

Received November 29, 2020, accepted December 16, 2020, date of publication December 23, 2020, date of current version January 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046787

# Diversified Semantic Attention Model for Fine-Grained Entity Typing

YANFENG HU<sup>1,2</sup>, XUE QIAO<sup>1,2</sup>, LUO XING<sup>3</sup>, AND CHEN PENG<sup>1,2</sup>

<sup>1</sup>Institute of Electronics, Chinese Academy of Sciences, Suzhou 215123, China

<sup>2</sup>Key Laboratory of Intelligent Aerospace Big Data Application Technology, Suzhou 215123, China

<sup>3</sup>Software Engineering Institute, East China Normal University, Shanghai 200062, China

Corresponding author: Xue Qiao (xqiao@mail.ie.ac.cn)

**ABSTRACT** Fine-grained entity typing, which aims to assign specific types to entity mentions in text, is attracting increasing attention in the field of natural language processing (NLP). However, it is quite a challenging problem due to the highly ambiguous nature of many entity mentions. Most existing entity typing methods based on attention mechanism generally extract the salient features separately from the entity mention and the contextual words. However, these approaches suffer from *two main limitations*: (1) They ignore the rich information contained by entity mentions when applying the attention mechanisms. (2) They do not consider the diversity of attention processes which can be beneficial in finding the discriminative features. To address these issues, we propose the *diversified semantic attention model (DSAM)* for fine-grained entity typing, and the main novelties are: (1) It explicitly pursues the diversity of attention and is able to maximally gather discriminative information. (2) It integrates two level attentions—the *mention-level attention* and the *context-level attention*—to jointly capture the rich information from mentions and contexts to enhance their mutual promotions. (3) It combines the *attention maps constraint* and the *attention segments constrain* to exploit the subtle semantic differences for distinguishing the subtypes. Importantly, the proposed DSAM approach can be trained end-to-end without employing ad-hoc features or post-processing. Extensive experiments using three benchmark datasets demonstrated that our DSAM approach achieves competitive performance compared to the current state-of-the-art methods used for fine-grained entity typing.

**INDEX TERMS** Fine-grained entity typing, diversified semantic attention model (DSAM), long shot-term memory (LSTM), mention-aware attention mechanism.

## I. INTRODUCTION

### A. RESEARCH BACKGROUND

Named entity recognition (NER) [1]–[4] is a basic task in natural language processing (NLP), aiming to jointly resolve the boundaries and types of named entities in a document. In this paper, we focus on the task of named entity typing, which is to assign types or labels to the detected entity mentions. Existing studies related to entity typing can be categorized into two groups: the *coarse-grained entity typing* and the *fine-grained entity typing*. Fine-grained entity typing is a challenging task which aims to classify entity mentions into a large set of fine-grained subtypes. However, the coarse-grained entity typing only needs to label the entity with a more general type, such

as *Person*, *Location*, *Organization* and *Time*. In most cases, coarse-grained typing is too general and not exact enough for many tasks. An example of a question answering task is shown in Fig. 1. The candidate answers to the question “Who is the author of the bestselling Harry Potter series?” are “J.K. Rowling”, “James Cameron” and “Albert Einstein”, which is difficult to make a choice if they are all classified as the coarse-grained type *Person*; it would be much easier to draw the answer as “J.K. Rowling”, if the candidate answers are fine-grained classified into *Author*, *Artist*, and *Scientist* respectively. Although the coarse-grained types have been extended to around one hundred in recent works, they are still in different levels with extremely uneven granularity.

Typically, fine-grained entity typing contains several hundreds of types that are arranged into a hierarchical structure. In addition, entity mentions can be set into different

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu<sup>1</sup>.




Question	Who is the author of the bestselling Harry Potter series?		
Answer candidates	 J.K. Rowling	 James Cameron	 Albert Einstein
Coarse-grained entity typing	Person	Person	Person
Fine-grained entity typing	Person / Author	Person / Artist	Person / Scientist

FIGURE 1. Coarse-grained entity typing vs. fine-grained entity typing.

types based on their local contexts. As shown in Fig. 2, three sentences mention the same entity “Stephen Hawking” which actually belongs to three different fine-grained types. Our goal is to classify the first entity mention “Stephen Hawking” to *Person/Author*, the second one to *Person*, and *Person/Scientist* for the third case. The fine-grained types of the entity mentions can be a great asset for many high-level NLP applications, such as question answering [5], entity linking [6], relation extraction [7], reading comprehension [8], and knowledge completion [9].

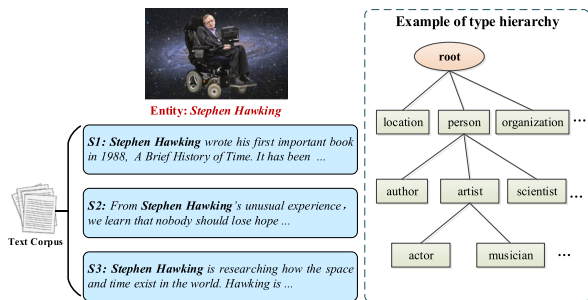


FIGURE 2. An illustration of fine-grained entity typing.

The main difficulty for fine-grained entity typing comes from the fact that entity mentions with different local contexts may have subtle semantic differences which are difficult to identify. Traditional methods solve this problem by carrying out extensive feature engineering, such as hand-crafted features and external resources [10]. However, these methods rely heavily on hand-crafted features, which limits their capabilities of generalizing uncommon or unseen entity mentions. To eliminate the use of hand-crafted features, researchers begin focusing on learning representations of entity mention and its context. Previous methods [11], [12] generally learn feature representations with pre-trained fixed word embeddings, such as Word2Vec and GloVe. Despite of achieving compelling results by using fixed word embeddings, these methods still suffer from limitations for their incapability in capturing word semantic meanings in different contexts. **This is the first limitation.**

To address the limitation of fixed word embeddings, the contextualized word representations (e.g., ELMo) were proposed by Peters *et al.* [13] and have been extensively used in recent works [14]–[16]. Following this elegant recipe, we employ contextualized word representations to better represent out-of-vocabulary words as well as capture context-aware word semantics. Besides the contextualized

word representations, local semantics of sentences often play a significant role in distinguishing subtypes of entity mention. Attention mechanism has been widely used in existing neural-based methods [17]–[19] to learn the most relevant semantic information in text. However, these methods only apply attention mechanism to the context but ignore the rich information in the entity mention. **This is the second limitation.**

Meanwhile, in the annotating process, assigning labels to entity mentions can be regarded as a two-step attention process: human annotators generally first focus their attention on decisive words in text, instead of processing the entire text at once, and then different decisive words over time are combined to build up the global semantics of the text. An intuitive idea is to convert the problem of finding different attentive words simultaneously to finding them multiple times. Recently, many attentive methods [11], [12], [18], [19] have achieved promising performance. Shimaoka *et al.* [18] introduced an attention-based Long Short-Term Memory (LSTM) network to allow the model focusing on relevant expressions. Moreover, they [19] incorporated hand-crafted features into the attention-based neural models, and the performance on the FIGER dataset and OntoNotes dataset was improved. However, it is still difficult for existing attention-based methods to find multiple discriminative words simultaneously. **This is the third limitation.**

To address the above three limitations, we propose a diversified semantic attention model (DSAM) to classify entity mentions into fine-grained subtypes. Particularly, the proposed DSAM initially generates multiple attention segments to extract the diversified features for attention. Furthermore, the long short-term memory (LSTM) network integrated with a mention-aware attention mechanism is imposed to the sequential attention segments from coarse to fine resolution. Then, a dynamic feature representation is built up by incrementally combining the information from different contexts and lengths of the sentences. Finally, the general and the local semantic features of the sentence are captured with this representation to facilitate the fine-grained entity classification.

Fig. 3 presents the research motivation of this work. The first example shows the impact of mention attention. Existing method misclassifies the entity mention “British Defense Ministry” as *Location* probably because “British” usually appears in location mentions, while it can be correctly classified as *Organization* when higher weights are assigned to “Defense Ministry”. In Example #2 and #3, although “Tim Breene” and “WCRS” occur in the same sentence, the model should use “chief executive officer” to help classify “Tim Breene” as *Person*, but focus on “executive” and “new agency” to determine that “WCRS” should be an *Organization*.

**B. CONTRIBUTIONS**

The main novelties and contributions of our DSAM approach can be summarized as follows:

#1 ... its purchase by GEC may heighten [ORGANIZATION British Defense Ministry] worries about concentration in the country's defense industry...

#2 ... Both will report to [PERSON Tim Breene], a former WCRS executive who will be chief executive officer at the new agency ...

#3 ... Both will report to Tim Breene, a former [ORGANIZATION WCRS] executive who will be chief executive officer at the new agency ...

