}$$

In Equations 4-6, $W_{Qh}^1, W_{Kh}^1, W_{Vh}^1 \in \mathbb{R}^{d_{md} \times d_k}$, $W_O \in \mathbb{R}^{d_{md} \times d_{md}}$ are parameters of the model, $d_k = d_{md}/d_{head}$ where d_{head} denotes the number of heads in the attention. As a result, we get $V^t = \{V_i^t\}_{i=1}^{l_p}$ where V_i^t denotes the vector of i-th paragraph in the t-th entity document. After that, we obtain a fixed-length representation vector U^t by a weighted-pooling layer. Formally, the weighted-pooling layer is shown as follows.

$$e_i^t = \tanh(W_1 V_i^t + b_1) \tag{7}$$

$$a_i^t = \frac{\exp\left(e_i^t\right)}{\sum_j \exp e_j^t} \tag{8}$$

$$U^t = \sum_i a_i^t \cdot V_i^t \tag{9}$$

where W_1,b_1 are trainable parameters. Since this representation is generated by combining the information of the mention context and entity description, we call it the mention-aware entity representation. Meanwhile, the weights can indicate the importance of the entity paragraphs given the mention context. Finally, we put the encoding U^t to a feed-forward neural network FFNN_1 and get the score of t-th candidate entity using a softmax function.

$$\hat{s}^t = \text{FFNN}_1(U^t) \tag{10}$$

$$s^t = \frac{\exp\left(\hat{s}^t\right)}{\sum_j^T \exp\left(\hat{s}^j\right)} \tag{11}$$

We employ cross-entropy as our loss function. The loss is calculated as follows.

$$L_u = -\left(y^t \log s^t + (1 - y^t)\log(1 - s^t)\right) \tag{12}$$

In Equation 12, $y^t \in \{0, 1\}$ and y^t equaling to 1 means that t-th entity is the gold entity otherwise it equals to 0.

Bidirectional Multi-paragraph Reading

Although the Uni-MPR model can promote performance by reading more paragraphs in the entity documents, its performance is still constrained. For the case in Figure 1(b), Uni-MPR model still cannot get a correct result because the key information exists in other paragraphs in the mention document other than the mention context that the model can read. Therefore, a desirable model should read across multiple paragraphs in both mention and entity documents. We need to leverage more textual information in the mention documents besides the mention context defined previously. A naïve method is to take multiple paragraphs in the mention document as input directly and apply the Uni-MPR model once per paragraph. However, directly applying the Uni-MPR model on multiple paragraphs in the mention document needs to read each mention-entity paragraph pair which brings high time complexity. Besides, not all the paragraphs of the mention document are related to the given mention. Treating them equally may incur noises.

To solve the above issues, we introduce a bidirectional multi-paragraph reading model, i.e., Bi-MPR model. Specifically, we firstly obtain the mention-aware entity representation in the same way as the Uni-MPR model and then use it as a "query vector" to backward match multiple paragraphs in the mention document. In this way, our proposed model obviate the need of matching each mention-entity paragraph pair. In addition, it emphasizes the importance of the mention context which is encoded in the mention-aware entity representation. Thus, it only needs to perform the matching between the fixed entity representation U^t and multiple mention document paragraphs.

In detail, after removing the previously defined mention context from the mention document, we truncate l_q paragraphs in the rest of the mention document and collect them as input, each of which has n tokens, that is, $C' = \{C_i'\}_{i=1}^{l_q}, C_i' = \{C_{ij}'\}_{j=1}^{n}$, and encode them using BERT. We can obtain

$$Q_i = \text{BERT}(C_i') \tag{13}$$

where $Q_i \in \mathbb{R}^{l_q \times d_{md}}$ is the output at the position of [CLS]. Then, an inter-paragraph attention is performed to model paragraph-wise dependence and the updated representation of each paragraph $\tilde{V}_i^t$ is obtained. Formally,

$$\tilde{H}_i = \text{Attention}(Q_i W_{Qh}^2, Q_i W_{Kh}^2, Q_i W_{Vh}^2) \tag{14}$$

$$\tilde{V}_i = \text{Concat}\left(\tilde{H}_{i1}, ..., \tilde{H}_{ih}\right) W_O' \tag{15}$$

