Employing Gated Attention and Multi-similarities to Resolve Document-level Chinese Event Coreference

Haoyi Cheng, Peifeng Li, Qiaoming Zhu
Natural Language Processing Lab, School of Computer Science and Technology
Soochow University
Suzhou, China
29029323@qq.com, {pfli, qmzhu}@suda.edu.cn

***Abstract*—Event coreference resolution is a challenging task. To address the issues of the influence on event-independent information in event mentions and the flexible and diverse sentence structure in Chinese language, this paper introduces a GANN (Gated Attention Neural Networks) model to document-level Chinese event coreference resolution. GANN introduces a gated attention mechanism to select event-related information from event mentions and then filter noisy information. Moreover, GANN not only uses a single Cosine distance to calculate the linear distance between two event mentions, but also introduces multi-mechanisms, i.e., Bilinear distance and Single Layer Network, to further calculate the linear and nonlinear distances. The experimental results on the ACE 2005 Chinese corpus illustrate that our model GANN outperforms the state-of-the-art baselines.**

Keywords-gated attention; multi-similarities; chinese event coreference

I. INTRODUCTION

Event coreference resolution is an important task in NLP, which is the foundation of many NLP tasks, such as topic detection [1], information extraction [2], and reading comprehension [3]. In recent years, most studies focused on entity coreference resolution, and only a few concern event coreference resolution because it is a challenging task. The task of event coreference resolution is to determine which event mentions (a phrase or sentence within which an event is described) in texts refer to the same real-world event and then cluster them to a unique coreferential event chain. Take the following two event mentions as examples:

S1：NHK报道日本检方1号针对发生在东海村的油燃加工厂不慎外泄事故**起诉** 6 名工厂的员工。(1st, Japanese prosecutors **sued** six employees of the oil fuel processing plant for the leakage accidents in Donghai. - from NHK.)

S2：因为东海村油燃料加工厂去年发生事故而被**起诉**的 6 名员工，包括…。(Six employees were **sued** for the accident at the oil fuel processing plant in Donghai last year, including ….)。

The event mention in S1 whose event trigger (the main word that most clearly expresses the occurrence of an event) is "起诉" (sue) and the mention in S2 with the trigger "起诉" (sue) refer to the same real-word event, a *Justice* event, and they are coreferential event mentions. Event mentions that refer to the same event can appear both within a document and across multiple documents. Hence, event coreference resolution is usually divided into document-level and cross-document level. This paper focuses on document-level Chinese event coreference resolution, which is critical to further cross-document event coreference resolution.

Most previous studies on Chinese document-level event coreference resolution were based on feature engineering, which used lots of manual features and could not capture the semantics hiding in event mentions. Currently, neural network models were introduced to English event coreference resolution and achieved success. Krause [4] and Fang [5] employed the CNN model and the multiple decomposable attention networks to resolve document-level event coreference. However, there are two issues in above neural network models. The first is that they cannot eliminate the influence of noises derived from those event-independent information (e.g., "NHK" in S1) event mentions. The second is that they are not suitable for Chinese event coreference resolution for the language-independent. Compared with English, the event mentions in Chinese are more complex and have the Chinese characteristics, such as discourse-driven, pro-drop, zero entity coreference, etc.

To address the above two issues, this paper introduces a novel GANN (Gated Attention Neural Network) model to the task of document-level Chinese event coreference resolution. Firstly, GANN introduces a gated attention mechanism to select the event-related information from event mentions and then filter noisy information. Secondly, GANN not only uses a single Cosine distance to calculate the linear distance between two event mentions, but also introduces multi-mechanisms, i.e., Bilinear distance and Single Layer Network, to further calculate the linear and nonlinear distance. Hence, linear distance and nonlinear distance can complement each other. The experimental results on the ACE 2005 Chinese corpus illustrate that our model GANN outperforms the state-of-the-art baselines.