**FIGURE 3.** Mention and context attention visualization. Darker background color indicates higher attention score.

- **Diversified Semantic Attention Model.** Most existing studies ignore the diversity of attentive features for fine-grained entity typing, while the diversified attentive features are highly beneficial to capturing the discriminative semantic information. For addressing this problem, we propose the DSAM approach to discover the global semantics as well as the discriminative nuances from a sentence to enhance entity type identification. It integrates two types of features: the *coarse-grained global feature* focuses on the whole representation of a sentence and the *fine-grained diversified attention feature* focuses on the distinguishing semantic differences. By combining the two types of features, an incremental sentence representation with various attention features is dynamically built up, from which subtle semantic differences can be captured accurately. As far as we know, the proposed DSAM is the first method to exploit the diversity of the semantic attention for fine-grained entity typing.
- **Mention-aware Attention Mechanism.** Most attentive methods fail to consider the potential information of entity mentions when performing an attention mechanism, while some words in the entity mention may provide more useful information for entity typing. For addressing this problem, a two-step mention-aware attention mechanism is proposed to capture the rich information from mentions and contexts as much as possible. Particularly, it calculates the attention scores for contexts in a mention-aware manner, allowing the model to grasp different information for different entity mentions. Two levels of attentions are jointly considered: the *mention-level attention* and the *context-level attention*, both of which are highly useful in capturing important information and are complementary in improving the performance of fine-grained entity typing.
- **Diversity Constraint Model.** To ensure the diversity in the attention process, we propose an attentive feature generation approach driven by a diversity

constraint model. It highlights the saliency of attentive features and enhances their discrimination to ensure that the generated attention maps are highly representative. It combines two types of constraints: the *attention maps constraint* enforces that the generated attention maps are highly representative, and the *attention segments constraint* reduces the overlapping proportions among the segments as well as highlights the saliency of the segments, which eliminates the redundancy and enhances the discrimination of the generated segments. Combination of these two constraints not only significantly promotes the selection of discriminative feature, but also achieves a notable improvement on fine-grained entity typing.

- **End-to-end Trainable Model.** Another superiority of our model lies at its end-to-end framework. It is jointly trained, from scratch, by optimizing the probability of the output using a variant of loss function. In summary, we propose an elegant neural-based model that attempts fine-grained entity typing end-to-end without providing ad-hoc features or employing post-processing.

### C. PAPER ORGANIZATION

The rest of the paper is organized as follows: Section II discusses the studies related to fine-grained entity typing; Section III elaborates the details of the proposed DSAM approach; Section IV presents our experimental settings and results; Section V provides a conclusion of our paper.

## II. RELATED WORK

Fine-grained entity typing has been widely studied in recent years. Existing methods can be categorized into three groups: distant supervision-based methods, neural-based methods, and attention-based methods.

### A. DISTANT SUPERVISION-BASED METHODS

For most existing fine-grained entity typing methods, they often use distant supervision to generate training examples and assume that all candidate types generated in this manner are correct. Ling *et al.* [7] were the first to adopt distant supervision method [20] for fine-grained entity typing. They derived 112 types from Freebase and automatically created the training data from Wikipedia. From then on, many works begin to focus on reducing label noise induced by distant supervision. Gillick *et al.* [21] proposed three types of pruning heuristics to constrain label noise. Ren *et al.* [10] designed a novel partial-label loss to further reduce the label noise. Moreover, Xu *et al.* [22] introduced a method of normalization of hierarchical loss to reduce specific types of noise. A recent study [23] introduced a penalty term in the optimization process to effectively diminish the side effect of the label noise and confirmation bias. However, these distant supervision-based methods aggressively filter training examples and may cause performance degradation. To address this problem, recent studies converted fine-grained typing into the task of a graph-based semi-supervised classification.

For example, Jin *et al.* [24] used links between entities to construct an entity graph, and jointly utilized entity features and a graph structure to make an entity type inference. Later, they further proposed a novel architecture [25] consisting of three graph convolutional networks to capture different kinds of semantic correlations between entities to refine entity types. Although distant supervision-based methods provide an efficient way to annotate training data, they ignore the local contextual information associated with entities and limit its usage in context-aware applications.

### B. NEURAL-BASED METHODS

In recent years, thanks to the rapid development of deep learning methods and artificial intelligence, many neural-based methods have been proposed and achieved pleasant results. These methods used multiple neural networks to learn semantic information of entities from local context for a better entity classification. Dong *et al.* [26] first proposed a hybrid neural network architecture comprised of a mention model and a context model. The mention model obtains the vector representations of entity mentions by using recursive neural network (RNN), while the context model derives the hidden representations of context words by employing multi-layer perceptron. After that, Karn *et al.* [27] introduced an encoder-decoder neural model to infer entity types, which can be trained end-to-end. Different from previous methods that obtained the entity context information through a fixed window, Liu *et al.* [28] introduced a novel entity typing method with sliding window context and dynamic global information. Despite of achieving promising results, these neural-based methods are still limited by ignoring type hierarchy in inferring process. Therefore, Ren *et al.* [23] employed a neural network, called hierarchical inference model, to infer entity types layer by layer, in order to maximally capture entity information, the mention as well as its context aspects.

### C. ATTENTION-BASED METHODS

Inspired by recent advances in neural machine translation [29], attention mechanism has been widely used in neural fine-grained entity typing methods [11], [18], [19], [30] to weight context words. Choi *et al.* [11] presented an neural architecture that resembles the Attentive NER model [19] to improve the sentence and mention representations; they also introduced a new multitask objective to handle multiple sources of supervision. Xin *et al.* [12] introduced a knowledge attention mechanism to capture important context words and improve the quality of context representation; they used the entity representation obtained from the external knowledge base as the attention query. While these attention-based methods outperform previous methods which only use sparse binary features [7], [21] or distributed representations [31], they suffer from that the attention on context is computed solely upon the context, considering no alignment to the entity. To overcome this drawback, Zhang *et al.* [32] proposed a neural architecture which learns more context-aware representations by using a better attention mechanism and

taking advantage of semantic discourse information available in the document as well as sentence-level contexts. Recently, the graph-based algorithm combining with attention mechanism is developed to integrate entity feature and structural information. Xiong *et al.* [33] encoded both global label co-occurrence statistics and word-level similarities by using a graph-enhanced model equipped with an attention-based matching module. Unlike Xiong *et al.* [33] operating under Euclidean assumption, López *et al.* [34] imposed a hyperbolic geometry to enrich the hierarchical information, and applied a self-attentive encoder to get the context representation. Furthermore, Lin *et al.* [19] developed a hybrid model that incorporates latent type representation in addition to binary relevance, which is able to capture inter-dependencies between entity types. Unlike above mentioned attention-based methods, we propose a diversified semantic attention model to pursue the diversity of attention and collect distinguishing contextual information to improve the performance of fine-grained typing.

## III. THE PROPOSED APPROACH

The overview of our DSAM approach is shown in Fig. 4. It includes three components: attention segments generation, diversified semantic attention, and classification. First, our DSAM approach divides the input sentences into several segments of different lengths through the attention segments generation component. Then, these segments are fed into the following diversified semantic attention component to predicted the attention maps. This highlights import words or phrases within each segment and maximizes the information gained across multiple attention segments. Furthermore, a two-step mention-aware attention mechanism is proposed to focus on important information in mentions and contexts. Different from previous attentive methods that focus on a single distinguishing feature, our DSAM approach jointly extracts diverse features by a novel loss function. Meanwhile, the attentive features are dynamically pooled from the generated attention maps and accumulated into the diversified attention model. Finally, the type of entity mention will be predicted at each time step, and all the prediction results will be merged in an average manner to get the final prediction results.

### A. DIVERSIFIED ATTENTION SEGMENTS GENERATION

To make the attentive features diversified, an attention segments generation method is proposed for dividing the input sentences into multiple attention segments with each segment containing different words and being of different lengths. Some of the attention segments contain entity mentions, while others include only some of the context words. These attention segments provide abundant candidates of the original sentence, which is beneficial for capturing multiple discriminative semantic features to achieve better fine-grained accuracy of entity typing.

The generation of attention segments with different lengths and strides is shown in Fig. 5, where the “stride” refers to



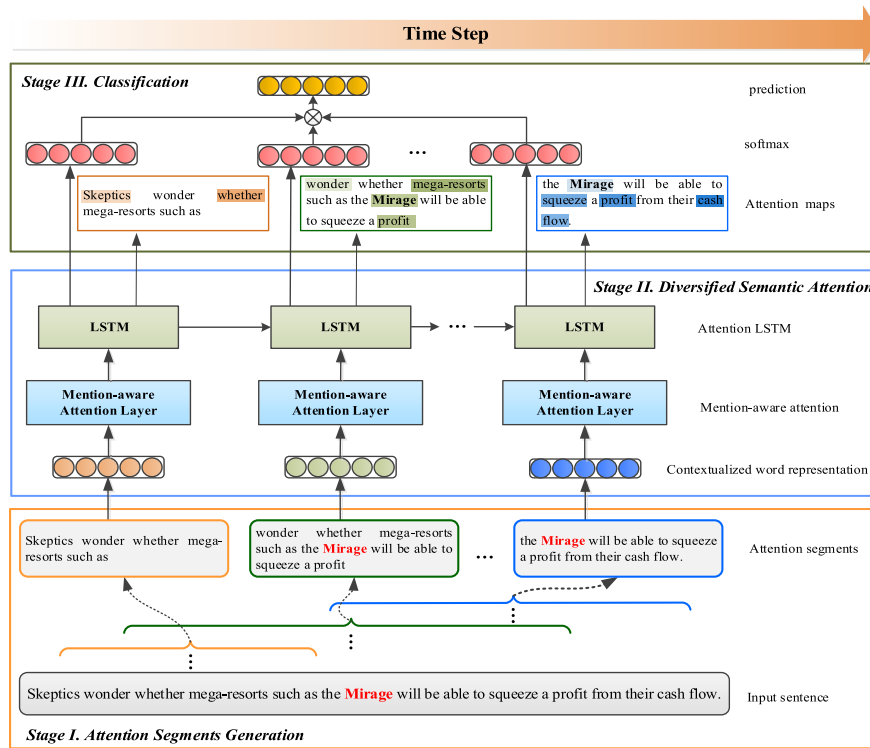


FIGURE 4. An overview of the proposed diversified semantic attention model (DSAM) approach.

