IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 16, 2019, accepted July 20, 2019, date of publication July 30, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932041

Distant Supervision for Relation Extraction via Piecewise Attention and Bag-Level Contextual Inference

VAN-THUY PHI[1,4], JOAN SANTOSO[2,3], (Member, IEEE), VAN-HIEN TRAN[1], HIROYUKI SHINDO[1,4], MASASHI SHIMBO[1,4], AND YUJI MATSUMOTO[1,4]

[1]Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan
[2]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[3]Department of Informatics, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya 60284, Indonesia
[4]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan

Corresponding authors: Van-Thuy Phi (phi.thuy.ph8@is.naist.jp) and Joan Santoso (joan@stts.edu)

This work was supported in part by the Japan Science and Technology Agency (JST) CREST under Grant JPMJCR1513, in part by the Japan Society for the Promotion of Science (JSPS) Kakenhi under Grant 15H02749, and in part by the Japan Society for the Promotion of Science (JSPS) Kakenhi under Grant JP18K18109.

ABSTRACT Distant supervision (DS) has become an efficient approach for relation extraction (RE) to alleviate the lack of labeled examples in supervised learning. In this paper, we propose a novel neural RE model that combines a bidirectional gated recurrent unit model with a form of hierarchical attention that is better suited to RE. We demonstrate that an additional attention mechanism called *piecewise attention*, which builds itself upon segment level representations, significantly enhances the performance of the distantly supervised relation extraction task. Our piecewise attention mechanism not only captures crucial segments in each sentence but also reflects the direction of relations between two entities. Furthermore, we propose a *contextual inference* method that can infer the most likely positive examples of an entity pair in bags with very limited contextual information. In addition, we provide an annotated dataset without *false positive* examples based on the Riedel testing dataset, and report on the actual performance of several RE models. The experimental results show that our proposed methods outperform the previous state-of-the-art baselines on both original and annotated datasets for the distantly supervised RE task.

INDEX TERMS Relation extraction, distant supervision, piecewise attention, bidirectional gated recurrent unit (BiGRU).

I. INTRODUCTION

Distant supervision (DS) is a class of weakly supervised methods [1] and has become a popular approach for relation extraction (RE) to alleviate the lack of labeled examples in supervised learning. DS is an effective approach to scale RE to very large corpora that contain thousands of relations without any labels on the text.

The term "*distant supervision*" was formally used by Mintz *et al.* [2] as a method of utilizing existing structured facts for obtaining training data without the manual labeling of examples. For the RE task, DS makes use of an already existing knowledge base (KB) such as Freebase or a domain-specific KB to label entity pairs automatically in the text. This is then used to extract features and train a machine learning classifier. The original "*DS assumption*" is that *if two entities participate in a known Freebase relation, any sentence that contains these two entities might express that relation*. For example, Freebase contains the fact that <*Tokyo, is the capital of, Japan*>. We consider this fact and label each pair of "*Tokyo*" and "*Japan*" that appear in the same sentence as a positive example for the "*/location/country/capital*" relation. By aligning KB facts with texts, DS provides coherent positive training examples and avoids the high cost and human effort of manual annotation. Such large datasets allow for learning more complex models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). However, DS often introduces noise to the generated training data. This approach can generate *false positives*, as not every mention of an entity pair in a sentence means that a relation is also expressed. As a result, DS is still limited by the quality of the training data, and noise existing in positively

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

labeled data may affect the performance of the supervised learning.

Recently, neural networks have been widely explored in distantly supervised RE and achieved state-of-the-art results. Zeng *et al.* [3] treated RE as a problem of multi-instance learning to relax the strong assumption of DS: they assumed that ''at least one document in the bag expresses the relation of the entity pair.'' Then, they divided the original input sentence into three segments by the positions of two entities and used piecewise max pooling to automatically learn relevant features using a piecewise CNN (*PCNN*). Lin *et al.* [4] addressed the shortcoming of the previous model, which only used the most relevant sentence from the bag. They proposed using sentence-level attention to capture the importance of each sentence, and then leveraging large amounts of useful data and information that is expressed by all sentences in each bag. Currently, *PCNN+ATT*, proposed by Lin *et al.* [4], is one of the state-of-the-art neural-network-based RE models.

In this study, we propose a novel neural RE model that combines a bidirectional gated recurrent unit (BiGRU) sequence model with a form of hierarchical attention that is better suited to RE. Our model consists of two attention modules: a piecewise attention that builds itself upon segment-level representations, and a sentence-level attention that builds itself upon sentence-level representations in each bag. Our piecewise attention not only captures crucial segments in each sentence but also reflects forward and backward directions of a sentence for better understanding the target relations between two entities.

The primary goal of RE under DS is to determine the relation for a given bag, i.e., between a given pair of entities. Hence, we propose using a contextual inference method that can infer the most likely positive examples of an entity pair in bags with very limited contextual information (i.e., for a bag with only a few sentences). Our inference method increases the number of positive examples and intentionally covers more contexts for target bags by using the similarity between entity pairs in positively labeled data. In addition, we provide an annotated dataset for the distantly supervised RE task, which is based on the most commonly used dataset developed by Riedel *et al.* [5], and report on the actual performance of several RE models. All experimental results show that our proposed methods outperform previous state-of-the-art baselines on both original and annotated datasets.

Our contributions can be summarized as follows: (a) a novel BiGRU model combined with an additional attention mechanism called *piecewise attention* for distantly supervised RE; (b) a contextual inference method for improving bag label prediction; (c) an annotated dataset of 5,863 sentences,[1] which is checked by annotators for false positive examples; and (d) experimental results showing that the proposed models outperform various state-of-the-art baselines on both original and annotated datasets for the distantly supervised RE task.

II. DISTANTLY SUPERVISED RE TASK

A. BACKGROUND

The original assumption of DS [2] indicated that all sentences containing a known relation (e.g., in Freebase) might be potential *true positive* relation mentions. This assumption is too strong and causes the issue of incorrect labels. Consequently, it will deteriorate the performance of a model trained on such noisy data. *At-least-one* models make a relaxed DS assumption [5]: *if two entities participate in a relation, at least one sentence that mentions these two entities might express that relation*. In this case, at least one mention is considered as a true positive.

Ridel *et al.* [5], Hoffmann *et al.* [6], and Surdeanu *et al.* [7] introduced a series of models casting DS as a multiple-instance learning problem [8]. In this multi-instance setting, the training set contains many entity-pair bags, and each bag consists of many *relation mentions*. Each relation mention is an occurrence of a pair of entities with the source sentence .[2] The labels of the bags are known; however, the labels of the relation mentions in these bags are unknown.

A DS system has several key differences from traditional supervised RE systems. First, the primary goal of a DS system is to determine whether a relation between a given pair of entities is expressed *somewhere* in the text, and not necessarily where it is expressed [5]. In other words, a DS system should predict labels for *relations* (i.e., entity pair labels), not *relation mentions* (i.e., sentence labels). By contrast, the objective of standard supervised RE systems is to classify relation mentions (i.e., a sentence mentioned a specific entity pair). One of the most important benefits of focusing on relations instead of relation mentions is that it allows us to aggregate evidence for a relation from several places in the corpus. Second, in standard supervised learning, the gold annotations of all training sentences are given, whereas in DS, only entity pair labels are provided. This, however, may serve as a challenge because DS generates many noisy mentions that do not support target relations.

B. PROBLEM DEFINITION

1) DISTANT SUPERVISION (DS)

We are given a corpus C and a KB K that contains known tuples (e_1, r, e_2) in which $r \in R$ (the set of relations we are interested in) and (e_1, e_2) is an entity pair that expresses the relation r. The labeling procedure of DS is as follows: we align K to C; and for a tuple (e_1, r, e_2) in K, all sentences (relation mention candidates) in C that simultaneously mention both entities e_1 and e_2 constitute a bag and are deemed as having the relation r. This generates a dataset that has labels on the entity-pair (bag) level with (possibly noisy) positive examples. Previous works typically assumed that if the argument entity pair (e_1, e_2) does not appear in K

[1] We release our annotated dataset at: https://github.com/pvthuy/distantly-supervised-RE

[2] We used the original term *relation mention* as used in [5].

as holding a relation, all of the corresponding relation mentions in C are automatically annotated as negative examples (i.e., with "*NA*" labels).