II. RELATED WORK

Event coreference is much less studied in comparison to the large number of studies on entity coreference [6]. Early work on document-level event coreference resolution used traditional methods, such as probability-based models [7] and graph-based models [8], most of them derived from the entity coreference literature and focused on English. Chen [9] proposed a pairwise event coreference model on various kinds of annotating event attributes. Especially, it took the four basic attributes of events (tense, polarity, molarity and genericity) as its features. Liu [10] also used a SVM-based pairwise model on annotating information and manual features (e.g., semantic information of trigger words, distance between event mentions). Since the above

978-1-7281-5014-7/19/$31.00

pairwise classifiers do not take into account the document-level global information, there are many conflicts when pairwise results are transformed into coreferential event clusters. Lu and Ng [11] proposed a joint inference model based on Markov logic networks to correct the mistakes from the pairwise event coreference resolver. Liu [12] presented a unified graph framework to conduct event coreference by using many features (e.g., frame features, argument features and event mention distances). As for the task of Chinses event coreference resolution, there is only one literature concerned it. Teng [13] trained a Maximum Entropy Model on a large number of features extracted by the external tools.

Recently, neural networks have been widely-used in various English NLP applications. Krause introduced the Convolutional Neural Network (CNN) to event coreference and they used many annotating entity and event information. Fang introduced a multiple decomposable attention network from different views, i.e., event mention, event arguments and trigger context. Moreover, it applied document-level global inference mechanism to further resolve the coreference chains. This is the first paper to apply a neural network to the task of document-level Chinses event coreference resolution.

III. GANN FOR EVENT COREFERENCE RESOLUTION

In this section, we first introduce the framework of our gated attention neural network called GANN and then describe its components, i.e., input layer, gated attention layer, similarity calculation layer and output layer.

A. Overview

Krause used a CNN to dig out the contextual information of words and it only considered local information between words regardless of the relationship between event mention pairs. Fang used an attention mechanism to extract relatively important features, but there are noises in those extracted features. To address the above issues and the characteristics of Chinese language, this paper introduces a novel neural network GANN to resolve document-level Chinese event coreference and its architecture is shown in Figure 1.