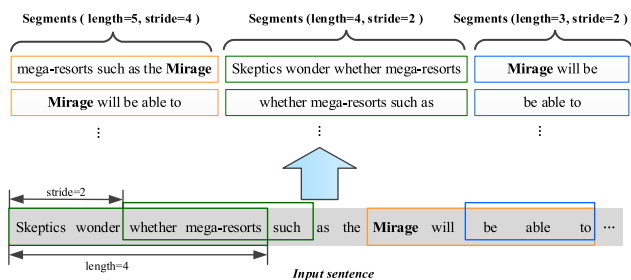


FIGURE 5. An illustration of our method for generating the attention segments.

the number of words that moves in each step on the original sentence, and “length” refers to the number of words that contained in each step. The length and stride jointly determine the number of attention segments to be generated. The attention segments will be cropped according to the defined length and stride along the direction of the input sequence. Clearly, for a long length, a long stride will produce a small number of segments. Inversely, given a shorter length, a short stride will generate more local segments.

Following the strategy above, the multiple attention segments generated will cover most of the information in the input sentence. Each attention segment will contain different words and be of different lengths. All of the generated attention segments will be jointly organized into a sequence, with the long segments placed ahead of the short segments. In this way, the global semantic information of sentences will

first be focused on, and then the local semantic differences of sentences will be subsequently captured.

### B. DIVERSIFIED SEMANTIC ATTENTION MODEL

Most existing attention-based methods [18], [19], [22] devote much to the learning of the discriminative features, but ignore the diversity in the attention process. The diverse attentions are important in simultaneously finding multiple semantically discriminative words to learn meaningful and representative features of the entity mentioned. To solve this important problem, a novel diversified semantic attention model is proposed in this paper. It converts the problem of simultaneously extracting different attentive features into extracting them at multiple times throughout the process. As shown in Fig. 6, the proposed diversified semantic attention model consists of two components: the attention map prediction and the attentive feature integration.

#### 1) ATTENTION MAP PREDICTION

In this stage (see Fig. 7), the pre-trained contextualized word representations [13] are first adopted to represent an input sentence. Then, a two-step mention-aware attention mechanism is employed to extract the most relevant features from the sentence to form the attentive features.

**Attention Segment Encoder:** Since we often need to determine the types, and especially the subtypes, according to the context, it is indispensable to collect contextual information for making the correct decision on a classification. Therefore, instead of using the fixed word embeddings, we adopt

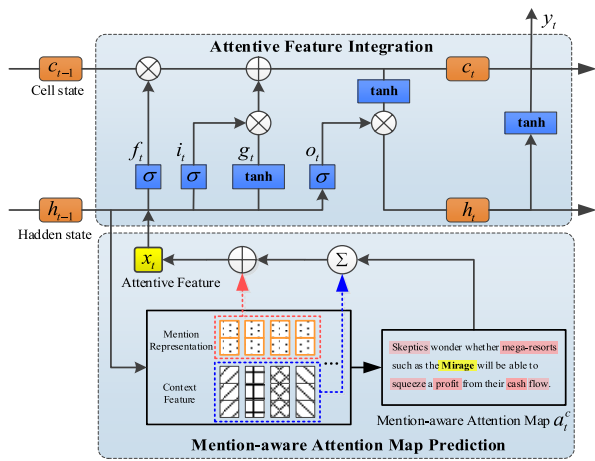


FIGURE 6. An illustration of the diversified semantic attention model.

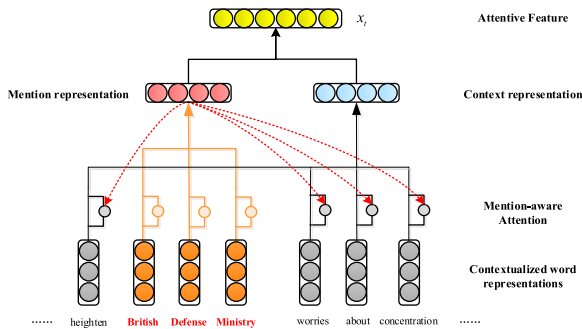


FIGURE 7. An illustration of the prediction of the mention-aware attention map.

contextualized word representations which can capture word semantics in different contexts. Let  $R_t = \{r_{t,1}, r_{t,2}, \dots, r_{t,L}\} \in \mathbb{R}^{L \times d_r}$  denotes the contextualized word representations of an attention segment  $S_t = \{w_{t,1}, w_{t,2}, \dots, w_{t,L}\}$  at time step  $t$ , where  $r_{t,i} \in \mathbb{R}^{d_r}$  is the  $d_r$ -dimensional representation corresponding to the  $i$ -th word  $w_{t,i}$  and  $L$  is the length of the attention segment  $S_t$ .

**Mention Representation:** Previous attentive models [18], [19], [22] only apply attention mechanism to the context. However, some words in an entity mention, such as “Defense Ministry” in Figure 7, may provide more useful information for determining the type. In order to focus on more informative words, we represent an entity mention  $m$  consisting of  $M$  words with an attention mechanism as follows:

$$m = \sum_i^M a_i^m r_i \quad (1)$$

where  $r_i$  is the contextualized representation of the  $i$ -th word in the entity mention  $m$ , and the attention score  $a_i^m$  of the entity mention  $m$  is computed as follows:

$$a_i^m = \text{soft max}(e_i^m) = \frac{\exp(e_i^m)}{\sum_{j=1}^N \exp(e_j^m)} \quad (2)$$

$$e_i^m = v \tanh(W^m r_i) \quad (3)$$

where  $W^m \in \mathbb{R}^{d_a \times d_r}$  and  $v \in \mathbb{R}^{d_a}$  are the trained parameters, and  $d_a$  is the dimension of the hidden attention layer.

**Mention-aware Context Representation:** Given the context  $c_t$  of an entity mention  $m$  in an attention segment  $S_t$ , we form its representation with a mention-aware attention mechanism as follows:

$$c_t = \sum_{i=1}^C a_{t,i}^c r_{t,i} \quad (4)$$

where  $C$  is the number of contextual words, and the attention map  $a_{t,i}^c$  is defined as follows:

$$a_{t,i}^c = \frac{\exp(W_i^h h_{t-1} + W_i^C (r_{t,i} \oplus m))}{\sum_{j=1}^C \exp(W_j^h h_{t-1} + W_j^C (r_{t,j} \oplus m))} \quad (5)$$

where  $W_i^h$  denotes the connection weights from the previous LSTM hidden state  $h_{t-1}$  to the  $i$ -th score of the attention map, and  $\oplus$  represents concatenation. Similarly,  $W_i^C$  represents the weights from the contextualized word vectors  $r_{t,i}$  to the  $i$ -th score of the attention map.

Finally, by concatenating the mention representation  $m$  and the context representation  $c_t$ , the attentive feature  $x_t$  is formed as the input of LSTM unit:

$$x_t = m \oplus c_t \quad (6)$$

## 2) ATTENTION FEATURE INTEGRATION

In this stage, a LSTM network is adopted to integrate the attentive features, since it can solve sequential modelling by learning patterns with a wider range of temporal dependencies. As shown in the first row of Fig. 6, a basic LSTM unit consists of a single memory cell, an input activation function, and three gates (input gate  $i_t$ , forget gate  $f_t$ , and output gate  $o_t$ ). The input gate  $i_t$  allows an incoming signal to alter the state of the memory cell or to block it. The forget gate  $f_t$  controls what to be remembered and what to be forgotten by the cell and somehow can avoid the gradient from vanishing or exploding when back propagating through time. Finally, the output gate  $o_t$  allows the state of the memory cell to have an effect on the other neurons or to prevent it. Basically, the memory cell and gates in a LSTM block are defined as follows:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} Z \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \tanh(c_t)$$

where  $c_t$  denotes the cell state,  $h_t$  denotes the hidden state, the operator  $\odot$  denotes element-wise multiplication, and  $Z$  denotes the parameters of the LSTM. We use  $\sigma$  and  $\tanh$  to denote the sigmoid activation function and the hyperbolic tangent activation function, respectively.

### 3) ATTENTION MODEL INITIALIZATION

By using a *Multi-Layer Perceptron* (MLP), the cell and the hidden states of LSTM are initialized as follows:

$$c_0 = f_{init,c} \left( \frac{1}{TL} \sum_{t=1}^T \sum_{i=1}^L r_{t,i} \right) \quad (8)$$

$$h_0 = f_{init,h} \left( \frac{1}{TL} \sum_{t=1}^T \sum_{i=1}^L r_{t,i} \right) \quad (9)$$

where  $f_{init,c}$  and  $f_{init,h}$  refer to the functions employed by two MLPs, respectively.  $T$  is the total number of time steps. Specifically, the initial values  $c_0$  and  $h_0$  are used to predict the first attention map  $a_1^c$  for computing the initial attentive feature  $x_1$ .

### 4) FINAL PREDICTION

The hidden state  $h_t$  in Eq. (7) is used as the feature to predict the types of entity mentions. Meanwhile, it also has a guiding effect on the prediction of the attention map at next time step. Through the newly predicated attention maps, the new attentive features will be dynamically merged. Over time steps, the predict process will be recursively implemented. Finally, we average the prediction results over all time steps to obtain the final prediction result.

### C. DIVERSITY CONSTRAINT MODEL

The proposed diversified semantic attention model has been demonstrated that it is capable of exploiting the local and subtle discrimination to distinguish the subtypes, and uses neither prior knowledge nor external resources. However, when the input segments at different time steps are the same, the attention maps produced at each time step may be quite similar. This leads to the result that the attention across different time steps will not gain additional information for improved accuracy of the classification.