2) DISTANTLY SUPERVISED RE

The distantly supervised RE task can be formalized as follows: We are given a training set T that contains N entity-pair bags $(B_1, B_2, \ldots, B_N)$. The n-th bag consists of n_b sentences (or relation mentions) $\{x_1, x_2, \ldots, x_{n_b}\}$ and the relation label r for a given entity pair (e_1, e_2). An RE model M is trained with training set T to select valid sentences based on r for each bag. In the testing phase, our goal is to predict which relation types are expressed in the unseen bags, given all sentences in which both entities are mentioned in a large collection of unlabeled documents.

III. METHODOLOGY

The distantly supervised RE task is formulated as multi-instance learning. In this section, we introduce a novel neural RE model that combines a BiGRU sequence model with a form of hierarchical attention that effectively incorporates the *piecewise* and *sentence-level* attentions. Furthermore, we propose to use a *contextual inference* method that can infer the most likely positive examples of an entity pair in bags with limited contextual information without using any external knowledge resources or human annotations.

Our model takes input as an entity pair (e_1, e_2) and a bag $B = \{x_1, x_2, \ldots, x_{n_b}\}$ for (e_1, e_2), and predicts the probability $p(r|e1, e2)$ corresponding to the relation label r, $\forall r \in R$ (R is the set of relation labels). Our model consists of two main components:

- **Sentence Encoder** Given a sentence in $x \in B$, which contains two target entities, the sentence encoder outputs a distributed representation $\mathbf{x}$ of the sentence.
- **Bag Encoder** Given the encoding of each sentence in the bag for the entity pair (e_1, e_2), the bag encoder aims to learn a representation of the given bag, which is fed to a softmax classifier.

We briefly present the components of our model below. Each component will be described in detail in subsequent sections.

A. SENTENCE ENCODER

The overall architecture of the sentence encoder is depicted in Fig. 1, with the original sentence as the input to our model. Our sentence encoder has an embedding layer, two BiGRU layers, and a piecewise attention layer. These key modules are analyzed as follows.

1) EMBEDDING LAYER

Following previous work, we transform each input word of the source sentence into a combination of *word embedding* and *position embedding* in the embedding layer.

Word embeddings (WEs) aim to represent words as low-dimensional dense vectors. They can capture syntactic and semantic properties of words, such as in [9]. An embedding lookup table is first used to map words in the sentence into real-valued vectors. Word representations are encoded by column vectors in an embedding matrix $\mathbf{E} \in \mathbb{R}^{d^w \times |V|}$, where d^w is the dimensionality of the embedding space and $|V|$ is the size of the vocabulary.

Position embedding (PE) [10] is used to specify the positional information of the current word with respect to two target entities e_1 and e_2. Therefore, we define two lookup tables with two position embedding matrices $\mathbf{P_1}$ and $\mathbf{P_2}$, where $\mathbf{P_i} \in \mathbb{R}^{d^p \times |L|}$ (L is the maximum distance between any words of the sentence and two entities, and d^p is the dimension of the position embedding). $\mathbf{P_1}$ and $\mathbf{P_2}$ are randomly initialized. We then transform each relative distance (from the i-th word to e_1 or e_2) into a real-valued vector by looking up the position embedding matrices.

We concatenate the word and position embeddings as the input of the network. For a given sentence composed of k words, $x = \{w_1, w_2, \ldots, w_k\}$, we transform each word w_i into a real-valued vector. Then, x is fed into the next layer as $\mathbf{e}^x = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$. If the size of the word representation is d_w and that of the position representation is d_p, then the size of a word vector is $d_w + 2d_p$.

2) 1ST BiGRU LAYER

The role of the sentence encoder is to read the input sentence and construct an informative sentence representation. RNNs have been widely exploited to deal with variable-length sequence input. RNNs can learn long dependencies, but in practice they tend to be biased toward their most recent inputs in the sequence [11]. Long short-term memory networks (LSTMs) [12] incorporate a memory cell to combat this issue and avoid the vanishing gradient problem.

A gated recurrent unit (GRU) [13] is a simpler variant of the LSTM and was found to achieve better performance than the LSTM on some tasks [14]. A single-direction GRU has one drawback of not using the contextual information from the future words. A BiGRU exploits both the previous and future contexts by processing the sequence in two directions, and generates two independent sequences of GRU output vectors. Given the input sequence $\mathbf{e}^x = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$, we employ a BiGRU as the recurrent unit, where the GRU is defined as

$$\mathbf{z}_i = \sigma(\mathbf{W}_z[\mathbf{e}_i; \mathbf{h}_{i-1}]), \tag{1}$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r[\mathbf{e}_i; \mathbf{h}_{i-1}]), \tag{2}$$

$$\tilde{\mathbf{h}}_i = \tanh(\mathbf{W}_h[\mathbf{e}_i; \mathbf{r}_i \odot \mathbf{h}_{i-1}]), \tag{3}$$

$$\mathbf{h}_i = (1 - \mathbf{z}_i) \odot \mathbf{h}_{i-1} + \mathbf{z}_i \odot \tilde{\mathbf{h}}_i, \tag{4}$$

where $\mathbf{W}_z$, $\mathbf{W}_r$ and $\mathbf{W}_h$ are weight matrices, σ is a sigmoid function, and $\odot$ is an element-wise multiplication operator. Initially, for $t = 0$, the output vector is $\mathbf{h}_0 = 0$.

Inspired by the *PCNN* model [3], we divide the original input sentence x into three segments by the positions of two entities e_1 and e_2. Fig. 1 illustrates these three segments, namely *PRED*, *MID*, and *POST* in our model. Let *En*1*pos* and *En*2*pos* be the positions of two entities in x. The input

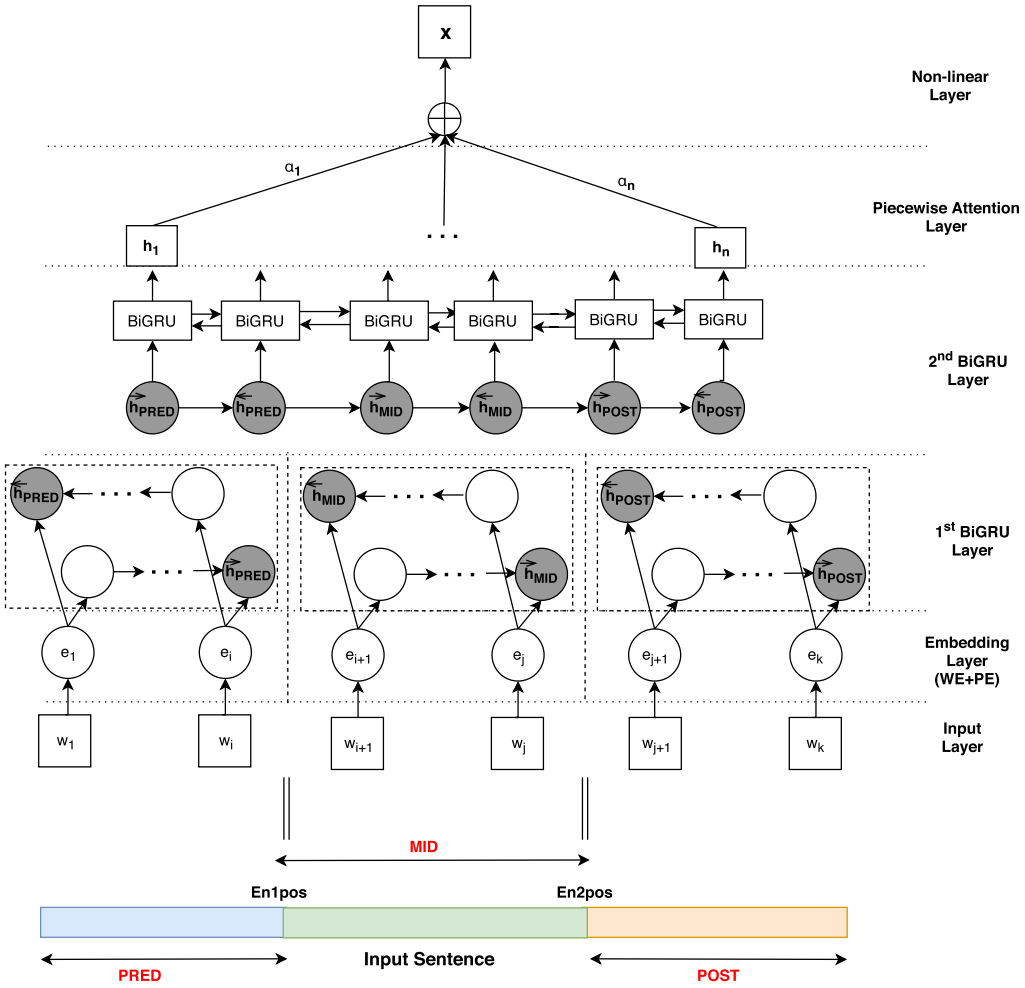


FIGURE 1. Architecture of our BiGRU model with piecewise attention used for sentence encoder.