Then, U^t is regarded as the query and $\tilde{V}_i$ is regarded as the representation of each paragraph of the mention document, and the backward matching is implemented by an entity-mention attention. This attention is in essence a multi-head attention using U^t and $\tilde{V}_i$ as input. Formally,

$$G_i^t = \text{Attention}(U^t W_{Qh}^3, \tilde{V}_i W_{Kh}^3, \tilde{V}_i W_{Vh}^3) \quad (16)$$

$$\tilde{U}^t = \text{Concat}\left(G_1^t, ..., G_{d_{head}}^t\right) W_O' \quad (17)$$

The resulting $\tilde{U}_i^t$ contains the matching information incorporating multiple paragraphs in the mention document. In the end, we calculate the entity compatibility score by a feed-forward neural network using the concatenation of U^t and $\tilde{U}^t$ as input. Formally, the score is calculated as follows.

$$\hat{s}^t = \text{FFNN}_2\left(\text{Concat}(U^t, \tilde{U}^t)\right) \quad (18)$$

$$s^t = \frac{\exp\left(\hat{s}^t\right)}{\sum_j^T \exp\left(\hat{s}^j\right)} \quad (19)$$

Same as Uni-MPR, we adopt cross-entropy as the loss function.

Domain	Length	Ent. Spans	Ent. Tokens
Training			
American Football	665.06	47.83	85.86
Doctor Who	264.14	33.54	49.30
Fallout	229.68	20.28	35.03
Final Fantasy	497.35	35.77	59.70
Military	870.55	32.93	59.67
Pro Wrestling	639.53	50.11	88.71
Star Wars	379.76	53.60	79.43
World of Warcraft	242.16	28.23	40.69
Validation			
Coronation Street	264.55	6.15	14.51
Muppets	161.50	18.71	32.98
Ice Hockey	282.62	23.48	45.05
Elder Scrolls	269.03	18.40	30.18
Forgotten Realms	257.29	20.97	30.81
Lego	223.86	17.47	24.87
Star Trek	393.21	47.82	69.86
YuGiOh	643.68	62.15	96.42

Table 1: Average length, entity spans, entity tokens of the documents in each domain.

Experiments

Experimental Setup

Dataset We conduct our experiments on the dataset which is proposed in (Logeswaran et al. 2019) and built using documents on Wikia*. Wikia communities consist of online encyclopedias, each one specializing in a particular subject such as a fictional universe. In this dataset, 8 domains are used for training, 4 for validation and 4 for testing. The training set has 49,275 labeled mentions while the validation and test sets both have 10,000 mentions.

Model Settings For a fair comparison, we initialize our model using a publicly available uncased base version of BERT and readers can refer to (Devlin et al. 2019) for details.

In the pre-training stage, in addition to randomly masking all of the tokens, we adopt the WEM strategy which masks randomly selected entities occurring in the text. We firstly tokenize the documents using the WordPiece tokenizer and then split the documents to 256-token paragraphs.

For tokenized paragraphs, we randomly replace 15% of the tokens and 50% of the entities with `[MASK]` for prediction. Specifically, we extract the entity spans using the longest prefix matching with the given domain-specific entity dictionary. To deal with multi-word entities spanning the boundary of two adjacent paragraphs, we slightly modify the boundaries of such paragraphs to include the whole entity while keeping the length of each paragraph not exceeding 256. For some entities with disambiguation titles (e.g., "Breton (Online)" and "Breton (Skyrim)" in the "Elder Scrolls" domain), we only match the words outside the brackets. Table 1 shows the average length, numbers of entity spans and numbers of the entity tokens of the domain-specific documents. We use Adam optimizer(Kingma and Ba 2015) with a learning rate of 2e-5 and warmup over the first 10% of total 10000 steps. The batch size is 16.