To illustrate this issue more intuitively, the attention maps of a sentence generated at different time steps are visualized in Fig. 8, in which vanilla (i.e., non-diversified) attention model and diversified semantic attention model are used separately. Given the same input sentence, the non-diversified attention model always focuses its attention on the same words of the sentence across all time steps (see Fig. 8(a)). Although the attentive words, such as “action-movie” and “star”, are discriminative for recognizing the type of “Arnold Schwarzenegger” from other types, they are insufficient to differentiate it from other semantically similar types such as *director* and *actress*, since the semantic differences among these types are subtle. In addition, it is difficult to learn useful discriminative information for classification as there are few attentive words. Fig. 8(b) shows the diversified semantic attention maps after imposing the multiple attention segments generation and diversity penalty. The semantic attention map at each time step is diversified, from global to local, which is very reasonable for fine-grained entity mentions.

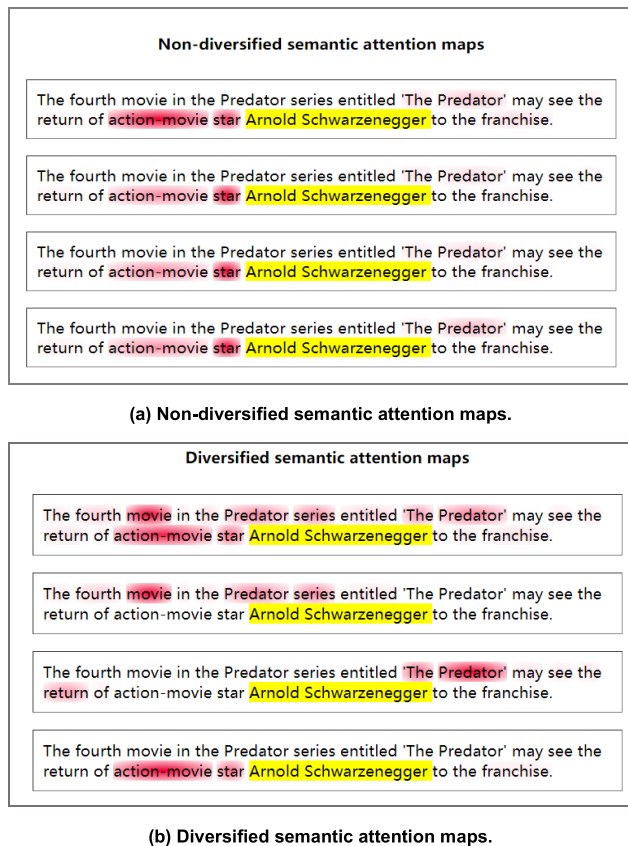


FIGURE 8. Attention maps generated by non-diversified semantic attention model versus diversified semantic attention model.

On the other hand, existing attention-based methods only consider minimizing the classification loss during the attention process, but fail to take the information gain into account. The classification loss function is defined as follows:

$$L_c = - \sum_{i=1}^{N^c} y_i \log \hat{y}_i \quad (10)$$

where  $y_i$  indicates whether the entity mention belongs to type  $i$ ,  $N^c$  is the number of types and  $\hat{y}_i$  is the probability of type  $i$ . Although such a strategy performs well in classifying entities with significant semantic differences, its performance is reduced when the semantic differences among the entities become quite subtle. For addressing this problem, it is necessary to collect sufficient semantic information from multiple features for making the correct decisions on the classification, which requires the attention process to be diverse. Therefore, we propose a diversity constraint model to ensure the diversity of attentive features. Two constraints are jointly considered: the *attention maps constraint* defines the correlation between temporally adjacent attention maps, and the *attention segments constraint* defines the overlapping proportions of the attention segments that are temporally neighbors.

1) ATTENTION MAPS CONSTRAINT

Ignoring the correlation between temporally adjacent attention maps causes that the generated attention maps may be quite similar, which decreases the diversity of the attentive features. Therefore, we propose an intuitive diversity metric to compute the correlation between attention maps that are temporally adjacent, which is defined as follows:

$$\Delta L_{maps} = \frac{1}{T} \sum_{t=2}^T \sum_{i=1}^C a_{t-1}^c \cdot a_{t,i}^c \quad (11)$$

where  $a_{t,i}^c$  is the  $i$ -th attention score of the mention-aware attention map after conducting a softmax on the context words at time  $t$ , and  $T$  is the total number of time steps. In general,  $\Delta L_{maps}$  will get a large value if two neighboring attention maps are similar.

2) ATTENTION SEGMENTS CONSTRAINT

Ignoring the relationship between attention segments that are temporal neighbors leads to the problem that the generated attention segments may largely overlap with each other, and that some discriminative segments will be ignored. Therefore, we impose a segment constraint on the support words of the attention segments, which restrains the overlapped proportion of the attention segments that are temporal neighbors to be smaller than a threshold. The segment constraint is defined as follows:

$$\Delta L_{segments} = \frac{|Supp[S_{t-1}] \cap Supp[S_t]|}{K} \quad (12)$$

where  $Supp[S_t]$  refers to the support words of the attention segment  $S_t$ , which is used to select the attentive words, and  $|Supp[S_{t-1}] \cap Supp[S_t]|$  is the number of intersection words between  $Supp[S_{t-1}]$  and  $Supp[S_t]$ .  $K$  is the length of the original sentence.

3) THE LOSS FUNCTION

The final loss function considers the combination of the classification loss and diversified constraint model, which is defined as follows:

$$L = - \sum_{t=1}^T \sum_{i=1}^{N^c} y_{t,i} \log \hat{y}_{t,i} + \lambda \Delta L_{maps},$$

$$s.t. \Delta L_{segments} < \beta, \forall t = 2, \dots, T \quad (13)$$

where  $y_{t,i}$  indicates the one-hot vector of type probabilities at time step  $t$ ,  $\hat{y}_{t,i}$  indicates the probability of type  $i$  at time step  $t$ ,  $\lambda$  is a penalty coefficient to control the diversity of the neighboring attention maps, and  $\beta$  is a given threshold. The diversity constraint model aims to enhance the diversity of attention, which consists of two items: The first item realized by  $\Delta L_{maps}$  aims to maximize the diversity of the attention maps. The second item realized by  $\Delta L_{segments}$  aims to reduce the overlapping proportions among attention segments that are temporal neighbors, where  $Supp[S_{t-1}] \cap Supp[S_t]$  ensures the generated attention segments have the least overlap.

IV. EXPERIMENTS

In this section, we evaluate the performance of our DSAM approach for fine-grained entity typing. First, the benchmark datasets and the details of the implementation of our DSAM approach are introduced. Then, we perform the model ablation studies to investigate the contribution of each component to the model. In addition, the proposed DSAM approach is compared with the current state-of-the-art methods, and the diversified attention maps produced are visualized in an intuitive way to demonstrate the superiority of the DSAM.

A. DATASETS

Three public datasets are adopted for the experiments. Table 1 summarizes the statistics of the three datasets.

TABLE 1. Statistics of the datasets.

Dataset	FIGER	OntoNotes	BBN
# Types	113	89	47
# Documents	780,549	13,109	2,311
# Sentences	1.51M	143,709	48,899
# Training entity mentions	2.60M	220,000	165,665
# Validation entity mentions	90,000	3,342	18,408
# Testing entity mentions	563	9604	46,018
Max hierarchy depth	2	3	2

**FIGER [7]:** It is the most widely-used dataset for fine-grained entity typing. FIGER contains 2,690,563 entity mentions of 113 different entity types organized in a 2-level hierarchy. It is divided as follows: 2,600,000 entity mentions are for training purposes, 90,000 entity mentions are for validation purposes, and 563 entity mentions are for testing purposes. Specifically, the training and development sets are automatically generated from Wikipedia articles with distant supervision, whereas the testing set is a manually annotated dataset from news reports.

**OntoNotes [21]:** It contains 232,946 entity mentions of 89 news types, among which 220,000 entity mentions are for training purposes, 3,342 entity mentions are for validation purposes, and 9604 entity mentions are for testing purposes. In the training and development data, each entity mention in sentences is automatically linked to Freebase by using DBpedia spotlight. In the test data, each entity mention is manually annotated from 77 news documents of the OntoNotes corpus.

**BBN [35]:** It consists of 48,899 sentences from 2,311 Wall Street Journal articles, which are entirely manually annotated with entity types. One sentence may contain several entity mentions. It is divided as follows: 165,665 entity mentions are for training purposes, 18,408 entity mentions are for validation purposes, and 46,018 entity mentions are for testing purposes.

B. IMPLEMENTATION DETAILS

In this subsection, we will describe the implementation details of our proposed DSAM approach. All sentences were first normalized by a padding strategy to ensure that they were all the same length. For the generation of the attention segment, three different lengths (i.e., 20, 15 and 10) were



used on FIGER and OntoNotes datasets to generate segments from the normalized sentences. For the BBN dataset, two different lengths (i.e., 20 and 15) were used to generate segments. Finally, we obtained large numbers of attention segments for these lengths, and all these attention segments were normalized to the same length. The sequence of these segments was arranged as follows: the first batch of segments was from length of 20, followed by the segments from length of 15, and the last batch of segments was from length of 10.