sequence $\mathbf{e}^x = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$ of the BiGRU layer is divided into three independent subsequences:

$$\mathbf{e}_x^{\text{PRED}} = [\mathbf{e}_1, \ldots, \mathbf{e}_{En1pos}], \tag{5}$$

$$\mathbf{e}_x^{\text{MID}} = [\mathbf{e}_{En1pos}, \ldots, \mathbf{e}_{En2pos}], \tag{6}$$

$$\mathbf{e}_x^{\text{POST}} = [\mathbf{e}_{En2pos}, \ldots, \mathbf{e}_k]. \tag{7}$$

The repetitions of entities in Eq. (5), Eq. (6), and Eq. (7) mark the opening or closing of a coherent piece of text, and help our models extract informative distinct features over these adjacent text spans. Then, the first BiGRU layer processes each segment (PRED| MID| POST) separately. Concretely, the BiGRU consists of a forward GRU and a backward GRU. The forward GRU reads the input from left to right and generates a sequence of hidden states, e.g., $(\overrightarrow{\mathbf{h}}_1, \ldots, \overrightarrow{\mathbf{h}}_{En1pos})$ for $\mathbf{e}_x^{\text{PRED}}$. The backward GRU reads the input in reverse from right to left, and results in another sequence of hidden states, e.g., $(\overleftarrow{\mathbf{h}}_1, \ldots, \overleftarrow{\mathbf{h}}_{En1pos})$ for $\mathbf{e}_x^{\text{PRED}}$. The i-th hidden state is defined as

$$\overrightarrow{\mathbf{h}}_i = \text{GRU}(\mathbf{e}_i, \overrightarrow{\mathbf{h}}_{i-1}), \tag{8}$$

$$\overleftarrow{\mathbf{h}}_i = \text{GRU}(\mathbf{e}_i, \overleftarrow{\mathbf{h}}_{i+1}). \tag{9}$$

3) 2ND BiGRU LAYER

The 1st BiGRU model sequentially takes each word in the input sentence, extracts its information, and embeds it into a semantic vector. Owing to its ability to capture long-term memory, the BiGRU accumulates increasingly richer information as it goes through the sentence. The entire representation can be obtained as the final hidden state of the last word or time step. We retain the final forward and backward hidden states of each segment separately from the 1st BiGRU, and then feed them into the 2nd BiGRU layer.

Let $\overrightarrow{\mathbf{h}}_{PRED}$ and $\overleftarrow{\mathbf{h}}_{PRED}$ be the two final hidden states of the forward and backward directions generated for the *PRED* segment, respectively, and similarly for the other segments. As shown in Fig. 1, we put these hidden states together in order of their occurrences in the input sentence to establish a direction-aware sequence:

$$(\overrightarrow{\mathbf{h}}_{PRED}, \overleftarrow{\mathbf{h}}_{PRED}, \overrightarrow{\mathbf{h}}_{MID}, \overleftarrow{\mathbf{h}}_{MID}, \overrightarrow{\mathbf{h}}_{POST}, \overleftarrow{\mathbf{h}}_{POST}). \tag{10}$$

The 2nd BiGRU takes the above sequence as the entire input, and can build up progressively higher-level representations of sequence data. Thus, it is more effective than the single-layer BiGRU encoder.

4) PIECEWISE ATTENTION LAYER

The attention mechanism was introduced by [15] in order to stress the target words step by step in machine translation. Recently, it was transferred to other tasks including distantly supervised RE. Lin *et al.* [4] proposed a sentence-level attention scheme to select informative sentences from each bag. Jat *et al.* [16] recently introduced a model with sentence-level attention integrated with word-level attention to further explore the importance of different words in each sentence.

The word-level attention mechanism is a straightforward method to extract specific words that are important to the meaning of a sentence. However, a drawback of this method as an approach for distantly supervised RE is that it is difficult to take the directionality of target relations into account. For example, we may know that two entities e_1 and e_2 should be related in a relation r (the relation is not symmetric in general), but we cannot really infer whether the tuple (e_1, r, e_2) or (e_2, r, e_1) is correct without focusing on the right context in a given sentence.

All of the segments in an input sentence might provide necessary information to RE. However, it is obvious that not all segments contribute equally to the sentence meaning for different relations. For example, considering three cases from the Riedel dataset with two entities are in boldface, and the important segments are underlined:

> <S1>. (*/people/person/nationality*) mr. burns said the indian foreign secretary, **shiv_shankar_menon**$_{<e_1>}$, had been invited to Washington for talks early next month, and mr. burns planned then to travel to **india**$_{<e_2>}$.
> <S2>. (*/location/location/contains*) kelly air force base closed in the 1990 's, but **san_antonio**$_{<e_1>}$ is still ringed by three air force installations as well as **brooke_army_medical_center**$_{<e_2>}$ and fort sam houston, the army 's largest base through world war ii .
> <S3>. (*/people/person/children*) one, senator **evan_bayh**$_{<e_2>}$, above, son of former senator **birch_bayh**$_{<e_1>}$ of indiana, is testing the waters for a possible presidential bid in 2008 .

In the sentence <S1>, the left segment is more important than others to reflect the relation type */people/person/nationality*. In the sentence <S2>, the middle and right segments might provide the necessary information to the relation type */location/location/contains*. The right context in <S2> also supplement more useful information for predicting target relations. In the last example, the middle segment is the most important part related to the relation type */people/person/children*. In addition, the direction of the relation between two entities *birch_bayh* and *evan_bayh* in the sentence <S3> should be taken into account properly.

In our model, we therefore integrate a direction-aware attention layer over the 2nd BiGRU network to tackle the above challenges. We propose an additional attention mechanism called *piecewise attention*, which builds itself upon segment-level representations to improve the performance of the distantly supervised RE task. Our piecewise attention not only captures crucial segments in each sentence but also reflects the direction of the target relation in each segment.

As shown in Fig. 1, we obtain hidden state representations of the sentence by feeding the sequence (10) into the 2nd BiGRU:

$$\{\mathbf{h}_1, \ldots, \mathbf{h}_6\} = \text{BiGRU}(\{\overrightarrow{\mathbf{h}}_{PRED}, \ldots, \overleftarrow{\mathbf{h}}_{POST}\}), \tag{11}$$

where

$$\mathbf{h}_j = [\overrightarrow{\mathbf{h}}_j \oplus \overleftarrow{\mathbf{h}}_j]; \quad j = 1, 2, \ldots, 6, \tag{12}$$

and the number of hidden states produced by the 2nd BiGRU is 6, which is equal to the number of components of the input to the BiGRU in Eq. (11). Here, we use the element-wise sum (the symbol $\oplus$ in Eq. (12)) to combine the forward and backward pass outputs.

Next, we apply the attention mechanism at the segment level to assign a weight α_i to each hidden vector $\mathbf{h}_i$ generated by the BiGRU network, and pay more attention to the informative segment. The piecewise attention α_i is given by

$$\mathbf{h}'_i = \tanh(\mathbf{h}_i), \tag{13}$$

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{h}'_i)}{\sum_k \exp(\mathbf{w}^\top \mathbf{h}'_k)}, \tag{14}$$

where $\mathbf{w}$ is a parameter vector to be trained, and $\mathbf{w}^\top$ is a transpose.

Finally, we aggregate the representation of these direction-aware segments to construct the sentence representation. The final sentence vector $\mathbf{x}$ is computed as a weighted sum of hidden states $\{\mathbf{h}_1, \ldots, \mathbf{h}_6\}$ as follows:

$$\mathbf{x} = \sum_{i=1}^{6} \alpha_i \mathbf{h}_i. \tag{15}$$

B. BAG ENCODER

Following previous work [4], we use selective attention to deemphasize noisy sentences in the given bag. By using the sentence-level attention over sentences, a representation for the entire bag is learned. The details are described below.