During the fine-tuning stage, the length of paragraph m,n and the number of paragraphs l_p,l_q influence the performance significantly in terms of accuracy and inference time. We experiment several parameter settings of these parameters. Since the BERT-based baseline is under a setting of $m = n = 128$, for the sake of fairness, we set $m = n = 128$, $l_p = l_q = 2$ when comparing with the existing BERT-based baseline.

Performance Comparison

We choose six models, i.e., the baseline, baseline+WEM, Uni-MPR(w/o WEM), Uni-MPR, Bi-MPR(w/o WEM) and Bi-MPR, and conduct experiments to observe their accuracy in the zero-shot entity linking task. The accuracy of earlier existing work leveraging Bi-LSTMs or CNNs like deep-ed (Ganea and Hofmann 2017) and CDTE(Gupta, Singh, and Roth 2017), is excerpted from (Logeswaran et al. 2019). Readers can refer (Logeswaran et al. 2019) for more details. Here, baseline+WEM refers to the same model as the baseline but pre-trained by the WEM strategy. Default Uni-MPR model and Bi-MPR model are pre-trained by WEM. To show the gain of the mutli-paragraph reading mechanism individually, we also list the performance of the models without WEM pre-training (i.e., Uni-MPR(w/o WEM) and Bi-MPR(w/o WEM)).

Since our work focuses on ranking the candidate entities rather than generating the candidates, we conduct experiments on the normalized collections of the dataset where top-64 candidates of each mention are retrieved by BM25 algorithm. Table 2 shows the accuracy of different models on each domain of validation and test sets.

As shown in Table 2, our proposed models outperform the

*https://www.wikia.com

Model	Coronation Street	Muppets	Ice Hockey	Elder Scrolls	Macro Acc.	Micro Acc.
deep-ed	-	-	-	-	-	26.96
CDTE	-	-	-	-	-	27.03
Baseline	82.82	81.59	75.34	72.52	78.07	76.5
Baseline+WEM	84.21	81.55	74.29	72.66	78.18	76.52
Uni-MPR(w/o WEM)	85.31	80.78	78.07	73.74	79.48	77.83
Uni-MPR	88.27	**82.83**	78.12	74.00	80.81	78.8
Bi-MPR(w/o WEM)	86.47	82.47	78.07	74.87	80.41	78.78
Bi-MPR	**90.12**	82.50	**79.37**	**75.92**	**81.98**	**80.1**
Model	Forgotten Realms	Lego	Star Trek	YuGiOh	Macro Acc.	Micro Acc.
Baseline	85.60	76.90	75.80	67.22	76.38	74.21
Baseline(WEM)	86.21	78.23	77.66	66.47	77.14	74.98
Uni-MPR(w/o WEM)	85.55	77.42	78.23	68.29	77.62	75.78
Uni-MPR	87.25	78.57	80.56	67.31	78.42	76.65
Bi-MPR(w/o WEM)	89.09	77.18	79.20	**69.98**	78.61	76.70
Bi-MPR	**89.60**	**80.50**	**81.04**	68.74	**79.97**	**77.85**

Table 2: Accuracy on the validation set and test set

Model	HO	MC	AS	LO
Baseline	87.64	77.27	75.89	71.46
Baseline+WEM	89.90	78.82	76.02	71.03
Uni-MPR	91.43	79.07	75.60	73.53
Bi-MPR	92.84	81.93	77.37	73.88

Table 3: Accuracy on the category-specific test subsets including High Overlap(HO), Multiple Categories(MC), Ambiguous Substring(AS), Low Overlap(LO).

baseline model on average. In particular, adding WEM while remaining the model structure same as the baseline model can improve the performance, which shows that the WEM strategy can learn better representations than the vanilla random masking. Furthermore, we note that the Uni-MPR and Bi-MPR models make substantial improvements on the domains whose documents are relatively long (e.g., Elder Scrolls) and behave not so salient in domains with relatively short documents (e.g., Muppets). This phenomenon occurs because the entity description length of Baseline(WEM) is 128 which is nearly equal to the average length of short documents (e.g., Muppets documents have an average length of 161.5).