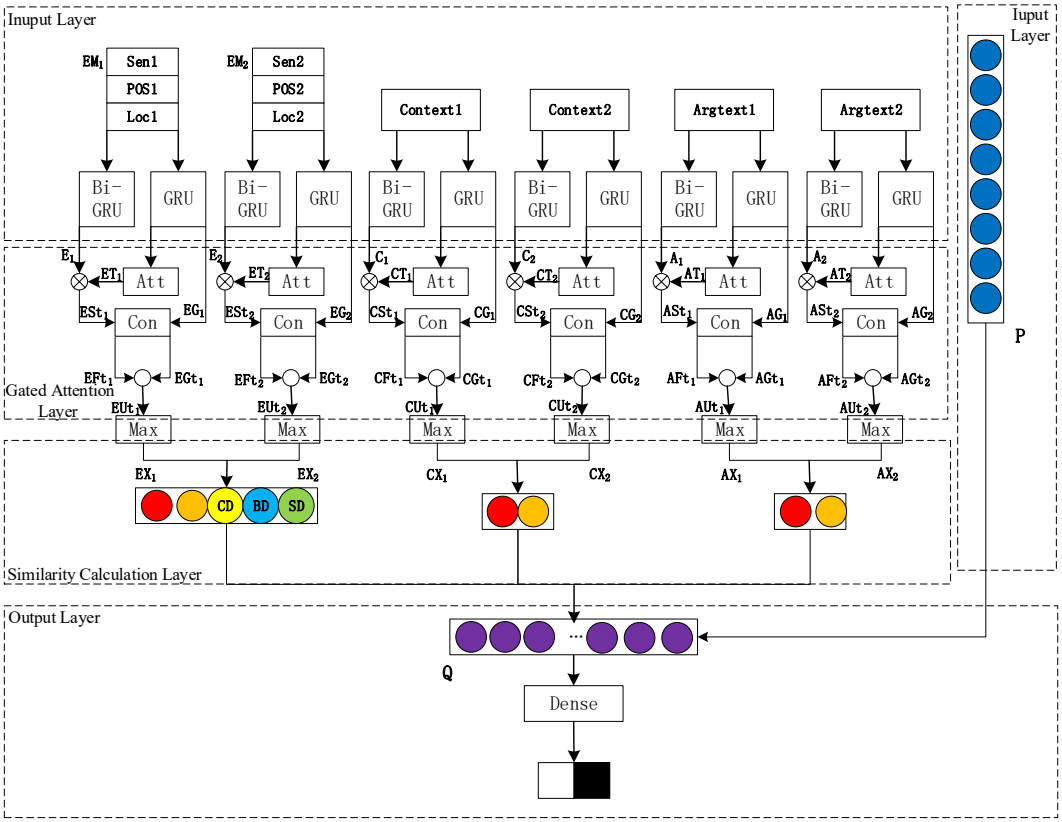

Figure1. The structure of GANN

GANN mainly introduces an attention mechanism to extract event-related features and then uses a gated mechanism to filters those event-independent information. Finally, it uses multiple similarity distances and pairwise features to resolve document-level Chinese event coreference. GANN is mainly divided into the following four parts: 1) input layer, 2) gated attention layer, 3) similarity calculation layer, and 4) output layer. In the input layer, various kinds of event mention representations (e.g., event mention, POS, location, context of the trigger, arguments and pairwise features) are encoded as vectors by using Bi-GRU and GRU. In the gated attention layer, a self-attention mechanism is introduced to extract event-related features from event mentions and then a gating mechanism is used to remove those event-independent information to reduce noisy information and simplify computational complexity. In the similarity calculation layer, the Cosine distance, Bilinear distance and Single Layer Network are used to obtain the linear distance and nonlinear distance between two event mentions. In the output layer, GANN determines whether a pair of event mentions is coreferential or not.

B. Input

Following Krause and Fang, the input of GANN is two event mentions e_1 and e_2 with annotated triggers, event types/subtypes, event arguments, and event attributes (e.g., modality), etc. We extract the follows features from these two event mentions as follows.

1) **Sentential features**: the words in event mention (SF1), their POS tagged by NLTK tools (SF2), and their positions (SF3);

2) **Context features**: the context around the trigger (the windows size is set to 3) (CF1) and the argument list of the event mention (CF2);

3) **Pairwise features**: the comparison results of the event trigger (PF1), type (PF2), subtype (PF3), modality (PF4), polarity (PF5), genericity (PF6) and tense (PF7), respectively. If the above attributes of an event mention pair have the same annotated label, its corresponding feature will be assigned 1; otherwise 0.

Pre-trained Wikipedia 300-dimensional word vector matrix **M** is used as the training matrix to encode SF1, SF2, CF1 and CF2 to four vectors **Sen**, **POS**, **Context** and **Argtext**, respectively. In Figure 1, **Sen1**, **POS1**, **Context1** and **Argtext1** are the vectors extracted from the event mention e_1, while **Sen2**, **POS2**, **Context2** and **Argtext2** are the vectors from the event mention e_2. Besides, the feature SF3 is encoded to 50-dimensions location vectors **Loc** by a random word embedding matrix.

For two event mentions e_1 and e_2, we merge their **Sen**, **POS** and **Loc** vectors to a vector **EM**, respectively, as follows.

$$\boldsymbol{EM_i} = Concat(\boldsymbol{Sen_i}, \boldsymbol{Pos_i}, \boldsymbol{Loc_i}) \quad (i = 1,2) \qquad (1)$$

We use Bi-GRU to encode $\mathbf{EM_i}$, $\mathbf{Context_i}$ and $\mathbf{Argtext_i}$ to get the new vectors $\mathbf{E_i}$, $\mathbf{C_i}$ and $\mathbf{A_i}$, and use GRU to encode $\mathbf{EM_i}$, $\mathbf{Context_i}$ and $\mathbf{Argtext_i}$ to get the new vectors $\mathbf{EG_i}$, $\mathbf{CG_i}$ and $\mathbf{AG_i}$. Since the pairwise features PF1-PF7 are binary numbers, we merge these pairwise features and the distance between two event mentions to a vector **P**.

C. Gated Attention

Attention mechanism [14] is a useful method in many NLP applications and can reweight each word in an event mention when we apply it to event coreference resolution. In this paper, we design a gated attention network to combine the attention mechanism and gating mechanism. This network uses the attention mechanism to reweight all words in the event mention, the context of the trigger and the argument list, and then uses the gating mechanism to reduce noisy information.

The vectors $\mathbf{EM_1}$, $\mathbf{EM_2}$, $\mathbf{Context_1}$, $\mathbf{Context_2}$, $\mathbf{Argtext_1}$ and $\mathbf{Argtext_2}$ are input into the self-attention layer as follows.