1) SENTENCE-LEVEL ATTENTION LAYER

In our model, the piecewise attention and the sentence-level attention are complemented to deemphasize the noisy samples. The sentence-level attention layer assigns higher weights to valid sentences and lower weights to invalid ones in a particular bag $B = \{x_1, x_2, \ldots, x_{n_b}\}$. The sentence-level attention β_i for the sentence vector $\mathbf{x}_i$ can be computed by

$$s_i = \mathbf{x}_i^\top \mathbf{A}\mathbf{r}, \tag{16}$$

$$\beta_i = \frac{\exp(s_i)}{\sum_k \exp(s_k)}, \tag{17}$$

where $\mathbf{A}$ denotes a diagonal weight matrix, $\mathbf{r}$ is a parameter vector related to relation r, and the query-based function s_i scores how well the input sentence x_i and the relation r match.

The final representation $\mathbf{b}$ for a given bag is computed as a weighted sum of its sentence vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n_b}\}$:

$$\mathbf{b} = \sum_{i=1}^{n_b} \beta_i \mathbf{x}_i. \tag{18}$$

where n_b is the number of sentences in the n-th bag.

2) CLASSIFICATION AND TRAINING

The bag vector $\mathbf{B}$ extracted from the segments and sentences of a bag B is a high-level representation of that bag and can be used as features for relation classification. Then, $\mathbf{B}$ is passed to a softmax layer to predict the probability distribution corresponding to the relation labels. The conditional probability of the i-th relation is

$$p(r_i|B;\theta) = \frac{\exp(\mathbf{o}_i)}{\sum_k \exp(\mathbf{o}_k)}, \tag{19}$$

where θ denotes all parameters of our model, and $\mathbf{o} = \mathbf{Mb} + \mathbf{d}$ comprises the scores of all possible relations ($o \in \mathbb{R}^{|N|}$, where $\mathbf{M}$ is the representation matrix, d is a bias vector, and N denotes the number of relations).

We define the objective function using cross-entropy at the bag level [4]:

$$J(\theta) = \sum_{i=1}^{n_b} \log p(r_i|B;\theta) \tag{20}$$

In addition, we adopt the dropout strategy [17] and use stochastic gradient descent (SGD) to optimize our models.

C. BAG-LEVEL CONTEXTUAL INFERENCE METHOD

The advantage of distantly supervised RE lies in aggregating features from multiple sentences for the same entity pair. However, in many cases, there are insufficient number of sentences for a particular entity pair because of the limited coverage of the text corpus (e.g., when aligning the KB with that corpus, we can not acquire many sentences for rare entity names, such as person/location names). For example, in the testing set developed by Riedel *et al.* [5], which is the most widely used dataset for the distantly supervised RE task, there are 74, 857 entity pairs that correspond to only one sentence around 3/4 overall entity pairs [4]. Therefore, it is desirable to infer more sentences for that entity pair. In addition, few sentence may not cover the diversity of the context for predicting the bag's label. More contexts may increase the confidence score of the prediction.

Using a small number of sentences in each test bag may affect the accuracy of the prediction in the testing phase. We therefore propose using a *contextual inference* method that can infer the most likely positive examples of an entity pair in test bags with limited contextual information without using any external corpora or KBs. The target bags are those containing only one or very few sentences in the testing phase.

For example, consider the following two sentences:

$\mathbf{s}_1$: "…***Tokyo*** *is located in* ***Japan*** …" <in training data>

$\mathbf{s}_2$: "…***Paris*** *is the capital of* ***France*** …" <in testing data>

In the above example, the sentence s_1 belongs to the bag *(Tokyo, Japan)* in the training set, and the s_2 is in the bag *(Paris, France)* in the testing set. Our assumption is that *if these two bags have a high similarity, their two entity pairs can be replaced by each other to form new sentences that may cover more contexts for the target relations*. One of the new examples can be produced by this assumption is "*Paris is located in France*."

We use the cosine function to measure the similarity of two bags. Each bag is represented by the embedding difference between its entity vectors [18], e.g., the bag *(Tokyo, Japan)* corresponds to vec("*Japan*") - vec("*Tokyo*"). The similarity between two bags (e_1, e_2) and (x_1, x_2) is defined as

$$\begin{aligned}&\text{Sim}((e_1, e_2), (x_1, x_2))\\ &\quad = \cos([vec(e_2) - vec(e_1)], [vec(x_2) - vec(x_1)])\end{aligned} \tag{21}$$

Our bag-level contextual inference method is described in Algorithm 1. We leverage the given training data to generate *artificial* sentences, and hence increase the number of positive examples for each bag in the testing phase. It is expected that the newly generated sentences will share a similar semantic meaning with the target bag and provide supporting contexts for prediction. Our inference method aims to find high-quality sentences and avoid noise added to the target bags. It can be integrated in our proposed BiGRU-based model. To the best of our knowledge, our contextual inference method is the first approach that can infer more examples for the target relations leveraging the similarity of two bags, without using any external resources in the distantly supervised RE task.

Algorithm 1 Bag-Level Contextual Inference

1 For each target bag (e_1, e_2) in a testing set (e.g., bags with only one sentence):
2 Find top-k similar bags to (e_1, e_2) from training set according to Eq. (21). Each sentence s in these similar bags has the form (x_1, c, x_2), where x_1, x_2 are two entities, and c is the context in s.
3 A new artificial sentence s' is generated with the form (e_1, c, e_2) by joining (e_1, e_2) and c.
4 Retain a maximum number of sentences s' (e.g., 5) added to the bag (e_1, e_2).
5 Include the newly generated sentences s' in the bag (e_1, e_2) to support the prediction.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

1) RIEDEL DATASET

We use the Riedel dataset introduced in [5], which is the most commonly used dataset for the distantly supervised RE task. It was generated automatically by aligning New York Times (NYT) articles with the Freebase KB. Articles from 2005–2006 are used as training, and articles from 2007 are

used as testing. The training data contain 522, 611 sentences, 281, 270 entity pairs, and 18, 252 relational facts. The testing data contain 172, 448 sentences, 96, 678 entity pairs, and 1, 950 relational facts. In total, there are 53 relation labels including the *NA* relation in this dataset. However, this automatically generated dataset could be incorrect owing to the limitation of the DS assumption.

2) OUR ANNOTATED DATASET

A training dataset for DS is created with the following simple rule: If a sentence mentions two entities e_1 and e_2 and they are known to have a relation r (according to a KB such as Freebase), the sentence must be put in a bag for the relation r between entities e_1 and e_2. Nevertheless, this rule may produce many *false positive* sentences in a bag, as e_1 and e_2 may have occurred in the same sentence merely by chance. Consequently, the existence of false positive sentences in a bag can hurt the performance of RE models.

We therefore provide an annotated dataset to guarantee the quality of the data and report on the real performance of various RE systems. The Riedel testing set comprises 172, 448 sentences, and 6, 444 of them are labeled as *positive* examples by the DS assumption. As some of them appear several times, we use 5, 863 unique positive examples for our annotation. To the best of our knowledge, our current work is the first that provides such a high number of annotated sentences for the distantly supervised RE task.

TABLE 1. Details of the second stage of our annotation process.

		Annotator 2	
		False positive	True positive
Annotator 1	False positive	1,529	46
	True positive	42	4,246

In the first stage, we request two annotators to check independently if 5, 863 sentences express the target relations. Second, the two annotators discuss the disagreed labels in order to reach a consensus. The details of the second stage of our annotation process are listed in Table 1. There are 1, 529 sentences where both annotators are marked as ''false positive'' and 4, 246 sentences marked as ''No'' (i.e., true positive). The Cohen's kappa coefficient on our annotation is 0.96, which indicates a strong agreement between annotators. For 88 sentences (1.5%) for which the two annotators cannot reach an agreement, another participant is involved in the decision-making process. Finally, 1, 575 of 5, 863 sentences (26.86%) are judged as false positive by three annotators.