Further, we analyze the linking accuracy on different categories. There are four categories including "High Overlap" in which the mention text is identical to the entity name, "Multiple Categories" in which the entity name contains a disambiguation phrase (e.g., "Breton (Online)"), "Ambiguous substring" in which the mention is a substring of the entity name, and "Low Overlap" that all other mentions are subordinate to. Correspondingly, the test set can be divided into four subsets. Table 3 shows the accuracy of the mentions in test subsets. We find that, our models improve more in "High Overlap" and "Multiple Categories" than other two categories. We conjecture that the latter two categories require more complex reasoning which needs to be addressed in the future. It is somewhat surprising that in "Low Overlap", the performance of baseline+WEM is a bit worse than the baseline. We suppose that the mentions in "Low Overlap" are not obviously related to the entity names. As a result, the model cannot benefit from the WEM pre-training strategy.

Impact of Input Length

In our proposed models, there are two kinds of parameters controlling the input length. One is entity related parameters including the paragraph number l_p and the length of paragraph m. The other is mention related parameters including the paragraph number l_q and the length of paragraph n. In Bi-MPR model, we treat the mention document and the entity document with the same importance, that is, $m = n, l_p = l_q$. Therefore, the total input length of tokens is $(m + n) \times l_p$. In Uni-MPR model, $l_q = 1$ since we do not add any extra paragraphs from mention document.

In the view of total input text length, same input text length can be composed by different settings. For example, a Bi-MPR model whose $m = n = 64, l_p = l_q = 4$ takes 256 tokens from the mention document and the entity document as input. While another Bi-MPR model whose $m = n = 128, l_p = l_q = 2$ accepts 256 tokens, too. From this perspective of view, the BERT-based baseline model is a special case of our proposed Bi-MPR model which has $m = n = 128, l_p = l_q = 1$.

However, the performance of above models differs largely in terms of accuracy and inference time. To investigate the influence of the parameters, we conduct experiments using models under different parameter settings which are shown in Table 4. The accuracy and inference time of the models under different settings are shown in Figure 4 and Figure 5. We list the inference time of each mention-entity pair running on CPU and GPU, respectively. The CPU computations were run on a Intel Xeon Processor 5118 CPU. The GPU computations were run on a single Nvidia Tesla V100 GPU. Notice that all the above models are pre-trained by WEM. In this section, we experiment the models with input length up to 256.

No.	Model	Length Setting
1	Uni-MPR	$m = n = 64, l_p = 2, l_q = 1$
2	Uni-MPR	$m = n = 64, l_p = 4, l_q = 1$
3	Uni-MPR	$m = n = 128, l_p = 2, l_q = 1$
4	Bi-MPR	$m = n = 64, l_p = l_q = 2$
5	Bi-MPR	$m = n = 64, l_p = l_q = 4$
6	Bi-MPR	$m = n = 128, l_p = l_q = 2$

Table 4: Different length settings.

From Figure 4, we can find that, under the fixed setting of paragraph lengths(m and n), models leveraging more paragraphs (more l_p or l_q) achieve higher accuracy, which indicates that extending the input length and adopting the multi-paragraph reading mechanism are feasible and effective.

However, the accuracy of model No. 5(Bi-MPR, $m = n = 64, l_p = l_q = 4$) cannot reach the one of model No. 6(Bi-MPR, $m = n = 128, l_p = l_q = 2$) and model No. 3(Uni-MPR, $m = n = 128, l_p = 2, l_q = 1$), although they have same input length of the entity document. The major difference between them is that the later models send more tokens to the initial matching step in BERT (i.e., $m = n = 128$ vs $m = n = 64$). As a result, the cross-attention between tokens in mention and entity documents is performed more effectively than the former models.