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left(\boldsymbol{QK^T}/\sqrt{\boldsymbol{d_k}}\right)\boldsymbol{V} \qquad (2)$$

$$\begin{aligned}\boldsymbol{ET_i} &= Attention(\boldsymbol{EG_i}, \boldsymbol{EG_i}, \boldsymbol{EG_i}) \\ &= Softmax\left(\boldsymbol{EG_i EG_i}^T/\sqrt{\boldsymbol{d_k}}\right)\boldsymbol{EG_i} \quad (i = 1,2)\end{aligned} \qquad (3)$$

$$\begin{aligned}\boldsymbol{CT_i} &= Attention(\boldsymbol{CG_i}, \boldsymbol{CG_i}, \boldsymbol{CG_i}) \\ &= Softmax\left(\boldsymbol{CG_i CG_i}^T/\sqrt{\boldsymbol{d_k}}\right)\boldsymbol{CG_i} \quad (i = 1,2)\end{aligned} \qquad (4)$$

$$\begin{aligned}\boldsymbol{AT_i} &= Attention(\boldsymbol{AG_i}, \boldsymbol{AG_i}, \boldsymbol{AG_i}) \\ &= Softmax\left(\boldsymbol{AG_i AG_i}^T/\sqrt{\boldsymbol{d_k}}\right)\boldsymbol{AG_i} \quad (i = 1,2)\end{aligned} \qquad (5)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are the input vectors, and $\sqrt{\mathbf{d_k}}$ is used to limit on the value of the inner product and prevent it from being too large. The encoding process of the function *Attention()* is to match each word vector in the sequence Q with each word vector in the sequence K to obtain a similarity, and then multiply this similarity by the sequence V to better preserve the original input information and represent the overall information. Hence, we obtain the important information vectors $\mathbf{ET_i}$, $\mathbf{CT_i}$ and $\mathbf{AT_i}$, respectively.

The vector $\mathbf{ET_i}$ is multiplied by the vectors $\mathbf{E_i}$, the vector $\mathbf{CT_i}$ is multiplied by the vectors $\mathbf{C_i}$, and the vector $\mathbf{AT_i}$ is multiplied by the vectors $\mathbf{A_i}$ as follows. This operation can amplify the vector gaps between words in event mention to make those event-related words more important than those event-independent words.

$$\boldsymbol{ESt_i} = \boldsymbol{ET_i E_i} \quad (i = 1,2) \qquad (6)$$

$$\boldsymbol{CSt_i} = \boldsymbol{CT_i C_i} \quad (i = 1,2) \qquad (7)$$

$$\boldsymbol{ASt_i} = \boldsymbol{AT_i A_i} \quad (i = 1,2) \qquad (8)$$

We concatenate the vector $\mathbf{EG_i}$, which encoded by GRU, and the vector $\mathbf{ESt_i}$, and then input their result to the fully connected layer using the activation functions *tanh* and *sigmoid* to obtain the weighted vectors $\mathbf{EFt_i}$ and $\mathbf{EGt_i}$, respectively, as follows. Using the similar equations, we can also get the weighted vectors $\mathbf{CFt_i}$ and $\mathbf{CGt_i}$ from $\mathbf{CG_i}$ and $\mathbf{CSt_i}$, and get the weighted vectors $\mathbf{AFt_i}$ and $\mathbf{AGt_i}$ from $\mathbf{AG_i}$ and $\mathbf{ASt_i}$, respectively.

$$\boldsymbol{EFt_i} = tanh\left(\boldsymbol{W^f}[\boldsymbol{EG_i}, \boldsymbol{ESt_i}]\right) \quad (i = 1,2) \qquad (9)$$

$$\boldsymbol{EGt_i} = sigmoid(\boldsymbol{W^g}[\boldsymbol{EG_i}, \boldsymbol{ESt_i}]) \quad (i = 1,2) \qquad (10)$$

where $\mathbf{W^f}$ and $\mathbf{W^g}$ are the parameter matrix and vector, respectively.

