Learning Event-aware Measures for Event Coreference Resolution

Yao Yao[1,2], **Zuchao Li**[3,*] **and Hai Zhao**[1,2,*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, 430072, P. R. China
yaoyao27@sjtu.edu.cn, zcli-charlie@whu.edu.cn,
zhaohai@cs.sjtu.edu.cn

Abstract

Researchers are witnessing knowledge-inspired natural language processing shifts the focus from entity-level to event-level, whereas event coreference resolution is one of the core challenges. This paper proposes a novel model for within-document event coreference resolution. On the basis of event but not entity as before, our model learns and integrates multiple representations from both event alone and event pair. For the former, we introduce multiple linguistics-motivated event alone features for more discriminative event representations. For the latter, we consider multiple similarity measures to capture the distinctions of event pairs. Our proposed model achieves new state-of-the-art on the ACE 2005 benchmark, demonstrating the effectiveness of our proposed framework.

1 Introduction

Knowledge-inspired natural language processing (NLP) may be generally conducted on the basis of entities. However, researchers are realizing that it is the form of "event" that can more comprehensively depict the knowledge clues in NLP tasks (Lee et al., 2012), among which coreference resolution is a fundamental task for either entity or event. Within-document event coreference resolution aims at finding all event mentions that refer to the same real-world event in a document (Lu and Ng, 2018).

Figure 1 shows an example of event coreference resolution from ACE 2005 (Walker et al., 2006) dataset. In the event coreference resolution, the trigger of an event mention usually refers to the word or phrase that describes the event with the most clarity and the event mention is typically the sentence containing the given trigger. As shown in the given example of Figure 1, the word *taken* triggers the *Transport* event. The words *wedding* and *marry* trigger the *Marry* event and these event mentions are coreferent because they both refer to the same real-world *Marry* event.

… attend his daughter ‘ s **wedding**$_{\{event1\}}$ ceremony …
… to be **taken**$_{\{event2\}}$ home later in the afternoon
to **marry**$_{\{event1\}}$ his eldest daughter …

Event	Type	Polarity	Modality	Genericity	Tense
event1	Life:Marry	Positive	Asserted	Specific	Future
event2	Movement: Transport	Positive	Other	Specific	Future

Figure 1: An event coreference resolution example from ACE 2005 dataset.

Compared with entity-level coreference resolution, event coreference resolution is more challenging despite their similarity (Lu and Ng, 2018). The main reasons are: (1) an event contains more complex syntactic information than an entity; (2) event coreference resolution suffers from error propagation since it has more upstream tasks; (3) there are far less triggers than entities in a document. Therefore, simply applying entity coreference resolution to event coreference resolution is unsatisfactory.

Many works utilize event linguistic features to address above challenges. In most cases, as shown in the above example, events that are coreferent have the same event features, such as type, polarity, and modality. Therefore, many studies (Lu and Ng, 2017b; Yu et al., 2020a; Lai et al., 2021; Lu et al., 2022) have suggested that incorporating event features into event coreference resolution is effective.

Despite of the developing trend of utilizing event features, the semantic similarity measures for event coreference resolution are less studied. Most of the works simply used element-wise multiplication to measure the similarity between two spans (Lee et al., 2017; Li et al., 2018, 2021) for entity-level

[*]Corresponding author. [†] Equal contribution. This work was supported by the Key Projects of National Natural Science Foundation of China (U1836222 and 61733011) and the Fundamental Research Funds for the Central Universities (No. 2042023kf0133).

Findings of the Association for Computational Linguistics: ACL 2023, pages 13542–13556
July 9-14, 2023 ©2023 Association for Computational Linguistics

NLP tasks. However, incorporating multiple similarity measures interestingly shows helpful in many NLP tasks other than our concerned coreference resolution. For instance, He and Lin (2016) proposed a deep pairwise word interaction model to measure semantic textual similarity between two text pieces, which directly utilized multiple similarity measures for answer selection task and semantic textual similarity measuring task. Liu et al. (2020) designed a U-shaped rewritten network to incorporate multiple similarity measures for context rewriting task.

Compared to entity-level coreference resolution, event-level coreference resolution focuses more on the cross-sentence semantic relationships, which are more complex and flexible. Therefore, to better address the event coreference resolution problem, we need to measure the relationships between events in more diverse and comprehensive ways. Inspired by this, in this paper, we propose a novel model for within-document event coreference resolution which integrates multiple measures from both event alone and event pair representations.