Although the models sending more text to the cross-attention module can achieve higher accuracy, their inference time increases far more than the models with more paragraphs. In other words, here is a trade-off between the accuracy and the efficiency. Notice that the increased accuracy is not salient in comparison to the increased time consumption, the models with larger l_p, l_q and smaller m, n could be more preferable in practice.

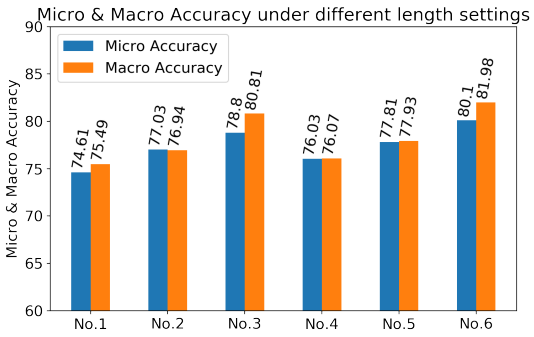

Figure 4: Accuracy on the validation set under different length parameters settings.

Related Work

Given a mention, the goal of an entity linking model is to find the corresponding entity from a collection of entities. In a classical entity linking task, the mentions are required to be linked to entities in a general knowledge base which provides various kind of information (Cucerzan 2007; Hoffart et al. 2011; Ratinov et al. 2011; Guo and Barbosa 2014).

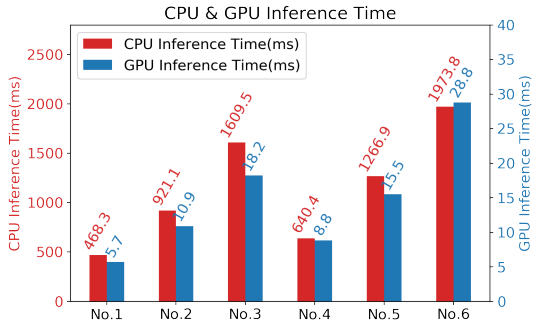

Figure 5: Inference time on the validation set under different length parameters settings.

With the help of the comprehensive knowledge base, most of the state-of-the-art models targeting on such task make use of information other than texts (Kundu et al. 2018; Raiman and Raiman 2018; Ganea and Hofmann 2017). Although these models have promising performance on this task, it is hard to adapt these models to specialized domains where no comprehensive knowledge base is provided.

Early studies employ handcrafted features to model the textual coherence between mention and entity (Ji and Grishman 2011; Shen, Wang, and Han 2015; Milne and Witten 2008; Chen and Ji 2011; Dredze et al. 2010). Recently, many studies resort to the methods based on neural networks which need no manual efforts. Usually, entity linking models based on deep learning usually encode two parts of text individually. For example, (Ganea and Hofmann 2017; Kolitsas, Ganea, and Hofmann 2018) use pre-trained entity embeddings by modeling the co-occurrences of the words and the entities. (Yamada et al. 2016) learns the embedding of words and named entities together. Moreover, neural networks (Fang et al. 2019; Gupta, Singh, and Roth 2017; Xue et al. 2019) are adopted to encode mention contexts and entity descriptions. However, the semantic relationships are not fully exploited due to the absence of cross-attention mechanism and the limited representation ability of the encoders. Even if models like (Logeswaran et al. 2019; Fang et al. 2020) employ pre-trained BERT model which performs cross-attention, it does not fully capture the semantics which spread around multiple paragraphs.

Conclusion

Zero-shot entity linking task forces the models to link mentions to unseen entities and leverage only textual information which challenges the generalization and text comprehension ability of entity linking models. Usually, the evidence for linking the golden entity could scatter in different paragraphs of a document which is hard to collect and comprehend. Focusing on this phenomena, we present a new bi-directional multi-paragraph reading model which can capture long-range text dependence between mention and entity documents but restrict the increasing amount of inference time in an acceptable range. The experimental results on the challenging zero-shot entity linking dataset show our model achieves state-of-the-art performance in different domains.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2017YFC0803300)

References

Amplayo, R. K.; Lim, S.; and Hwang, S. 2018. Entity Commonsense Representation for Neural Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 697–707. Association for Computational Linguistics.