3) EXPERIMENTAL SETTINGS

We follow the parameter settings that are similar to those used in previous baselines [3], [4] in order to evaluate the effectiveness of our proposed methods. We use the word embeddings trained on the NYT corpus. The entities consisting of multiple tokens are considered as a single token. The dimensions for the word embedding (WE) and position embedding (PE) are set to 50 and 5, respectively. We use the maximum relative distance $L = 100$ in the position embedding, which is randomly initialized. The BiGRU hidden unit size is set to 230. We use a dropout with probability $p = 0.5$ and learning rate $\lambda = 0.01$ for the SGD.

For the bag-level inference method, we also use the skip-gram word2vec model to measure the similarities between different bags. The target bags are those with only 1 sentence. The maximum number of sentences added to each bag is 5. We tune the top-k similar bags to the target bag when our inference method is combined with others.

For evaluation, we report on the performance of models by using a precision-recall curve and top-N precision (P@N) metrics, which were commonly used in previous works.

4) COMPARED MODELS

To evaluate our proposed models, we compare them against the previous baselines for the distantly supervised RE task. All of the models are described as follows:

- **Mintz**: A multiclass logistic regression model [2].
- **MultiR**: A probabilistic graphical model for multi-instance learning [6].
- **MIMLRE**: A graphical model that jointly models multiple instances and multiple labels [7].
- **CNN+ONE**: A CNN-based RE model [10] with multi-instance learning [3].
- **CNN+ATT**: A CNN-based RE model [10] with sentence-level attention [4].
- **PCNN+ONE**: A CNN-based RE model [3] that uses piecewise max-pooling to generate the sentence representation.
- **PCNN+ATT**: A piecewise max-pooling over a CNN-based model to obtain the sentence representation, followed by sentence-level attention [4]. Currently, *PCNN+ATT* is one of the state-of-the-art neural-network-based RE models for this task.
- **PCNN+ATT+Inference**: The model *PCNN+ATT* combined with our *bag-level contextual inference method*.
- **BGWA**: A recent single-layer BiGRU-based RE model with word-level and sentence-level attention [16].
- **2BiGRU+PATT**: Our proposed model, which uses two BiGRU layers and *piecewise* attention.
- **2BiGRU+PATT+Inference**: Our proposed model *2BiGRU+PATT* combined with the *bag-level contextual inference method*.

We refer to three feature-based systems (Mintz, MultiR, and MIMLRE) as the traditional models, and

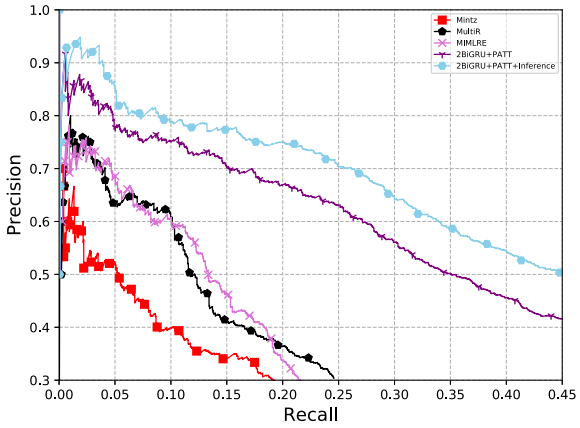


FIGURE 2. Performance comparison of the proposed model and traditional methods.

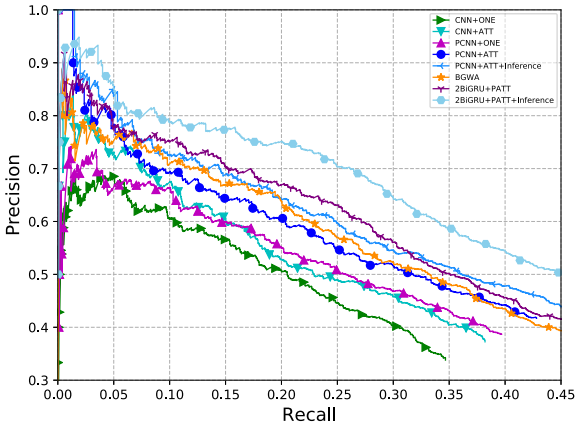


FIGURE 3. Performance comparison of the proposed model and state-of-the-art methods.

neural-network-based systems (CNN+ONE, CNN+ATT, PCNN+ONE, PCNN+ATT, PCNN+ATT+Inference, and BGWA) as the state-of-the-art models for comparison. An analysis of the results is provided in the next section.

B. EXPERIMENTAL RESULTS AND ANALYSIS

1) COMPARISON WITH TRADITIONAL METHODS

We evaluate our proposed models (2BiGRU+PATT and 2BiGRU+PATT+Inference) and compare them with three conventional feature-based methods (Mintz, MultiR, and MIMLRE) on the Riedel dataset. The precision-recall curve of each system is shown in Fig. 2. It is obvious that our proposed models significantly outperform all feature-based methods over the entire range of recall. When the recall is around 0.1, the performances of Mintz, MultiR, and MIMLRE drop quickly, while our models maintain high precision. All of the feature-based methods used human-designed features, which are time consuming and labor intensive. By contrast, our models can automatically learn the intrinsic features without human intervention from a large number of training examples.

2) EFFECTS OF OUR PROPOSED METHODS AND COMPARISON WITH STATE-OF-THE-ART MODELS

We compare our proposed models with two types of recent CNN-based models: the CNN model in [10] and the PCNN model in [3]) with at-least-one multi-instance learning (+ONE) used in [3] and the sentence-level attention (+ATT) used in [4]. *PCNN+ATT* is one of the state-of-the-art neural-network-based RE models reported in the Riedel dataset. The precision-recall curves of these models are presented in Fig. 3. The results show that our 2BiGRU+PATT model performs better than all CNN-based models to a significant extent, especially when compared to the state-of-the-art PCNN+ATT system. Our 2BiGRU+PATT+Inference model achieves the best performance among all of the methods. This demonstrates the effectiveness of our proposed models for the distantly supervised RE task.

We also compare our models with BGWA, which is a recent single-layer BiGRU-based RE model with word-level and the sentence-level attention [16]. From Fig. 3, we observe that the BGWA model achieves performance that is comparable to that of the PCNN+ATT model. BGWA is considered a baseline for evaluating the effectiveness of our piecewise attention as BGWA and 2BiGRU+PATT employ similar hierarchical attention networks (word-level or piecewise attention combined with sentence-level attention). The results indicate that the precision value of our 2BiGRU+PATT model is higher than that of the BGWA model when the recall value changes. This demonstrates the effect of using piecewise attention instead of word-level attention. Our new attention mechanism helps the RE models to focus on the right context in a given sentence and captures the directionality of nonsymmetric relations more efficiently.

Next, we compare the effects of integrating our bag-level contextual inference method into different systems. Our inference method boosts the performance of the PCNN+ATT system significantly and makes PCNN+ATT+Inference comparable to 2BiGRU+PATT. The inference method also enables the 2BiGRU+PATT+Inference model to achieve a large improvement compared to the 2BiGRU+PATT model. All of these examples show the superiority of our method against the state-of-the-art methods.

3) PERFORMANCE OF OUR ANNOTATED DATASET

In the Riedel testing set, there are 172, 448 sentences, and 6, 444 of them are labeled as *positive* examples by the DS assumption. We replace the labels of 6, 444 sentences in the Riedel testing set, which are checked by annotators, and refer to this as our annotated dataset. It means that we only changed the labels of false positive sentences to "*NA*" (i.e., true negative), and the total number of sentences is unchanged.

Fig. 4 shows the performance of our annotated dataset for three models: PCNN+ATT, 2BiGRU+PATT, and 2BiGRU+PATT+Inference. The "*" symbols denote the evaluations of our annotated dataset. It is observed that there are slight changes when the results are reported on the original and our annotated dataset. However, all of the systems are robust, and our 2BiGRU+PATT model performs even better on the annotated dataset. Our bag-level contextual

TABLE 2. P@N for RE in bags with different numbers of sentences; * symbols denote evaluations of our annotated dataset; One, two, and all denote number of sentences randomly selected from a bag; best scores are in boldface.