As far as we know, we are the first to introduce more comprehensive and distinguishing measures for event-level coreference resolution. Concretely, we introduce multiple linguistics-motivated event alone features for event alone representation and multiple similarity measures to capture the distinction of event pair. By conducting experiments on ACE-2005 benchmark, our proposed model shows a significant improvement and achieves a new state-of-the-art F1 score of 61.71% in end-to-end settings and 92.32% in all gold settings. Results demonstrate that our model can effectively models the event along features and event pair features, thus surpassing our baseline model and previous state-of-the-art models by large margins.

2 Related Work

Entities serve as the fundamental basis for conducting various knowledge-inspired natural language processing (NLP) tasks (Li et al., 2019, 2022; Yang et al., 2022; Stylianou and Vlahavas, 2021). Event coreference resolution, as an integral sub-task of natural language understanding, has gained increasing attention and finds applications in various domains. Most recent works of event coreference resolution can be categorized into two types: (1) joint models; (2) pipeline models.

In 2012, Lee et al. (2012) presented a joint model for cross-document coreference resolution. The model employed linear regression to build entity and event clusters and jointly solved events and entities references by handling both nominal and verbal mentions. Araki and Mitamura (2015) jointly formulated the event trigger extraction and event coreference as a problem of structured prediction to solve the error propagation problem. The authors utilized segment-based decoding with the multiple-beam search for event trigger identification and combined it with the best-first clustering for within-document event coreference resolution. Lu and Ng (2017a) proposed a joint model of trigger detection, event coreference resolution, and event anaphoricity determination. The model employed a structured conditional random field with two types of factors: (1) unary factors; (2) binary and ternary factors. Yu et al. (2020a) presented a Pairwise Representation Learning scheme for cross-document and within-document event coreference problems. The scheme jointly encoded text snippets pair by forwarding concatenated sentences into a transformer encoder and employed structured argument features to argument the pairwise representation.

Different from these works, in this paper, we adopt a more straightforward pipeline model conforming to the general idea of the task. Pipeline models mainly have two stages: event detection and event coreference resolution. Traditionally, pipeline models first detect event triggers, arguments, and event features and then apply coreference resolution to the predicted event mentions. Liu et al. (2016) first used bidirectional GRU to detect events and a logistic regression model to classify event features (i.e., realis). The model then employed the latent antecedent tree method to conduct coreference resolution. Choubey and Huang (2017) proposed a novel iterative approach for within-document and cross-document event coreference resolution. The method constructs event clusters gradually by exploiting inter-dependencies among event mentions in two stages. (Lai et al., 2021) employed OneIE (Lin et al., 2020) to extract event triggers and types. For other event features, the authors trained a simple classification model based on SpanBERT (Joshi et al., 2020). The model then used a context-dependent gated module (CDGM) to incorporate event features for event coreference resolution and a noisy training method to tackle the error propagation problem. Our work is closely related to (Lai et al., 2021). However, the methodologies are different. In our model, we found that a simple FFNN achieves a better result than the gating mechanism which questions the necessity for the CDGM gating method in event coreference resolution. We also innovatively introduce multiple similarity measures for event pair representation. With different ablation experiments and training settings, we verified that our proposed multiple measures learning model is effective and we hope our results can offer a new insight into not only event coreference resolution but also other fundamental tasks in NLP.

3 Method

3.1 Formalization

Our model focuses on within-document event coreference resolution. The input of our model is a document D containing n tokens. We then use a pre-tained language model to obtain the contextual embeddings of all the tokens in document D. Let $X = (x_1, ..., x_n)$ denote the token embeddings. For every word H_i, we use S^i_{token} and E^i_{token} to denote its subword tokens' start and end indices. Similarly, for the event e_i, $S^i_{trigger}$ and $E^i_{trigger}$ denote its event trigger's start and end indices.

3.2 Model Overview

Based on our formalization, we present a model flow for event coreference resolution. Our model first uses OneIE (Lin et al., 2020) to detect event triggers following the practice of previous work in a given document and classify its event type. Then a novel prompt-based event features prediction model generates every event alone features.