Chang, M. 2016. From Entity Linking to Question Answering - Recent Progress on Semantic Grounding Tasks. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, 2. The COLING 2016 Organizing Committee.

Chen, Z.; and Ji, H. 2011. Collaborative Ranking: A Case Study on Entity Linking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, 771–781. ACL.

Clark, C.; and Gardner, M. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 845–855. Melbourne, Australia: Association for Computational Linguistics.

Cucerzan, S. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 708–716. Prague, Czech Republic: Association for Computational Linguistics.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; and Hu, G. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *CoRR* abs/1906.08101.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dredze, M.; McNamee, P.; Rao, D.; Gerber, A.; and Finin, T. 2010. Entity Disambiguation for Knowledge Base Population. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, 277–285. Tsinghua University Press.

Fang, Z.; Cao, Y.; Li, Q.; Zhang, D.; Zhang, Z.; and Liu, Y. 2019. Joint Entity Linking with Deep Reinforcement Learning. In *The World Wide Web Conference*, WWW '19, 438–447. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Fang, Z.; Cao, Y.; Li, R.; Zhang, Z.; Liu, Y.; and Wang, S. 2020. High Quality Candidate Generation and Sequential Graph Attention Network for Entity Linking. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 640–650. ACM / IW3C2. doi:10.1145/3366423.3380146. URL https://doi.org/10.1145/3366423.3380146.

Ganea, O.-E.; and Hofmann, T. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2619–2629. Copenhagen, Denmark: Association for Computational Linguistics.

Guo, Z.; and Barbosa, D. 2014. Robust Entity Linking via Random Walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, 499–508. New York, NY, USA: Association for Computing Machinery.

Gupta, N.; Singh, S.; and Roth, D. 2017. Entity Linking via Joint Encoding of Types, Descriptions, and Context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2681–2690. Copenhagen, Denmark: Association for Computational Linguistics.

Hoffart, J.; Yosef, M. A.; Bordino, I.; Fürstenau, H.; Pinkal, M.; Spaniol, M.; Taneva, B.; Thater, S.; and Weikum, G. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 782–792. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Ji, H.; and Grishman, R. 2011. Knowledge Base Population: Successful Approaches and Challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1148–1158. Portland, Oregon, USA: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kolitsas, N.; Ganea, O.-E.; and Hofmann, T. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 519–529. Brussels, Belgium: Association for Computational Linguistics.

Kundu, G.; Sil, A.; Florian, R.; and Hamza, W. 2018. Neural Cross-Lingual Coreference Resolution And Its Application To Entity Linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 395–400. Melbourne, Australia: Association for Computational Linguistics.

Logeswaran, L.; Chang, M.-W.; Lee, K.; Toutanova, K.; Devlin, J.; and Lee, H. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguis-*

tics, 3449–3460. Florence, Italy: Association for Computational Linguistics.

Milne, D. N.; and Witten, I. H. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, 509–518. ACM.

Raiman, J.; and Raiman, O. 2018. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5406–5413. AAAI Press.

Ratinov, L.; Roth, D.; Downey, D.; and Anderson, M. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1375–1384. Portland, Oregon, USA: Association for Computational Linguistics.

Roth, D.; Ji, H.; Chang, M.; and Cassidy, T. 2014. Wikification and Beyond: The Challenges of Entity and Concept Grounding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Tutorial Abstracts*, 7. The Association for Computer Linguistics.

Shen, W.; Wang, J.; and Han, J. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 5998–6008.

Xue, M.; Cai, W.; Su, J.; Song, L.; Ge, Y.; Liu, Y.; and Wang, B. 2019. Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 5327–5333. ijcai.org.

Yamada, I.; Shindo, H.; Takeda, H.; and Takefuji, Y. 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*, 250–259. Berlin, Germany: Association for Computational Linguistics.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451. Florence, Italy: Association for Computational Linguistics.