We adopt the gating mechanism to obtain the event-related information flow $\mathbf{EUt_i}$ from the event mention as follows. Using the similar equations, we can also get the event-related information flow $\mathbf{CUt_i}$ from the context of the trigger and get the event-related information flow $\mathbf{AUt_i}$ from the argument list.

$$\boldsymbol{EUt_i} = \boldsymbol{EGt_i EFt_i} + (\mathbf{1} - \boldsymbol{EGt_i})\boldsymbol{EG_i} \quad (i = 1,2) \qquad (11)$$

Finally, we use the global maximum pooling to get the final event mention vector $\mathbf{EX_i}$, the context vector $\mathbf{CX_i}$ and the argument list vector $\mathbf{AX_i}$ as follows.

$$\boldsymbol{EX_i} = GlobalMax(\boldsymbol{EUt_i}) \quad (i = 1,2) \qquad (12)$$

$$\boldsymbol{CX_i} = GlobalMax(\boldsymbol{CUt_i}) \quad (i = 1,2) \qquad (13)$$

$$\boldsymbol{AX_i} = GlobalMax(\boldsymbol{AUt_i}) \quad (i = 1,2) \qquad (14)$$

D. Similarity Calculation

The above operations only processed each event mention itself and extracted its features. The principle of event coreference resolution is to measure the semantic similarity between an event mention pair. Hence, we must combine the information derived from an event mention pair to explore their relationship.

We use three distances to calculate the similarity between two event mentions. The Cosine distance (*Cos()*) calculates the angle between two vectors to measure the degree of similarity. Bilinear distance (*Binlear()*) is a simple way to incorporate the linear interactions between two vectors. Single Layer Network (*SLN()*) is to calculate the nonlinear distance between two vectors and we choose *tanh* as the activation function. Because of the existence of nonlinear activation functions, the Single Layer Network can capture nonlinear interactions between two event mentions. Since Bilinear distance focuses on capturing linear interactions and Single Layer Network focuses on capturing non-linear interactions, Single Layer Network can make up for the lack of Bilinear to some extent.

With Cosine distance, Bilinear distance and Single Layer Network, we measure the similarities and capture the linear and non-linear interactions between two event mentions as follows. Finally, we obtain the Cosine distance **CD**, the Bilinear distance **BD** and Single Layer Network distance **SD**.

$$\boldsymbol{CD}=Cos(\boldsymbol{EX_1},\boldsymbol{EX_2})=\frac{\boldsymbol{EX_1^T EX_2}}{\|\boldsymbol{EX_1}\|\cdot\|\boldsymbol{EX_2}\|} \quad (15)$$

$$\boldsymbol{BD} = Binliear(\boldsymbol{EX_1}, \boldsymbol{EX_2}) = \boldsymbol{EX_1^T W EX_2} \quad (16)$$

$$\boldsymbol{SD} = SLN(\boldsymbol{EX_1}, \boldsymbol{EX_2}) = Tanh(\boldsymbol{W}[\boldsymbol{EX_1}, \boldsymbol{EX_2}] + \boldsymbol{b}) \quad (17)$$

where **W** and $\boldsymbol{b}$ are the parameter matrix and vector, respectively.

E. Output

We merge the vectors using the global maximum pooling and three distances as follows.

$$\boldsymbol{Q} = Concat(\boldsymbol{EX_1}, \boldsymbol{EX_2}, \boldsymbol{CX_1}, \boldsymbol{CX_2}, \boldsymbol{AX_1}, \boldsymbol{AX_2}, \boldsymbol{CD}, \boldsymbol{BD}, \boldsymbol{SD}) \quad (18)$$

Then we merge the vector **Q** with the pairwise feature vector **P** to get the combined vector $\mathbf{V_f}$ as follows.

$$\boldsymbol{V_f} = Concat(\boldsymbol{Q}, \boldsymbol{P}) \quad (19)$$

The vector $\mathbf{V_f}$ is placed in a fully connected classifier which uses the activation function *Relu* as follows

$$\boldsymbol{V_h} = Relu(\boldsymbol{W^h V_f} + \boldsymbol{b}) \quad (20)$$

We get the confidence by the function *sigmoid* as follows.

$$\boldsymbol{score} = sigmoid(\boldsymbol{W^0 V_h} + \boldsymbol{b_0}) \quad (21)$$

We use the dropout in the fully connected layer to prevent over-fitting, which also improves robustness of the model, making the model easy to converge. Finally, following Fang, we use a closure to form an event chain that is determined as coreferential by the above classifier.

IV. EXPERIMENTS

In this section, we first introduce the dataset and experimental settings. Then, we report and analyze the experimental results on the ACE 2005 Chinese corpus.