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN+ATT	73.3	**69.2**	60.8	67.8	77.2	**71.6**	**66.1**	71.6	76.2	**73.1**	67.4	72.2
2BiGRU+PATT	**76.2**	66.7	**62.1**	**68.3**	**80.2**	69.2	65.8	**71.7**	**83.2**	**73.1**	**69.8**	**75.4**
PCNN+ATT*	70.0	63.0	58.7	63.9	76.0	**71.5**	65.0	70.8	75.0	71.5	66.3	70.9
2BiGRU+PATT*	**74.3**	**64.2**	**59.5**	**66.0**	**80.2**	68.7	**65.1**	**71.3**	**83.2**	**72.6**	**69.1**	**75.0**

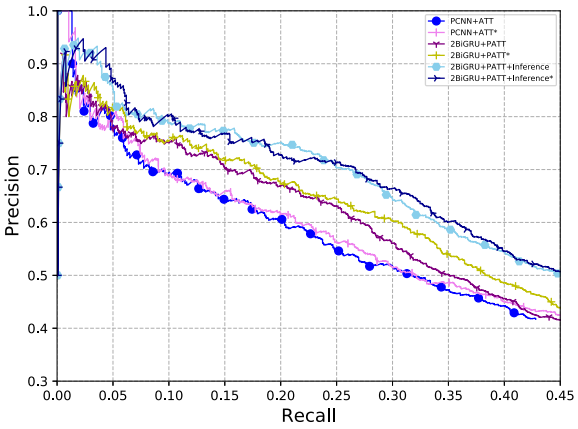


FIGURE 4. Performance of models on annotated dataset; * symbols denote evaluations of our annotated dataset.

method still shows its benefits and does not require any external resources of KBs. Furthermore, ours is the first work to report on the performance of various RE models on an annotated dataset with a high number of testing examples (5, 863) checked by humans.

4) EFFECT OF SENTENCE NUMBER

Following previous works, we also evaluate our methods with different numbers of sentences in the bags with more than one sentence. In this setting, one, two, or all sentences are (randomly) selected from each bag for comparison in the testing phase. We then report the P@100, P@200, P@300, and their mean for each model. The results are listed in Table 2. In all settings, our 2BiGRU+PATT model obtains higher average precision than the PCNN+ATT model, which demonstrates the efficacy of our method. These improvements are observed on both datasets to an extent of 3.2% (using all sentences in the Riedel dataset) and 4.1% (using all sentences in our annotated dataset). Using all of the sentences helps the models achieve the best results. However, adding sentences might result in more noise, which can affect the performance. This is illustrated in the ''One'' and ''Two'' settings. The 2BiGRU+PATT model using two sentences does not produce a higher improvement than when using only one sentence: 71.6 to 71.7% and 67.8 to 68.3% on the Riedel dataset, respectively; and 70.8 to 71.3% and 63.9 to 66.0% on our annotated dataset, respectively.

TABLE 3. P@N for RE in all bags; * symbols denote evaluations on our annotated dataset; best scores are in boldface.

Test Settings	All Bags			
P@N(%)	100	200	300	Mean
PCNN+ATT	81.0	71.0	69.3	73.8
PCNN+ATT+Inference	83.0	75.0	72.7	76.9
2BiGRU+PATT	82.2	75.6	73.8	77.2
2BiGRU+PATT+Inference	**87.1**	**81.1**	**78.1**	**82.1**
PCNN+ATT*	81.0	69.5	67.3	72.6
2BiGRU+PATT*	82.2	75.6	72.8	76.9
2BiGRU+PATT+Inference*	**86.1**	**79.6**	**76.7**	**80.8**

5) P@N IN ALL BAGS

The P@N results for all bags are presented in Table 3. We can see that our proposed methods show their advantages and achieve notable performance for all values of P@100, P@200, P@300, and Mean. For the Riedel dataset, our 2BiGRU+PATT model performs better than the PCNN+ATT model when the average precision increases from 73.8% to 77.2%, and performs in a similar manner for the models that use our inference method (76.9% to 82.1%). For our annotated dataset, the scores also improved remarkably: 72.6 to 76.9% when using our novel BiGRU-based model, and 72.6 to 80.8% when incorporating the additional inference method. All of the proposed methods still show their robustness on both datasets.

6) PARAMETER TUNING FOR OUR BAG-LEVEL CONTEXTUAL INFERENCE METHOD

For our bag-level contextual inference method, we tune the top-k similar bags (this is shown in Algorithm 1) to find the best performance of two models: PCNN+ATT+Inference and 2BiGRU+PATT+Inference. The average P@N ($N = 100, 200, 300$) results for all bags are used for comparison. Table 4 lists the numbers of similar bags and inferred sentences that were generated by our inference method. When the number of similar bags increases, the number of inferred sentences is incremented accordingly in most cases. The maximum number of sentences is 1, 807, which corresponds to 28.04% of the positive examples in the original Riedel testing dataset. When the number of similar pairs >= 15, the generated sentences are the same as for 14 since our method already generated all possible sentences for the bags with only one sentence.

TABLE 4. Parameter tuning for our bag-level context inference method; we only create data for bags with one sentence in testing set; maximum number of sentences added to each bag is five; when number of similar pairs >=15, generated sentences are same as for 14 since our method already generated all possible sentences for bags with one sentence; best score for each model is in boldface.

No. of Similar Bags (top-k)	1	2	3	4	5	6	7
No. of Inferred Sentences	756 (+11.73%)	1,302 (+20.20%)	1,526 (+23.68%)	1,656 (+25.70%)	1,713 (+26.58%)	1,748 (+27.13%)	1,769 (+27.45%)
PCNN+ATT+Inference - P@N (Mean)	76.7 ↑	**76.9** ↑	76.9 →	76.8 ↓	76.8 →	76.6 ↓	76.7 ↑
2BiGRU+PATT+Inference - P@N (Mean)	79.4 ↑	80.8 ↑	81.2 ↑	81.7 ↑	81.9 ↑	81.8 ↓	81.7 ↓
No. of Similar Bags (top-k)	8	9	10	11	12	13	14
No. of Inferred Sentences	1,786 (+27.72%)	1801 (+27.95%)	1801 (+27.95%)	1,802 (+27.96%)	1,803 (+27.98%)	1,803 (+27.98%)	1,807 (+28.04%)
PCNN+ATT+Inference - P@N (Mean)	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →
2BiGRU+PATT+Inference - P@N (Mean)	82.0 ↑	**82.1** ↑	82.1 →	82.1 →	82.1 →	82.1 →	82.1 →

TABLE 5. Some example results of our proposed models; correct predictions are in boldface.

1st Entity	2nd Entity	A Sentence in Bag	Gold Labels	Top-3 Predictions of 2BiGRU+PATT (probability)	Top-3 Predictions of 2BiGRU+PATT+Inference (probability)	Unknown Words (unknown entity is in *italics*)
jean-baptiste_colbert	france	a 17th-century eyewitness account of the coronation of a shah , written by jean chardin , a french jeweler , is inscribed to jean-baptiste_colbert , then the finance minister of france .	/people/person/nationality	**/people/person/nationality (0.965)** NA (0.023) /people/person/place_of_birth (0.005)	**/people/person/nationality (0.978)** NA (0.011) /people/person/place_of_birth (0.007)	17th-century, *jean-baptiste_colbert*
seyyed_hossein_nasr	george_washington_university	i am not apologetic about why the koran says this , said seyyed_hossein_nasr , an islamic scholar who teaches at george_washington_university .	/business/person/company	**/business/person/company (0.977)** NA (0.022) /people/person/religion (0.0004)	**/business/person/company (0.995)** NA (0.005) /people/person/religion (0.0001)	*seyyed_hossein_nasr*
nuevo_leon	mexico	on may 8 , representative marcy kaptur , an ohio democrat , and a dozen other legislators wrote to president felipe calderon of mexico and the governor of the state of nuevo_leon , of which monterrey is the capital , urging them to thoroughly investigate the killing and provide protection for the rest of the mexico staff of the farm workers ' union .	/location/administrative_division/country	**/location/administrative_division/country (0.548)** NA (0.348) /people/person/nationality (0.083)	**/location/administrative_division/country (0.545)** NA (0.338) /people/person/nationality (0.095)	kaptur, *nuevo_leon*
dylan_thomas	aeronwy_thomas	next year he is planning to publish the poetry of aeronwy_thomas , dylan_thomas 's daughter , and to bring her to the united states for a book tour along with the welsh poet and publisher peter thabit jones .	/people/person/children	NA (0.597) **/people/person/children (0.364)** /business/person/company (0.011)	NA (0.837) **/people/person/children (0.146)** /people/person/nationality (0.006)	*dylan_thomas*, thabit
canada	saskatchewan	if they have a residence in canada , they can buy farmland in saskatchewan through the agriculture development corporation , a private company , for a minimum buy-in of $ 20,000 .	/location/location/contains & /location/country/administrative_divisions	**/location/location/contains (0.790)** **/location/country/administrative_divisions (0.194)** NA (0.016)	**/location/location/contains (0.658)** **/location/country/administrative_divisions (0.336)** NA (0.005)	buy-in

The best average P@N score for each model is reported. The PCNN+ATT+Inference model reaches its best performance with top-k = 2, whereas our 2BiGRU+PATT+Inference model achieves the best result with top-k = 9. Compared to the original systems (which are listed in Table 3), the gap between 2BiGRU+PATT+Inference and 2BiGRU+PATT is higher than that of PCNN+ATT+Inference and PCNN+ATT: 82.1 to 77.2% compared with 76.9 to 73.8%, respectively. This is useful in practice because both models are beneficial when using the inference method to support the prediction. Our model shows its advantages and leverages the artificial data more efficiently.