After obtaining the predicted event triggers and event alone features, we use a multiple measures learning model for event coreference resolution. The model overview is shown in Figure 2. Our model consists of two parts: single mention encoding and multiple measures pair encoding. Firstly, the single mention encoding is employed to construct event-level representation for the given document D. Secondly, the multiple measures pair encoding aims at building event pair representations combined with multi-similarity and event alone features. Finally, we use a scoring layer to calculate the antecedent scores for every event pair.

4 Trigger Detection and Event Alone Feature Classification

Our model uses OneIE (Lin et al., 2020) to detect event triggers and their types and designs a model to predict other event alone features. The overview of event alone features prediction model can be seen in Figure 3. Specifically, we first insert special tokens <t> and </t> around the event trigger to highlight the target event. By doing so, we can transfer the original event feature prediction problem into a traditional text classification problem. To be more specific, we first use BERT-large-cased (Devlin et al., 2019) to build contextualized representations h^r for the input sentence X, which can be defined as:

$$h^r = \text{Encoder}(X)$$

We then utilize MLP to classify each sentence. The probability distribution of event feature labels is calculated by:

$$M^r = \text{Pooler}(h^r))$$

$$P(y|X) = \text{softmax}(\textbf{MLP}(M^r))$$

where $y \in Y$ and Y denotes a set of pre-defined event linguistic feature labels. We choose the label with the highest probability score among all the event feature labels.

4.1 Event-aware Representation Learning

Our model integrates multiple linguistics-motivated and attention-based feature learning for event alone representations and multiple similarity measures to capture the distinction of event pair.

4.1.1 Event Alone Representation

Event alone representation includes linguistic-aware properties integration and linguistic-agnostic attention-based feature for an event. For linguistic-aware properties in our consideration for this work, type, polarity, modality, genericity and tense, which can be seen in Figure 1, are included. Their properties are used as linguistic-aware event alone features in our model. For feature encoding, we employed the same embedding method used in Lai et al. (2021). Given event e_i feature k, our model use a simple embedding layer to obtain its feature representation F^k_i.

Besides, considering data-driven feature learning, we introduce a linguistic-agnostic attention-based feature extraction module. Inspired by Dobrovolskii (2021), we proposed an attention based

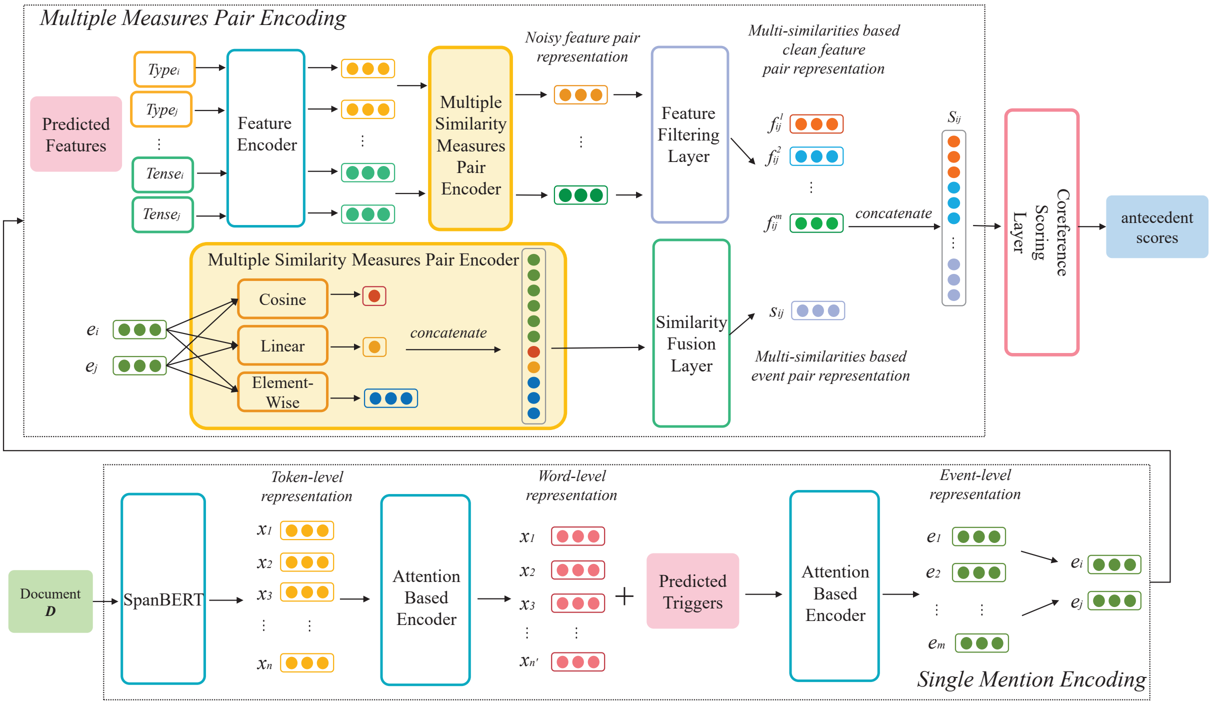


Figure 2: Overview of multiple measures learning model for event coreference resolution