Our implementation was based on PyTorch framework.<sup>1</sup> In the training phase, we selected different hyper-parameter settings for FIGER, OntoNotes and BBN separately, taking into consideration the differences between the three datasets. The hyper-parameters included the learning rate  $l_r$  for Adam Optimizer, the size of contextualized word representations  $d_w$ , the state size for LSTM layers  $d_s$ , the input dropout keep probability  $p_i$  and the output dropout keep probability  $p_o$  for LSTM layers, and  $L_2$  regularization parameter  $\lambda$ . All the hyper-parameters were selected by the  $k$ -fold cross-validation method. Considering the large scale of the training dataset, we set  $k$  as 10 to ensure a better selection of parameters. We randomly divided the training dataset into 10 mutually exclusive subsets of equal size, and conducted experiment 10 times. Finally, we picked the parameters that obtain the highest classification accuracy. The hyper-parameter settings for FIGER, OntoNotes and BBN datasets are shown in Table 2.

**TABLE 2.** The hyper-parameter settings.

Hyper-parameters	FIGER	OntoNotes	BBN
$l_r$	0.0002	0.0002	0.0005
$d_w$	300	300	300
$d_s$	180	440	200
$p_i$	0.7	0.5	0.5
$p_o$	0.9	0.5	0.5
$\lambda$	1	2	1

### C. COMPARISONS WITH THE CURRENT STATE-OF-THE-ART METHODS

This subsection presents the experimental results and analyses of our DSAM approach as well as the current state-of-the-art methods on three widely-used fine-grained entity typing datasets.

#### 1) COMPARING WITH THE CURRENT STATE-OF-THE-ART METHODS

Table 3 shows the results of the comparisons based on the FIGER, OntoNotes and BBN datasets. To evaluate the performance, Accuracy (Acc), Macro-F1 scores (Ma-F1), and Micro-F1 scores (Mi-F1) [7] were employed as the evaluation metrics in the experiments.

<sup>1</sup><https://pytorch.org/>

**Performance on the FIGER dataset:** Early works [10], [19] chose fixed word embeddings to represent mention as well as its context, and their performances were limited and much lower than our DSAM approach. Our approach was the best among all of the methods based on both Accuracy and Macro-F1 scores. It obtained a 0.54% higher Accuracy when compared with the best results of the NFETC [22] (69.44% vs. 68.9%). It is noted that the labels used in NFETC are handled by a variant of cross entropy loss function during the training phase, while our approach does not do any processing with the labels. Compared to the second highest Macro-F1 score of Attentive +LTR [36], our DSAM approach achieved a 0.29% higher score (83.29% vs. 83.00%). Our DSAM approach achieved a much higher Accuracy and Macro-F1 score than HMGCN [25] but performed worse based on the Micro-F1 score (69.44% vs. 57% in Accuracy, 83.29% vs. 79.8% based on the Macro-F1 score and 81.46% vs. 83.6% based on the Micro-F1 score). This is due to the fact that three GCN models proposed in the HMGCN provide complementary information, which makes its performance based on the Micro F1 score slightly better than ours.

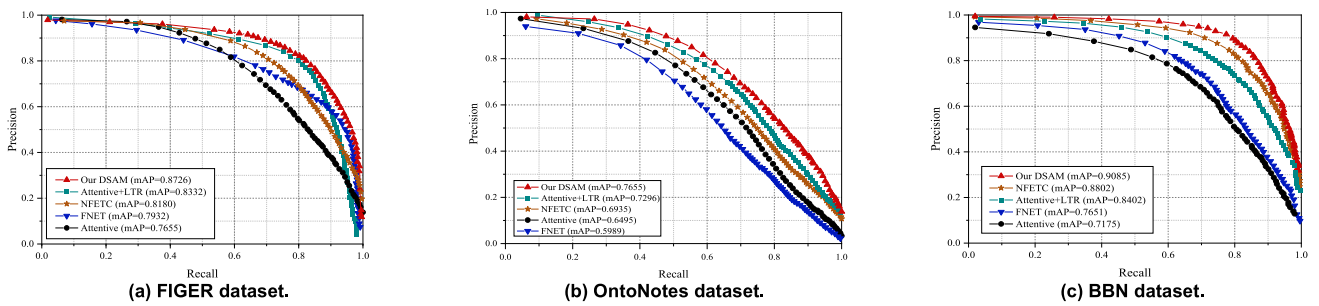
Our approach performed better than the methods that ignore the diversity of attention, such as Attentive [18], NEURAL [19] and Attentive + LTR [36]. In Attentive, an attention-based bi-directional LSTM network was adopted to achieve an Accuracy of 58.97% by composing representations for the context of each mention recursively, which was lower than our approach by 10.47%. Our DSAM approach improved on Accuracy by 9.76% compared to NEURAL, which combines an attentive neural model with hand-crafted features. This verifies the effectiveness of our proposed DSAM in gathering useful information for fine-grained entity typing.

Furthermore, our approach outperforms the methods which select different optimal thresholds for different types, such as MuLR [17], ACT [32] and LABELGCN [33]. Moreover, our approach outperformed methods that use knowledge bases, such as PLE [37], CUTE [38] and APE [24]. Neither knowledge bases nor external resources are used in our DSAM approach, which makes fine-grained entity typing move closer to practical application.

**Performance on OntoNotes and BBN datasets:** When applied to OntoNotes, our DSAM approach outperformed the other methods based on all metrics (66.06% in Accuracy, 83.07% on Macro-F1 scores and 78.19% on Micro-F1 scores respectively). DSAM delivered improvements of 0.86%, 0.17% and 0.25% in Accuracy, Macro F1 score and Micro F1 score, respectively, than the best results of the other methods with which it was compared. The pattern of the results after application to the BBN dataset was similar to those obtained using the FIGER dataset. Our DSAM approach achieved the best results based on the Macro F1 score (82.84%) and Micro F1 score (81.93%) of the current state-of-the-art methods. It achieved an Accuracy close to that of PLE [37]. The superiority of PLE in Accuracy

**TABLE 3.** Comparisons with the current state-of-the-art methods on FIGER, OntoNotes and BBN datasets.

Method	FIGER			OntoNotes			BBN		
	Acc(%)	Ma-F1(%)	Mi-F1(%)	Acc(%)	Ma-F1(%)	Mi-F1(%)	Acc(%)	Ma-F1(%)	Mi-F1(%)
AFET [10]	53.30	69.30	66.40	55.10	71.10	64.70	67.00	72.70	73.50
PLE [37]	59.90	76.30	74.90	57.20	71.50	66.10	<b>68.30</b>	74.40	74.70
CUTE [38]	53.10	74.30	78.20	51.80	67.90	70.20	52.40	68.50	71.70
Attentive [18]	58.97	77.96	74.94	51.74	70.98	64.91	64.72	76.46	71.49
NEURAL [19]	59.68	78.97	75.36	51.74	68.50	63.30	55.17	73.35	69.63
FNET [42]	65.80	81.20	77.40	52.20	68.50	63.30	60.40	74.10	75.70
MuLR [17]	54.80	77.60	81.20	60.80	70.80	67.10	61.00	75.20	77.50
FIGMENT [30]	56.30	78.50	81.90	60.10	71.32	66.60	60.50	74.50	76.90
APE [24]	51.50	72.20	75.60	61.20	73.99	68.40	61.10	76.00	78.20
NFETC [22]	68.90	81.90	79.00	60.20	76.40	70.20	62.85	79.06	78.24
ACT [32]	60.23	78.67	75.52	55.52	73.33	67.61	60.87	77.75	76.94
Attentive+LTR [36]	62.90	83.00	79.80	63.80	82.90	77.30	55.90	79.30	78.10
HS [34]	57.71	72.39	69.57	56.50	71.35	67.70	59.50	76.80	71.80
LABELGCN [33]	60.52	78.57	73.10	59.60	77.80	72.20	59.60	77.80	72.20
HMGCN [25]	57.00	79.80	<b>83.60</b>	65.23	79.68	77.94	63.14	79.86	80.31
Our DSAM approach	<b>69.44</b>	<b>83.29</b>	81.46	<b>66.06</b>	<b>83.07</b>	<b>78.19</b>	67.18	<b>82.84</b>	<b>81.93</b>

**FIGURE 9.** The precision-recall curves.

mainly comes from modeling the type correlation with entity-type facts in knowledge bases, which yields more accurate and complete type correlation statistics compared to our approach.

## 2) TESTING WITH THE PRECISION-RECALL CURVES

Since the Macro F1 and Micro F1 scores both rely on a decision threshold, the precision-recall curves [39] were introduced for more transparent comparisons. As shown in Fig. 9, the data points in each precision-recall curve are based on the validation performance provided by 55 equal-interval thresholds between 0 and 1. From Fig. 9, we can observe that: On the one hand, there are clear margins between our DSAM and the method without attention mechanism (FNET) after they were applied to the three datasets. This indicates the effectiveness of the attention mechanism in capturing discriminative features. On the other hand, our DSAM approach outperformed the other attention-based methods after they were applied to the three datasets, and it verified the superiority of our DSAM. The DSAM integrates the coarse-grained global feature and fine-grained diversified attention feature to boost the discriminative feature learning and enhance their complementarity.

## 3) TESTING AT DIFFERENT TYPE LEVELS

In addition, to further explore the classification performance variance with different type levels, we report the comparisons of the Accuracy at different levels of the entity type hierarchy in Fig. 10. From the results (Fig. 10) of the methods being applied to the three datasets, it is more difficult to distinguish among deeper (more fine-grained) types due to *small* variances among the different subtypes. However, despite this challenge, our DSAM approach still outperformed the other methods with which it was compared, and achieved a 7.8% higher Accuracy than the best result of the Attentive + LTR on level-3 types (45.72% vs. 37.92%). The gain mainly comes from the dynamic modeling of the diversity of attention, which is able to gather discriminative information to the maximal extent possible. In addition, our DSAM approach employs the diversity constraint model to exploit subtle and local discrimination for distinguishing the subtypes.