7) CASE STUDY

Table 5 shows five randomly selected example results of our proposed models from the Riedel testing data. For each case, we show the gold labels and the top-3 predictions of our *2BiGRU+PATT* and *2BiGRU+PATT+Inference* models, respectively. The values appeared in parentheses represent their corresponding probabilities. The correct predictions are in boldface.

We can see that our two proposed models produce reasonable predictions in the analysis for our relation extraction task. For four of five cases (except the 4-th case), our proposed models give high probabilities to the correct predictions. The contextual inference method can enhance the performance of our *2BiGRU+PATT* model with the help of supporting contexts and is useful in our task. Our *2BiGRU+PATT+Inference* model assigns comparable or higher scores to the correct predictions than the *2BiGRU+PATT* model.

In the last column of Table 5, we show the unknown words, which can not be found in our embedding matrix, in the corresponding sentence. The unknown entities are indicated in italics. An *unknown entity* affected significantly to the label of its bag for the short context, especially in the 4-th case. Since the is no meaningful text span between two

entities *dylan_thomas* and *aeronwy_thomas*, and the 1st entity's vector is missing from the embedding matrix, our models result in the second top-scoring predictions (i.e., */people/person/children*).

We checked the ratio of matched entities between the Riedel dataset and our embedding matrix. We use the word embeddings trained on the NYT corpus and keep the words which appear more than 100 times in the corpus as vocabulary. These word embeddings are similar to previous baselines [3], [4]. There are 69,040 unique entities appeared in the Riedel dataset. However, we found that only 22,515 of 69,040 entities (32.61%) matched in our embedding matrix. It suggests that a larger text corpus should be used to cover the high number of entities appeared in the Riedel dataset and improve the performance of our proposed models. In addition, the vector embeddings of Wikipedia concepts and entities, such as a person's name, an organization or a place can be trained using the character embedding, which handles infrequent words better than the word embedding as the latter suffers from lack of enough training opportunity for out-of-vocabulary words.

Figure 5 shows similar entity pairs involved in our contextual inference method from both training and testing portions in the Riedel dataset. Recall that for each target bag (e_1, e_2) in a testing set, our contextual inference method selects top-k similar bags to (e_1, e_2) from the training set according to Eq. (21). We selected 1,000 pairs between (e_1, e_2) and (x_1, x_2) that have highest similarity scores, and visualize these pairs using force-directed graph layout algorithms. Each entity pair (or a bag) is represented by a node, and similar entity pairs are linked by edges in the graph, which provides an overview of relationships among related bags.

In order to evaluate the quality of similar entity pairs chosen by our contextual inference method using the vector difference between entities' vectors, we randomly select 100 pairs between (e_1, e_2) and (x_1, x_2) (out of 1,000 pairs above), and check whether these two pairs indeed have a similar semantic relationship. For example, *(atlanta, high_museum_of_art);(chicago, art_institute_of_chicago)* is assigned as correct since these two pairs are similar according to the */location/location/contains* relationship. In total, 83 out of 100 cases (83.0%) are judged as correct by two annotators. It demonstrated that using the vector difference between e_1 and e_2, and x_1 and x_2 in Eq. (21) is effective for calculating the similarity between bags. Without any external corpora or KBs, our inference method showed its advantages and leveraged the training data efficiently.

For better understanding the reason of the incorrect inference, we also analyzed each entity name in 17 incorrect cases (out of 100 cases above). For example, *(kentucky, center_college);(mitch_mustain, arkansas)* is an incorrect example, where *mitch_mustain* is a person name, and others are locations or places. We found that 13 out of 17 incorrect cases (76.5%) contain at least one person name, while only 22 out of 83 correct cases (26.5%) have such entity type. It indicates that learning meaningful vector representations for person names is more difficult than for others. In the future work, we think that much efforts should be done to obtain better embeddings of rare entity names, such as the person names in the Riedel dataset.

Due to the diversity of relation types and limitations of model capabilities, we think that a small number of incorrect predictions are inevitable. In general, our proposed methods are very effective for improving the performance of the distantly supervised relation extraction systems.

V. RELATED WORK

The distantly supervised RE task aims at identifying the semantic relation of a sentence set expressed toward an entity pair or a bag level [2]. Ridel *et al.* [5], Hoffmann *et al.* [6], and Surdeanu *et al.* [7] introduced a series of models casting DS as a multiple-instance learning problem [8] to relax its original strong assumption.

Recently, neural networks have been widely explored in distantly supervised RE and achieved state-of-the-art results [3], [4], [10]. Most existing systems model the noisy DS process in the hidden layers by learning an informative sentence representation or features, and then selecting one or more valid relation mentions for RE. Zeng *et al.* [3] divided the original input sentence into three segments by the positions of two entities, and used piecewise max-pooling to automatically learn relevant features using a piecewise CNN (*PCNN*) model. Lin *et al.* [4] and Ji *et al.* [19] addressed the shortcoming of the PCNN model, which uses only the most relevant sentence from each bag. They proposed to use sentence-level attention to dynamically calculate the weights of multiple sentences, and then leverage large amounts of useful information from all sentences in each bag. Currently, *PCNN+ATT* [4] is one of the state-of-the-art neural-network-based RE models.

Zhou *et al.* [20] presented word-level attention integrated in a BiLSTM-based model and achieved significant improvements on SemEval2010 [21], which is a supervised dataset and cannot be used for the distantly supervised RE task. Yang *et al.* [22] and Jat *et al.* [16] combined the word-level and sentence-level attention mechanisms in their single-layer BiGRU-based models and showed that these performed better than the CNN/PCNN models.

We believe that using only sentence-level or the word-level attention might not be the optimal solution because the crucial information should be distributed to different segments in the input sentence. Therefore, in this work, we develop two-layer BiGRU-based models with a combination of piecewise and sentence-level attention in order to capture the significance of each piece of text as well as the directionality of nonsymmetric relations.

We also make another contribution by proposing a novel contextual inference method that can support the bags with very few examples. In addition, previous works usually evaluated RE systems in a held-out evaluation, which suffers from noise, e.g., in the Riedel dataset. Only a few works conducted manual evaluations with a small number of annotated