method to encode event mentions. We calculate event representations as weighted sum of their triggers' token embeddings, instead of simply averaging them (Lai et al., 2021). To be more specific, for a given word H_i , its word-level representation h_i is calculated as where W_j is a learnable weight:

$$h_i = \sum_{j=S^i_{token}}^{E^i_{token}} W_j \cdot x_j$$

We then use the same attention strategy to compute event representation e_i where W'_j is a learnable weight :

$$e_i = \sum_{j=S^i_{trigger}}^{E^i_{trigger}} W'_j \cdot h_j$$

By using this attention based encoder, our model can generate more accurate event representations with a better focus on important tokens.

4.1.2 Event Pair Representation

After obtaining event alone representations, we design a novel multiple similarity measures for measuring the coreference relationship. Given event representation e_i and e_j, we use the following similarity measures to calculate their event pair features:

Linear Similarity The linear similarity between e_i and e_j is defined as:

$$Sim^{(ij)}_{lin} = W^T_{lin}[e_i; e_j] + b$$

where W_{lin} is a vector of trainable weights, b is a bias parameter, and [;] is vector concatenation operation.

Cosine Similarity The cosine similarity between e_i and e_j is calculated by:

$$Sim^{(ij)}_{cos} = \frac{e_i \cdot e_j}{\|e_i\|_2 \cdot \|e_j\|_2}$$

where $\|\cdot\|_2$ denotes L2 norm.

Element-wise Similarity The element-wise similarity between e_i and e_j is calculated by:

$$Sim^{(ij)}_{ele} = e_i \odot e_j$$

where $\odot$ denotes element-wise multiplication.

Then, we concatenate all three measures mentioned with event representation to build event pair representation e_{ij} which is calculated by:

$$e_{ij} = [e_i; e_j; Sim^{(ij)}_{lin}; Sim^{(ij)}_{cos}; Sim^{(ij)}_{ele}]$$

Similarity Fusion Layer After obtaining the pair representation, our model then employed a similarity fusion layer to construct the multi-similarity

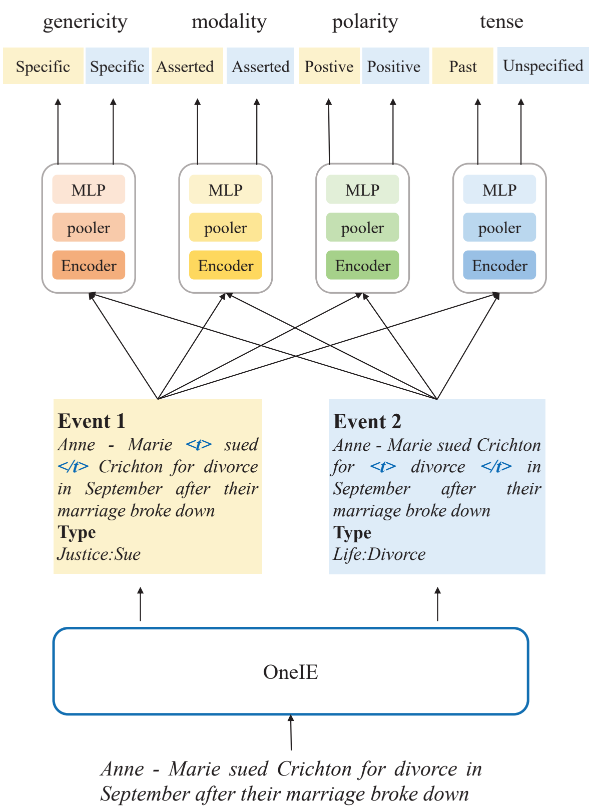

genericity
modality
polarity
tense
Specific
Specific
Asserted
Asserted
Postive
Positive
Past
Unspecified
MLP
pooler
Encoder
MLP
pooler
Encoder
MLP
pooler
Encoder
MLP
pooler
Encoder
Event 1
Anne - Marie <t> sued </t> Crichton for divorce in September after their marriage broke down
Type
Justice:Sue
Event 2
Anne - Marie sued Crichton for <t> divorce </t> in September after their marriage broke down
Type
Life:Divorce
OneIE
Anne - Marie sued Crichton for divorce in September after their marriage broke down