## D. MODEL ABLATION STUDIES

In this subsection, ablation studies are conducted on the proposed DSAM to evaluate the effectiveness of each individual component.

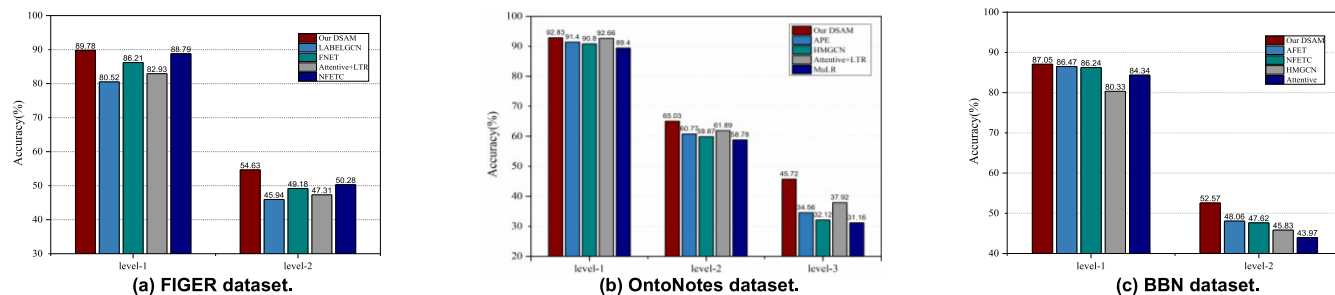


FIGURE 10. Accuracy on different type levels.

TABLE 4. Comparisons of typing results on FIGER, OntoNotes and BBN datasets.

Method	FIGER			OntoNotes			BBN		
	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)
Original	81.25	80.87	80.67	79.30	78.03	79.08	82.21	81.51	80.76
One-length-Att	82.17	81.34	81.09	80.58	79.40	79.57	82.47	81.84	81.02
Two-length-Att	83.09	81.82	82.01	82.45	81.46	80.28	<b>83.06</b>	<b>82.62</b>	<b>81.63</b>
Three-length-Att	<b>84.31</b>	<b>82.29</b>	<b>82.41</b>	<b>83.55</b>	<b>82.60</b>	<b>80.70</b>	82.86	82.35	81.42

1) EFFECTIVENESS OF THE DIVERSIFIED ATTENTION SEGMENTS

In our DSAM approach, the final classification result is generated by merging the prediction results of multiple attention segments across all time steps. Multi-length attention segments with different words contribute to the diversity of attention maps, which makes it possible to capture the semantics of an entity mention from coarse to fine granularity. To verify the effectiveness of the diversified attention segments, the experiments with different combinations of lengths were conducted on the three datasets. In Table 4, “Original” refers to the DSAM trained with the original sentence e, “One-length-Att” refers to the DSAM trained with the attention segments generated by a length of 20, “Two-length-Att” adopted lengths of 20 and 15, and “Three-length-Att” attended to the attention segments with lengths of 20, 15 and 10. The Macro-Precision (Ma-P), Macro-Recall (Ma-R) and the Area Under Curve (AUC) [40] were adopted as the evaluation metrics to comprehensively evaluate the performance, since they can reflect the relationship between the false negative rate and the true positive rate in a classification. From Table 4, we can observe that:

- Overall, the performance of the methods trained with diversified attention segments (i.e., “One-length-Att”, “Two-length-Att” and “Three-length-Att”) were better than “Original” for all metrics after the methods were applied to the three datasets. In particular, “Three-length-Att” obtained a 4.25% improvement in Macro-Precision and a 4.57% improvement in Macro-Recall compared to “Original” after being applied to the OntoNotes dataset. Superior performance of these methods trained with attention segments mainly comes from explicitly diversifying the attentive features, which is able to boost the capturing of discriminative semantics.

- The general pattern of the results after application of the methods to the FIGER and OntoNotes datasets was that the performance of “One-length-Att”, “Two-length-Att” and “Three-length-Att” were significantly improved, when more lengths of segments were involved in the proposed DSAM. Specifically, “Two-length-Att” achieved higher AUC results than “One-length-Att”, i.e., 82.01% vs. 81.09% after application to the FIGER dataset. Compared to “One-length-Att”, “Three-length-Att” improved much on the AUC, i.e., by 1.32% and 1.13% after application to the FIGER and OntoNotes datasets respectively. Comparing with “One-length-Att”, “Two-length-Att” and “Three-length-Att” gained performance from correctly capturing more discriminative words with different lengths. In general, the generated attention segments with long lengths tended to capture the global semantics, while local discriminative semantics will be attended to by short lengths.
- We observe that “Three-length-Att” did not always improve the performance, i.e., the performance of “Two-length-Att” was better than “Three-length-Att” after application to the BBN dataset (83.06% vs. 82.86% in Macro-Precision, 82.62% vs. 82.35% in Macro-Recall, and 81.63% vs. 81.42% in AUC). It shows that the attention segments generated by a short length cannot always provide additional information for fine-grained entity typing. For example, if the attention segments only contain a word such as “movie” or “starring” in a sentence, it is hard to assign the correct types for a certain entity from types such as director and actor, since “movie” and “starring” usually appears in these semantically similar types.

The observations above demonstrate the effectiveness of the proposed diversified attention segments in capturing

**TABLE 5.** Performance of mention-aware attention mechanism on FIGER, OntoNotes and BBN datasets.

Method	FIGER			OntoNotes			BBN		
	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)
Our DSAM	<b>84.31</b>	<b>82.29</b>	<b>82.41</b>	<b>83.55</b>	<b>82.60</b>	<b>80.70</b>	<b>83.06</b>	<b>82.62</b>	<b>81.63</b>
DSAM-dotAtt	82.56	81.93	81.36	81.75	81.31	79.16	82.18	82.04	81.09
DSAM-selfAtt	81.06	81.02	80.13	80.28	80.25	77.90	80.97	81.76	80.78
DSAM-NoAtt	77.04	79.71	77.78	74.88	77.09	76.02	79.93	80.13	79.59

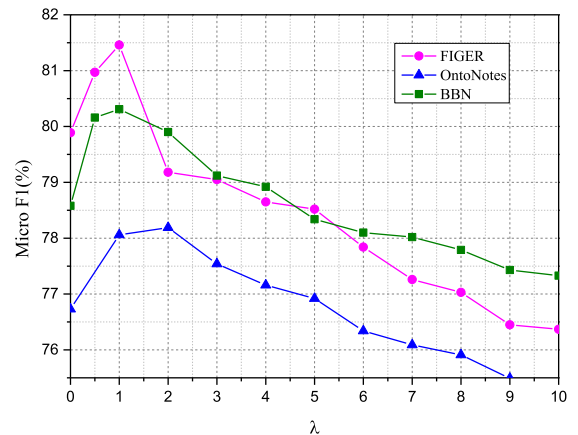
multiple discriminative semantics, which further confirms the superiority of our DSAM approach for fine-grained entity typing.

## 2) EFFECTIVENESS OF THE MENTION-AWARE ATTENTION MECHANISM

Compared with previous attentive methods, which only perform an attention mechanism for the entity mention contexts, we further propose a two-step mention-aware attention mechanism to focus on informative words in mentions and contexts. To study the effectiveness of the proposed mention-aware attention mechanism in our DSAM, we compared it with the DSAM adopting other attention mechanisms, such as the dot-product attention mechanism [41] and the self-attention mechanism [19], which are denoted as “DSAM-dotAtt” and “DSAM-selfAtt”. In addition, the DSAM without the attention mechanism was chosen as another variant of our DSAM, and denoted as “DSAM-NoAtt”. The comparison results after application to the three datasets are summarized in Table 5:

- Compared with DSAM-NoAtt, our DSAM gained performance by extracting the most relevant features via adopting the mention-aware attention mechanism, which assists in capturing the subtle semantic differences among subtypes.
- Based on all the evaluation metrics, our DSAM outperformed the other two variants (i.e., DSAM-dotAtt and DSAM-selfAtt) after they were applied to the three datasets. In particular, our DSAM obtained a 1.75% improvement on the Macro-Precision and a 0.36% enhancement on the Macro-Recall compared to the best variant DSAM-dotAtt after application to the FIGER dataset. The DSAM improved on the Macro-Precision by 3.27% compared to DSAM-selfAtt after application to the OntoNotes dataset. The enhancement mainly comes from jointly combining mention-level attention and context-level attention, which can provide more supplementary semantics compared to the methods adopting other attention mechanisms.

Therefore, the observations verify that the proposed mention-aware attention mechanism is more effective than other methods for fine-grained entity typing, due to the complementary information between an entity mention and its context.