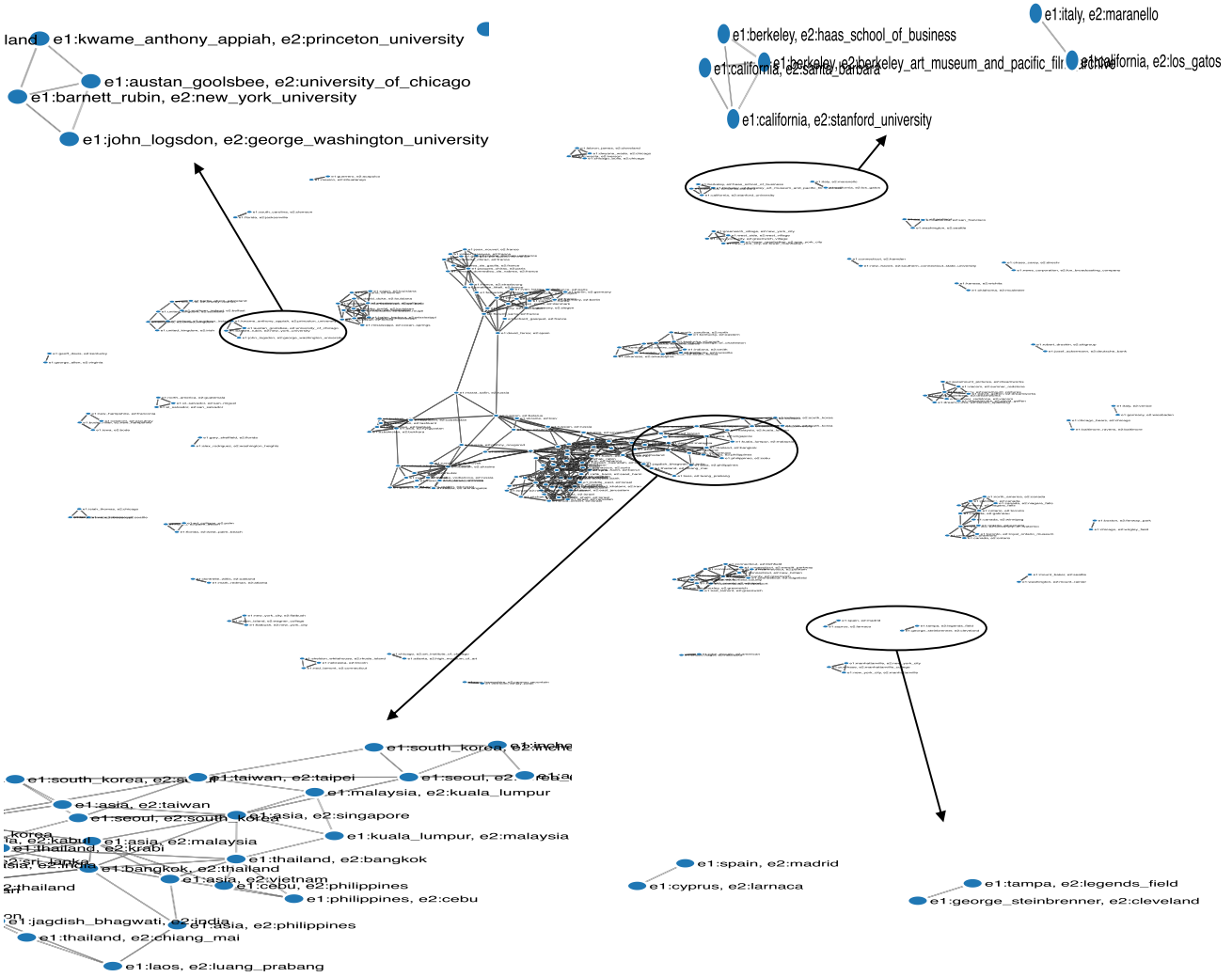


FIGURE 5. **Some similar entity pairs involved in our contextual inference method; each node represents an entity pair, and similar entity pairs are linked by edges.**

sentences (e.g., 500 in [19]). By providing an annotated dataset of non-false positive examples, the real performance of various RE systems can then be measured accurately.

VI. CONCLUSION AND FUTURE WORK

In this article, we proposed novel neural RE systems with two BiGRU layers and two attention modules: the piecewise and sentence-level attentions. We also presented a contextual inference method that can infer the most likely positive examples of an entity pair in bags with very limited contextual information without using any external KBs or corpora. The experimental results showed that our proposed models offer significant improvements over state-of-the-art methods on our newly created dataset and the Riedel dataset. Our dataset will be made publicly available for other researchers to use as a benchmark.

In the future, we plan to develop more sophisticated methods for measuring the similarity between entity-pair bags, such as using the shortest dependency path between the two entities instead of the full sentence to infer similar examples from external text corpora, and apply our methods to other domains such as biomedical or scientific articles in order to further benefit this task.

REFERENCES

[1] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," in *Proc. ISMB*, 1999, pp. 77–86.

[2] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. 47th Annu. Meeting ACL 4th IJCNLP AFNLP*. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2009, pp. 1003–1011. [Online]. Available: http://www.aclweb.org/anthology/P09-1113

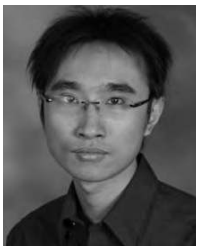
[3] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2015, pp. 1753–1762. [Online]. Available: http://www.aclweb.org/anthology/D15-1203

[4] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2016, pp. 2124–2133. [Online]. Available: http://www.aclweb.org/anthology/P16-1200

[5] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*. Berlin, Germany: Springer, 2010, pp. 148–163.

[6] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2011, pp. 541–550.
[7] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.* Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2012, pp. 455–465.
[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.
[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[10] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. COLING, 25th Int. Conf. Comput. Linguistics, Tech. Papers*, 2014, pp. 2335–2344.
[11] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[13] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Jun. 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078
[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 2014, *arXiv:1412.3555*. [Online]. Available: https://arxiv.org/abs/1412.3555
[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[16] S. Jat, S. Khandelwal, and P. Talukdar, "Improving distantly supervised relation extraction using word and entity based attention," Apr. 2018, *arXiv:1804.06987*. [Online]. Available: https://arxiv.org/abs/1804.06987
[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[18] V.-T. Phi and Y. Matsumoto, "Integrating word embedding offsets into the espresso system for part-whole relation extraction," in *Proc. 30th Pacific Asia Conf. Lang., Inf. Comput., Oral Papers*, 2016, pp. 173–181.
[19] G. Ji, K. Liu, S. He, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," in *Proc. AAAI*, 2017, pp. 3060–3066.
[20] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 207–212.
[21] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proc. Workshop Semantic Eval., Recent Achievements Future Directions*. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2009, pp. 94–99.
[22] L. Yang, T. L. J. Ng, C. Mooney, and R. Dong, "Multi-level attention-based neural networks for distant supervised relation extraction," in *Proc. 25th Irish Conf. Artif. Intell. Cogn. Sci.*, Dublin, Ireland: Insight Centre, Dec. 2017, pp. 1–12.

VAN-THUY PHI received the B.E. degree from the University of Engineering and Technology, Vietnam National University (VNU), in 2013, and the M.Sc. degree from the Nara Institute of Science and Technology (NAIST), in 2016, where he is currently pursuing the Ph.D. degree. His research interests include natural language processing and machine learning, in particular, information extraction.



JOAN SANTOSO received the bachelor's degree in computer science and the master's degree in information technology from Sekolah Tinggi Teknik Surabaya (STTS), in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember (ITS). He is also a Lecturer with the Department of Informatics, Institut Sains dan Teknologi Terpadu Surabaya (ISTTS). His research interest includes computational linguistics, information extraction, machine learning, and big data processing.

VAN-HIEN TRAN received the bachelor's degree in an honors program with the Faculty of Information Technology, University of Engineering and Technology, Vietnam National University (UET-VNU), in 2014. He is currently pursuing the master's degree with the Computational Linguistics Laboratory, Graduate of Information Science, Nara Institute of Science and Technology, Japan. From 2015 to 2016, he was a Teaching Assistant with the Data Science and Knowledge Technology Laboratory, UET-VNU. His research interests include natural language processing (especially on information extraction), text mining, and social network analysis.

HIROYUKI SHINDO received the B.E. and M.E. degrees from Waseda University, Japan, in 2007 and 2009, respectively, and the Ph.D. degree in engineering from the Nara Institute of Science and Technology (NAIST), Ikoma, Japan, in 2013. From 2009 to 2014, he was a Researcher with NTT Communication Science Laboratories. He is currently an Assistant Professor with the Graduate School of Information Science, NAIST. His research interests include machine learning and computational linguistics.

MASASHI SHIMBO received the M.E. and Ph.D. degrees in engineering from Kyoto University, in 1994 and 2000, respectively. He is currently an Associate Professor with the Nara Institute of Science and Technology. His research interests include machine learning, data mining, and information extraction from text.

YUJI MATSUMOTO received the M.Sc. and Ph.D. degrees in information science from Kyoto University, in 1979 and 1989, respectively. He joined the Machine Inference Section of the Electrotechnical Laboratory, in 1979. He served as an Academic Visitor with the Imperial College of Science and Technology, the Deputy Chief of the First Laboratory, ICOT, and an Associate Professor with Kyoto University. He is currently a Professor of information science with the Nara Institute of Science and Technology. His main research interests include natural language understanding and machine learning.