A. Experimental Setup

The ACE 2005 Chinese corpus is the only available dataset for the task of document-level Chinese event coreference resolution. Since a portion of the documents in this corpus do not have coreferential event pairs, following Teng, we remove these documents form the corpus and obtains a new corpus containing 445 documents. We enumerate all event mention pairs of any two event mentions in the same document, excluding those event pairs with different event subtypes. Finally, a total of 14394 event mentions pairs are extracted and the positive-negative ratio is about 1:5.

For fair comparison, we perform 5-fold cross-validation on the ACE corpus. In each turn, the whole corpus is divided into training set, validation set and test set on the ratio of 3:1:1. MUC [15], B^3 [16], $CEAF_e$ [17] and BLANC [18] are used to evaluate the performance of event coreference resolution. Among them, MUC is the most important metric and it is a score based on event links. B^3 is a score based on event nodes, which makes up for the MUC's ignorance of the evaluation of non-coreferential events. $CEAF_e$ is similar to B^3, but it adds entities to evaluate the performance of event coreference resolution. BLANC measures the average performance between non-coreferential events and coreferential events. The evaluation of the above four metrics can comprehensively reflect the model performance in event coreference resolution. Following the previous work, we also use the average score (AVG) of the above four metrics as comparison metric.

Follow the previous work [19], we pre-trained the word embeddings with *Word2Vec* on the Wikipedia Chinese corpus. The dimension of word vector is set to 300 and the dimension of the position vector is set to 50. To prevent overfitting, the value of Dropout is set to 0.2. Besides, GRU neurons are set to 100 dimensions, and Bi-GRU neurons are set to 50 dimensions. Finally, the model training round is set to 20.

B. Experimental Results

To verify the effectiveness of our model GANN, we introduce three state-of-the-art systems as baselines: 1)**Teng**: the only available system on document-level Chinese event coreference resolution using a traditional model; 2) **Krause**: a CNN model; and 3) **Fang**: a multiple decomposable attention network model. Since Krause and Fang are two systems on English, we obtain their codes and modify their system to fit Chinese. Table 1 shows the performance comparison of four systems on the ACE 2005 Chinese corpus.

TABLE I. PERFORMANCE COMPARISON OF FOUR MODELS

System	MUC	B^3	BLANC	$CEAF_e$	AVG
Teng	73.5	/	/	/	/
Krause	66.52	86.08	75.31	77.34	76.31
Fang	69.87	87.77	77.86	80.53	79
GANN	74.59	89.18	80.32	82.56	81.66

From Table 1, we can find out that:

1) Compared with the traditional model Teng, our GANN improves the metric MUC by 1.09. Teng used the external tools to extract a large number of features, and we only use a few features provided the corpus. This result verifies that the neural network models are more suitable for Chinese event coreference resolution than the traditional models.

2) Compared with the CNN model Krause, our GANN gains an improvement of 5.35 in AVG with all improvements in four metrics. The reason is that Krause only performed convolution operations to extract the local information in event mentions, and they did not reweight the important information in event mentions. Moreover, they also did not consider the relationships between an event mention pair. Our GANN not only introduces the gated attention mechanism to reweight and select event-related information, but also calculates the semantic similarity between an event mention pair on three distances. This also illustrates the usefulness of our gated attention mechanism and three distances.

3) Compared with another neural network model Fang, our GANN also outperforms it on all metrics from 1.41 to 4.72. Although Fang used the attention mechanism, but they did not filter the noisy information. Besides, Fang only used the Cosine distance to obtain the linear relation between an event mention pair, while GANN uses three distances to capture the linear and nonlinear relations.

Among all metrics, the largest improvement of GANN comes from MUC. The metrics MUC is the most important metric in event coreference resolution and it scores on the edges of undirected graphs, which usually related to the coreferential events. This result derives from the ability of the gated attention mechanism to reweight and select event-related information from event mentions. Hence, GANN can identify more positive examples in which have all kinds of noisy information.

C. Results Analysis

To analyze the effectiveness of each component in GANN, we implement four simplified versions for comparison. The results are showed in Figure 2, where 1) -GA is a model whose gated attention mechanism is replaced by Bi-GRU and CNN, 2) –B&S is a model only using Cosine distance, 3) –C&S is a model only using Bilinear distance, and 4) –C&B is a model only using Single Layer Network.