**FIGURE 11.** Performance of our DSAM with different  $\lambda$  on the three datasets.

## 3) EFFECTIVENESS OF THE DIVERSIFIED CONSTRAINT MODEL

To ensure the diversity in the attention process, the diversified constraint model is further employed to drive the prediction of the discriminative attention maps. The attention maps constraint ensures that the generated attention maps have high representativeness, while the attention segments constraint eliminates redundancy and enhances discrimination of the selected segments. Both are jointly employed to exploit the subtle semantic discrimination for distinguishing the subtypes. The effectiveness of the diversified constraint model is verified in the following paragraphs. In Table 6, “DSAM-AMC” refers to the DSAM only performing an attention maps constraint, “DSAM-ASC” refers to the DSAM only adopting an attention segments constraint, and “DSAM-NoDC” refers to the DSAM without considering the diversified constraints model. From Table 6, we can see that the performance of DSAM-AMC was better than DSAM-ASC in terms of all metrics after application to the three datasets, which shows that the effect of the attention maps constraint is stronger than that of the attention segments constraint. In addition, a combination of the two constraints (i.e., DSAM) further improved the classification performance. In particular, our DSAM improved DSAM-AMC’s AUC by 2.29% and DSAM-ASC’s Ma-P by 1.67% after application to the FIGER dataset. Compared to the best variant DSAM-AMS, our DSAM obtained over a 0.86% improvement on Ma-P and over a 1.05% enhancement on Ma-R after application to the OntoNotes dataset. Superior



Time Steps	Diversified Semantic Attention Maps	Gold	Prediction
t1	In an announcement to its staff last week , executives at Time Warner Inc. 's weekly magazine said Time will "dramatically de-emphasize " its use of electronic giveaways such as telephones in television subscription drives ; cut the circulation it guarantees advertisers by 300,000 , to four million ; and increase the cost of its annual subscription rate by about \$ 4 to \$ 55 .		
t2	In an announcement to its staff last week , executives at Time Warner Inc. 's weekly magazine said Time will "dramatically de-emphasize " its use of electronic giveaways such as telephones in television subscription drives ; cut the circulation it guarantees advertisers by 300,000 , to four million ; and increase the cost of its annual subscription rate by about \$ 4 to \$ 55 .		
t3	In an announcement to its staff last week , executives at Time Warner Inc. 's weekly magazine said Time will "dramatically de-emphasize " its use of electronic giveaways such as telephones in television subscription drives ; cut the circulation it guarantees advertisers by 300,000 , to four million ; and increase the cost of its annual subscription rate by about \$ 4 to \$ 55 .	/organization /organization/company /organization/company/news	/organization /organization/company /organization/company/news
t4	In an announcement to its staff last week , executives at Time Warner Inc. 's weekly magazine said Time will "dramatically de-emphasize " its use of electronic giveaways such as telephones in television subscription drives ; cut the circulation it guarantees advertisers by 300,000 , to four million ; and increase the cost of its annual subscription rate by about \$ 4 to \$ 55 .		
t5	In an announcement to its staff last week , executives at Time Warner Inc. 's weekly magazine said Time will "dramatically de-emphasize " its use of electronic giveaways such as telephones in television subscription drives ; cut the circulation it guarantees advertisers by 300,000 , to four million ; and increase the cost of its annual subscription rate by about \$ 4 to \$ 55 .		

(a) OntoNotes dataset.

Time Steps	Diversified Semantic Attention Maps	Gold	Prediction
t1	But , when coupled with the ACL tear suffered by forward Marjie Heard earlier this week and the increasing likelihood that freshman forward Talia Walton will miss the rest of the season and pursue a medical redshirt , the Huskies might be left with just two players taller than 6-foot that are ready to go against Washington State .		
t2	But , when coupled with the ACL tear suffered by forward Marjie Heard earlier this week and the increasing likelihood that freshman forward Talia Walton will miss the rest of the season and pursue a medical redshirt , the Huskies might be left with just two players taller than 6-foot that are ready to go against Washington State .		
t3	But , when coupled with the ACL tear suffered by forward Marjie Heard earlier this week and the increasing likelihood that freshman forward Talia Walton will miss the rest of the season and pursue a medical redshirt , the Huskies might be left with just two players taller than 6-foot that are ready to go against Washington State .	/organization /organization/sports_team	/organization /organization/sports_team
t4	But , when coupled with the ACL tear suffered by forward Marjie Heard earlier this week and the increasing likelihood that freshman forward Talia Walton will miss the rest of the season and pursue a medical redshirt , the Huskies might be left with just two players taller than 6-foot that are ready to go against Washington State .		
t5	But , when coupled with the ACL tear suffered by forward Marjie Heard earlier this week and the increasing likelihood that freshman forward Talia Walton will miss the rest of the season and pursue a medical redshirt , the Huskies might be left with just two players taller than 6-foot that are ready to go against Washington State .		

(b) FIGER dataset.

Time Steps	Diversified Semantic Attention Maps	Gold	Prediction
t1	The Chicago Mercantile Exchange , a major futures marketplace , yesterday announced the addition of another layer of trading halts designed to slow program traders during a rapidly falling stock market , and the Big Board is expected today to approve some additional restrictions on program trading .		
t2	The Chicago Mercantile Exchange , a major futures marketplace , yesterday announced the addition of another layer of trading halts designed to slow program traders during a rapidly falling stock market , and the Big Board is expected today to approve some additional restrictions on program trading .		
t3	The Chicago Mercantile Exchange , a major futures marketplace , yesterday announced the addition of another layer of trading halts designed to slow program traders during a rapidly falling stock market , and the Big Board is expected today to approve some additional restrictions on program trading .	/organization /organization/corporation	/organization /organization/corporation
t4	The Chicago Mercantile Exchange , a major futures marketplace , yesterday announced the addition of another layer of trading halts designed to slow program traders during a rapidly falling stock market , and the Big Board is expected today to approve some additional restrictions on program trading .		
t5	The Chicago Mercantile Exchange , a major futures marketplace , yesterday announced the addition of another layer of trading halts designed to slow program traders during a rapidly falling stock market , and the Big Board is expected today to approve some additional restrictions on program trading .		

(c) BBN dataset.

**FIGURE 12.** Three sentences and their attention maps at different time steps. The entity mentions are highlighted in yellow, while the attention segments (bold typeface) are highlighted in red. Darker background color indicates higher attention score.

**TABLE 6.** The performance of the diversified constraint model on FIGER, OntoNotes and BBN datasets.

Method	FIGER			OntoNotes			BBN		
	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)	Ma-P(%)	Ma-R(%)	AUC(%)
Our DSAM	<b>84.31</b>	<b>82.29</b>	<b>82.41</b>	<b>83.55</b>	<b>82.60</b>	<b>80.70</b>	<b>83.06</b>	<b>82.62</b>	<b>81.63</b>
DSAM-AMC	83.40	81.37	80.12	82.93	81.74	79.65	82.02	81.91	80.08
DSAM-ASC	82.64	80.66	79.09	81.46	80.56	79.29	81.53	80.17	79.21
DSAM-NoDC	80.26	79.40	78.50	79.54	78.82	78.72	78.92	78.26	77.62

performance of our DSAM demonstrates the effectiveness of the proposed diversified constraint model in promoting selection of the discriminative features.

#### 4) EFFECTIVENESS OF PARAMETER $\lambda$

Furthermore, Fig. 11 shows the performance sensitivity of DSAM with respect to  $\lambda$ —the tuning parameter in the

diversified constraint model to control the diversity of the neighboring attention maps—after application to the three datasets.

One can see that the performance of our DSAM improves as  $\lambda$  increases, and dramatically decreases as  $\lambda$  is large than 1, 2 and 1 after application to the three datasets respectively. In particular, our DSAM achieved the highest Micro F1 scores (81.46%, 78.19% and 80.31% respectively) after application to the three datasets when  $\lambda = 1$ ,  $\lambda = 2$  and  $\lambda = 1$  respectively. Consequently, we set  $\lambda = 1$ ,  $\lambda = 2$  and  $\lambda = 1$  throughout the experiments above for our DSAM when applied to the FIGER, OntoNotes and BBN datasets respectively.

## V. VISUALIZATION OF DIVERSIFIED SEMANTIC ATTENTION

We visualized the diversified attention maps generated through our DSAM approach for several instances selected from the three datasets, as shown in Fig. 12. Overall, the diversified attention maps were correctly captured across the different time steps. From Fig. 12(a), we can see that the attentive words such as “staff” and “executives” were first captured, and then “magazine” and “telephones” were focused on to help classify “Time” as *organization/company*. In the last few time steps, local discriminative words such as “television subscription”, “advertisers” and “annual subscription” received more attention, which assists in determining the subtype to be *news* instead of *broadcast*. Similarly, in Fig. 12(b), the attentive words such as “ACL” and “forward” were observed first. The discriminative words such as “season” and “players” were then sequentially captured from the next attention segments. For Fig. 12(c), the global semantics were first captured from the long-length attention segments. After shortening the attention segments, “traders”, “stock”, and “trading” were observed as well. Finally, the fine-grained types *organization/corporation* were assigned to the mention “Big Board”.

## VI. CONCLUSION

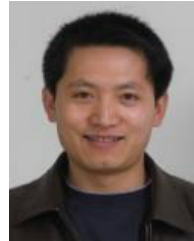
In this paper, we propose a diversified semantic model (DSAM) for fine-grained entity typing, which explicitly diversifies the semantic attentions for capturing multiple discriminative information. It dynamically captures important information at each time step through a recurrent mention-aware attention mechanism. As a second contribution, a diversified constraint model is proposed to drive the generation of attention maps, which combines two constraints: the attention maps constraint aims to ensure the high representativeness of generated attention maps, and the attention segments constraint aims to eliminate redundancy and highlights discrimination of the attention segments. Combination of both constraints yields a promotion on the capturing of the subtle and local discrimination. Importantly, our DSAM avoids using any prior knowledge and external resources to move closer to a practical application. Extensive experimental results demonstrate the effectiveness and

robustness of the proposed DSAM approach, when compared with over 10 current state-of-the-art approaches after being applied to three widely-used datasets.