TABLE II. COMPARISON OF GANN AND ITS FOUR SIMPLIFIED VERSIONS.

System	MUC	B^3	BLANC	$CEAF_e$	AVG
GANN	74.59	89.18	80.32	82.56	81.66
-GA	-5.53	-1.12	-2.61	-1.62	-2.72
-B&S	-3.73	-1.18	-2.73	-1.56	-2.3
-C&S	-4.38	-1.2	-2.29	-1.73	-1.11
-C&B	-3.27	-1.13	-2.52	-1.6	-0.7

Compared with GANN, the simplified version –GA reduces the AVG by 2.72 and this further ensures that the gated attention mechanism is helpful to resolve document-level Chinese event coreference.

The event-related information is often valuable for event coreference resolution. The attention mechanism plays a core role in GANN, which can reweight the features and then highlight the event-related information. Moreover, the gating mechanism can further filter event-independent information to enhance the representation of event mention. On the contrary, the Bi-GRU and CNN cannot enlarge the difference between event-related and event-independent information, and some noisy information will affect the discrimination of the model. Take the following two event mentions as examples.

S5：号称是最先进、设备最齐全的台北市 119 勤务中心今天正式**落成**启用。(119 service center of Taipei, which is the most advanced and fully equipped center, was officially **opened** today.)

S6：重新改建的勤务中心是在上午**落成**启用。(The rebuilt service center was **opened** in the morning.)

The gated attention mechanism can extract the event-related information both in S5 and S6 as “勤务中心落成启用” (the service center was opened) and ignore the rest noisy information. Another advantage of the gated attention mechanism is that it can reduce computational complexity and enhance the discriminative accuracy.

When we introduce only one distance to GANN, the results of –C&B, -C&S and –B&S in Figure 2 show that the combination of three linear and nonlinear distances can improve the performance of GANN. Different from the results in English event coreference resolution where Cosine distance plays a core role, nonlinear distance Single Layer Network is more suitable for Chinese event coreference resolution for its low performance drop (AVG: 0.7 vs 2.3). The reason derives from the characteristics of Chinese language in which an event has flexible and diverse expressions. The Cosine and Bilinear distances consider the semantic similarity between two event mentions and they do not consider the influence of a single word on the distance. On the contrary, nonlinear distance Single Layer Network fully considers the influence of a single word on the distance, and is suitable for Chinese for its flexible sentence structure. Furthermore, the results in Figure 2 also show that the combination of three distances can be complementary. If the three distances are used together, it is possible to comprehensively consider the relationships and interactions between event mentions in a multi-faceted manner, which is more helpful for the event coreference resolution.

The disadvantage of our GANN is that it cannot resolve the coreferential event mentions which have different words and structures. Currently, almost all coreference resolution models are similarity model, which relies heavily on the similarity between two event mentions. GANN is also a similarity model and it has the common problem in those similarity models. Take the following two event mentions as examples.

S7：凌晨 2 点南投分局出动了 90 名警力，一行人浩浩荡荡来到南投司法大厦前吴薇婉的**静坐**处，在经过分局长沟通说明之后，进行强制拆除。

(At 2 a.m, Nantou Branch dispatched 90 police officers. They went to the front of the Nantou Judicial Building where is the place of the sit-in **protest** for Wu Weiwan. After the director communicated with Wu, they removed all items by force.)

S8：今天早上她带着将近 30 多位的支持群众向警方**抗议**。

(This morning, she brought nearly 30 supporters to **protest** to the police.)

Since the words in event mentions S7 and S8 are almost completely different, this lead to the low similarity between the above two mentions, calculated by GANN. However, if the model can understand the context information according to the full text, it can understand that they are coreferential.

In this paper, we uses a few annotating features (e.g., arguments, tense and modality) to build a strong model. However, annotating these event information are time-consuming and laborious. Hence, how to resolve event coreference from less annotating features is a challenging task and it can be applied to real world applications. To make our neural network model more persuasive, we also evaluate our model on annotated event mentions, event types and triggers, but not use any other annotated information. We compare our GANN with the baseline Fang (the highest performance model in all baselines) and the results are shown in Table 3.