## REFERENCES

- [1] H. Zhu, C. He, Y. Fang, and W. Xiao, “Fine grained named entity recognition via Seq2seq framework,” *IEEE Access*, vol. 8, pp. 53953–53961, 2020, doi: [10.1109/ACCESS.2020.2980431](https://doi.org/10.1109/ACCESS.2020.2980431).
- [2] Y. Lou, T. Qian, F. Li, and D. Ji, “A graph attention model for dictionary-guided named entity recognition,” *IEEE Access*, vol. 8, pp. 71584–71592, 2020, doi: [10.1109/ACCESS.2020.2987399](https://doi.org/10.1109/ACCESS.2020.2987399).
- [3] X. Man and P. Yang, “Fine-grained Chinese named entity recognition in entertainment news using adversarial multi-task learning,” in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2019, pp. 1671–1675.
- [4] J. Li, A. Sun, and Y. Ma, “Neural named entity boundary detection,” *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: [10.1109/TKDE.2020.2981329](https://doi.org/10.1109/TKDE.2020.2981329).
- [5] C. Lee, Y.-G. Hwang, H.-J. Oh, S. Lim, J. Heo, C.-H. Lee, H.-J. Kim, J.-H. Wang, and M.-G. Jang, “Fine-grained named entity recognition using conditional random fields for question answering,” in *Proc. Asia Inf. Retr. Symp. (AIRS)*, Singapore, 2006, pp. 581–587.
- [6] T. Lin and O. Etzioni, “No noun phrase left behind: Detecting and typing unlinkable entities,” in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (CoNLL EMNLP)*, Jeju Island, South Korea, 2012, pp. 893–903.
- [7] X. Ling and D. S. Weld, “Fine-grained entity recognition,” in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, 2012, pp. 94–100.
- [8] H. Lin, L. Sun, and X. Han, “Reasoning with heterogeneous knowledge for commonsense machine comprehension,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2032–2043.
- [9] Y. Jia, W. Xu, P. Qin, and Z. Bao, “Fine-grained entity typing for knowledge base completion,” in *Proc. IEEE Int. Conf. Netw. Infrastruct. Digit. Content (IC-NIDC)*, Beijing, China, Sep. 2016, pp. 361–365.
- [10] X. Ren, W. He, M. Qu, L. Huang, H. Ji, and J. Han, “AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 1369–1378.
- [11] E. Choi, O. Levy, Y. Choi, and L. Zettlemoyer, “Ultra-fine entity typing,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, vol. 1, 2018, pp. 87–96.
- [12] J. Xin, Y. Lin, Z. Liu, and M. Sun, “Improving neural fine-grained entity typing with knowledge attention,” in *Proc. AAAI Conf. AI (AAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 5997–6004.
- [13] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, New Orleans, LA, USA, 2018, pp. 2227–2237.
- [14] U. Naseem and K. Musial, “DICE: Deep intelligent contextual embedding for Twitter sentiment analysis,” in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sydney, NSW, Australia, Sep. 2019, pp. 953–958.
- [15] U. Naseem, S. K. Khan, I. Razzak, and I. A. Hammed, “Hybrid words representation for airlines sentiment analysis,” in *Proc. Australas. Joint Conf. AI*, Cham, Switzerland: Springer, 2019, pp. 381–392.
- [16] U. Naseem, I. Razzak, K. Musial, and M. Imran, “Transformer based deep intelligent contextual embedding for Twitter sentiment analysis,” *Future Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020, doi: [10.1016/j.future.2020.06.050](https://doi.org/10.1016/j.future.2020.06.050).
- [17] Y. Yaghoobzadeh and H. Schütze, “Multi-level representations for fine-grained typing of knowledge base entities,” in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, vol. 1, 2017, pp. 578–589.
- [18] S. Shimaoka, P. Stenetorp, K. Inui, and S. Riedel, “An attentive neural architecture for fine-grained entity type classification,” in *Proc. 5th Workshop Automated Knowl. Base Construct.*, 2016, pp. 1–6. [Online]. Available: <https://arxiv.org/abs/1604.05525>
- [19] S. Shimaoka, P. Stenetorp, K. Inui, and S. Riedel, “Neural architectures for fine-grained entity type classification,” in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, vol. 1, 2017, pp. 1271–1280.
- [20] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP)*, Singapore, vol. 2, 2009, pp. 1003–1011.

- [21] D. Gillick, N. Lazic, K. Ganchev, J. Kirchner, and D. Huynh, "Context-dependent fine-grained entity type tagging," 2014, *arXiv:1412.1820*. [Online]. Available: <http://arxiv.org/abs/1412.1820>
- [22] P. Xu and D. Barbosa, "Neural fine-grained entity type classification with hierarchy-aware loss," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA, vol. 1, 2018, pp. 16–25.
- [23] Q. Ren, "Fine-grained entity typing with hierarchical inference," in *Proc. IEEE 4th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Chongqing, China, vol. 1, Jun. 2020, pp. 2552–2558, doi: [10.1109/ITNEC48623.2020.9085112](https://doi.org/10.1109/ITNEC48623.2020.9085112).
- [24] H. Jin, L. Hou, J. Li, and T. Dong, "Attributed and predictive entity embedding for fine-grained entity typing in knowledge bases," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, Santan Fe, NM, USA, 2018, pp. 282–292.
- [25] H. Jin, L. Hou, J. Li, and T. Dong, "Fine-grained entity typing via hierarchical multi graph convolutional networks," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 4970–4979.
- [26] L. Dong, F. Wei, H. Sun, M. Zhou, and K. Xu, "A hybrid neural model for type classification of entity mentions," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, 2015, pp. 1243–1249.
- [27] S. Kam, U. Wltinger, and H. Schütze, "End-to-end trainable attentive decoder for hierarchical entity classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, vol. 2, 2017, pp. 752–758.
- [28] J. Liu, L. Wang, M. Zhou, J. Wang, and S. Lee, "Fine-grained entity type classification with adaptive context," *Soft Comput.*, vol. 22, no. 13, pp. 4307–4318, Jul. 2018.
- [29] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, and J. Xie, "A hierarchy-to-sequence attentional neural machine translation model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 623–632, Mar. 2018, doi: [10.1109/TASLP.2018.2789721](https://doi.org/10.1109/TASLP.2018.2789721).
- [30] Y. Yaghoobzadeh, H. Adel, and H. Schuetze, "Corpus-level fine-grained entity typing," *J. Artif. Intell. Res.*, vol. 61, pp. 835–862, Apr. 2018, doi: [10.1613/jair.5601](https://doi.org/10.1613/jair.5601).
- [31] D. Yogatama, D. Gillick, and N. Lazic, "Embedding methods for fine grained entity type classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, vol. 2, 2015, pp. 291–296.
- [32] S. Zhang, K. Duh, and B. Van Durme, "Fine-grained entity typing through increased discourse context and adaptive classification thresholds," 2018, *arXiv:1804.08000*. [Online]. Available: <http://arxiv.org/abs/1804.08000>
- [33] W. Xiong, J. Wu, D. Lei, M. Yu, S. Chang, X. Guo, and W. Y. Wang, "Imposing label-relational inductive bias for extremely fine-grained entity typing," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL)*, Minneapolis, MI, USA, vol. 1, 2019, pp. 773–784.
- [34] F. López, B. Heinzerling, and M. Strube, "Fine-grained entity typing in hyperbolic space," in *Proc. 4th Workshop Represent. Learn. NLP (RePL NLP)*, Florence, Italy, 2019, pp. 169–180.
- [35] R. Weischedel and A. Brunstein, "BBN pronoun coreference and entity type corpus," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. LDC2005T33, 2005, doi: [10.35111/9fx9-gz10](https://doi.org/10.35111/9fx9-gz10).
- [36] Y. Lin and H. Ji, "An attentive fine-grained entity typing model with latent type representation," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 6197–6202.
- [37] X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, and J. Han, "Label noise reduction in entity typing by heterogeneous partial-label embedding," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 1825–1834.
- [38] B. Xu, Y. Zhang, J. Liang, Y. Xiao, S. Hwang, and W. Wang, "Cross-lingual type inference," in *Proc. Int. Conf. Database Syst. Adv. Appl. (DASFAA)*, Dallas, TX, USA, 2016, pp. 447–462.
- [39] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, Prague, Czech Republic, 2013, pp. 451–466, doi: [10.1007/978-3-642-40994-3\\_29](https://doi.org/10.1007/978-3-642-40994-3_29).
- [40] S. Narkhede, "Understanding AUC-ROC curve," *Towards Data Sci.*, vol. 2018, p. 26, Jun. 2018.
- [41] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1412–1421.
- [42] A. Abhishek, A. Anand, and A. Awekar, "Fine-grained entity type classification by jointly learning representations and label embeddings," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, vol. 1, 2017, pp. 797–807.



**YANFENG HU** received the B.S. degree from Xidian University, Xi'an, China, in 1999, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2005. He is currently a Researcher with the Institute of Electronics, Chinese Academy of Sciences, Suzhou, China. His research interests include natural language processing and remote sensing image understanding.



**XUE QIAO** received the M.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2016. She is currently a Technical Researcher with the Institute of Electronics, Chinese Academy of Sciences, Suzhou, China. Her current research interests include natural language processing, text mining, and knowledge graph.



**LUO XING** received the B.E. degree from the Jiangxi University of Finance and Economics, Nanchang, in 2018. He is currently pursuing the M.E. degree with the School of Software and Engineering, University of East China Normal University, Shanghai. His research interests include data mining and recommendation systems.



**CHEN PENG** received the B.Sc. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2015. He is currently an Associate Professor with the Institute of Electronics, Chinese Academy of Sciences, Suzhou. His research interests include natural-language understanding, knowledge graph, and spatio-temporal data mining.

• • •