TABLE III. EVALUATION ON LESS ANNOTATED FEATURES

System	MUC	B^3	BLANC	$CEAF_e$	AVG
GANN	56.45	82.17	67.4	71.77	69.45
Fang	-7.67	+1.37	-1.4	+1.59	-1.53

Table 3 shows that our GANN outperforms Fang by 1.53 in AVG, especially the largest improvement in MUC (7.67). This result also can further verify the effectiveness of GANN for the task of document-level Chinese event coreference resolution. Besides, we can also find that AVG drops from 81.66 to 69.45 (-12.21) without the annotated entity and event attributes.

V. CONCLUSIONS

In this paper, we introduce a novel GANN model to the task of document-level Chinese event coreference resolution. Firstly, GANN introduces a gated attention mechanism to select the event-related information from event mentions and then filter noisy information. Secondly, GANN not only uses a single Cosine distance to calculate the linear distance between two event mentions, but also introduces multi-mechanisms, i.e., Bilinear distance and Single Layer Network, to further calculate the linear and nonlinear distance. Hence, linear distance and nonlinear distance can complement each other. The experimental results on the ACE 2005 Chinese corpus illustrate that our model GANN outperforms the state-of-the-art baselines. In the future, we will focus on the end-to-end and cross-document event coreference resolution.

REFERENCES

[1] M. Fouad and M. Atyah, "Efficient Topic Detection System for Online Arabic News," Proc. IJCA, 2018, pp. 7-12.

[2] L. Cheng, H. Gao and H. Wang, "A News Event Extraction Method in Chinese and Thai Languages Based on Dependency Tree Elements Combined with Rules," Software Guide, 2018.

[3] S. Swayamdipta, A. Parikh and Tom Kwiatkowski, "Multi-Mention Learning for Reading Comprehension with Neural Cascades," Proc. ICLR, 2018.

[4] S. Krause, F. Xu, H. Uszkoreit and D, Weissenborn, "Event Linking with Sentential Features from Convolutional Neural Networks," Proc. CoNLL, 2016, pp. 239–249.

[5] J. Fang, G. Zhou and P. Li, "Employing Multiple Decomposable Attention Networks to Resolve Event Coreference," Proc. NLPCC, 2018, pp. 246-256.

[6] A. Haghighi and D. Klein, "Simple coreference resolution with rich syntactic and semantic features," Proc. EMNLP, 2009, pp. 1152-1161.

[7] B. Yang, C. Cardie and P. Frazier, "A Hierarchical Distance-dependent Bayesian Model for Event Coreference Resolution," Computer Science, 2015.

[8] Z. Chen and H. Ji, "Graph-based event coreference resolution," Proc. ACL, 2009, pp. 54–57.

[9] Z. Chen, H. Ji and R. Haralick, "A pairwise event coreference model, feature impact and evaluation for event coreference resolution," Proc. ACL, 2009, pp. 17-22.

[10] Z. Liu, J. Araki, E. Hovy and T. Mitamura, "Supervised with-document-level event coreference using information propagation," Proc. LREC, 2014, pp. 4539–4544.

[11] J. Lu and V. Ng, "Joint Learning for Event Coreference Resolution," Proc. ACL, 2017, pp. 90–101.

[12] Z. Liu, T. Mitamura and E. Hovy, "Graph-Based Decoding for Event Sequencing and Coreference Resolution," Computational Natural Language Learning, 2018.

[13] J. Teng, P. Li and Q. Zhu, "Global Inference for Co-reference Resolution between Chinese Events," Acta Scientiarum Naturalium Universitatis Pekinensis, pp. 97-103.

[14] A. Parikh, O. Tackstrom and J. Uszkoreit, "A Decomposable Attention Model for Natural Language Inference," Proc. EMNLP, 2016, pp. 2249–2255.

[15] M. Vilain, J. Burger and J, Aberdeen. "A Model-Theoretic Coreference Scoring Scheme," Proc. ACL, 1995, pp. 45-52.

[16] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," Proc. LREC, 1998, pp. 563–566.

[17] X. Luo, "On coreference resolution performance metrics," Proc. EMNLP, 2005, pp. 25-32.

[18] M. Recasens and E. Hovy, "BLANC: Implementing the Rand Index for Coreference Evaluation," Proc. NLE, 2011, pp. 485-510.

[19] S. Xu, P. Li, G. Zhou and Q. Zhu, "Employing Text Matching Network to Recognise Nuclearity in Chinese Discourse," Proc. EMNLP, 2018, pp. 525-535.