Figure 3: Overview of event alone features prediction model

based event pair representation s_{ij}, which is defined as:

$$s_{ij} = \mathbf{FFNN}_e(e_{ij})$$

where $\mathbf{FFNN}_e$ is a feed forward neural network mapping from $\mathbb{R}^{3\times d_1+2} \to \mathbb{R}^l$. By fusing event mentions and similarity measures all together, we are able to generate the event representation which is cleaner and contains richer information.

Feature Filtering Layer Similar to event pair encoding, for each event linguistic feature k, we too construct multi-similarity based feature pair representation $\hat{f}_{ij}^k$ by:

$$F_{ij}^k = [F_i^k; F_j^k; Sim_{lin}^{(ij)}; Sim_{cos}^{(ij)}; Sim_{ele}^{(ij)}]$$

$$\hat{f}_{ij}^k = \mathbf{FFNN}_f(F_{ij}^k)$$

where $\mathbf{FFNN}_f$ is a feed forward neural network mapping from $\mathbb{R}^{3\times d_2+2} \to \mathbb{R}^l$

Whereas predicted features can be noisy and hence cause the error propagation problem. To address this problem, we propose a feature filtering layer to clean the noisy features and focus on more important event alone features. For each feature k, we employed a feed forward neural network to filter the features.

$$f_{ij}^k = \mathbf{FFNN}_c^k(\hat{f}_{ij}^k)$$

By doing this, we are able to obtain clean event pair feature representations.

4.1.3 Coreference Scoring Layer

We then concatenate feature pair representations and event pair representation to get the multi-similarity based event-feature pair representation S_{ij}.

$$S_{ij} = [s_{ij}; f_{ij}^1; f_{ij}^2; ...; f_{ij}^m]$$

Next, we employ a coreference scoring layer to obtain antecedent scores between event i and event j where $\mathbf{FFNN}_s$ maps from $\mathbb{R}^{(m+1)\times l} \to \mathbb{R}$

$$A = \mathbf{FFNN}_s(S_{ij})$$

5 Experiments

5.1 Dataset

	Train	Dev	Test
#Docs	529	30	40
#Sent	19204	901	676
#Event	3342	327	293

Table 1: ACE 2005 dataset statistics (# denotes numbers)

We evaluate our model on ACE 2005 English dataset (Walker et al., 2006). ACE 2005 dataset adopts a strict notion of event coreference (Song et al., 2015). Specifically, two event mentions are coreferential if and only if "they had the same agent(s), patient(s), time, and location. We adopt the same split as Chen et al. (2015) and the detailed dataset statistics are shown in Table 1.

5.2 Model Setup

In order to have a fair comparison with Lai et al. (2021), we too use SpanBERT-base-cased (Joshi

Parameters	Value
Epochs	150
Batch size	16
Task learning rate	0.0001
Transformer learning rate	5e-5
Hidden size	500

Table 2: Training parameters for ACE 2005

Event Alone Features (k)	Noisy Probability (p_k)
Type	0.00
Polarity	0.00
Modality	0.15
Genericity	0.15
Tense	0.25

Table 3: noisy training probability for different features

et al., 2020) as the base Transformer encoder. We use AVG and CoNLL as our evaluation metrics. The AVG is the average F-score of four metrics (Lu and Ng, 2018): MUC, B^3, CEAF$_e$ and BLANC. The CoNLL score is the average of first three metrics. All models are trained for 150 epochs. The detailed training parameters can be seen in Table 2

According to Lai et al. (2021), simply training the model using predicted event alone features may hurt the performance since the prediction of test set contains more errors than that of the train set. Inspired by Lai et al. (2021), we too introduce the noisy training method. Specifically, for every predicted event alone features k, we randomly assign a new label with a probability of p_k. According to Lai et al. (2021), the probability p_k varies inversely with the discrepancy between the train and test accuracies. Our noisy training probability for different features can be seen in Table 3. The probability value is positively correlated with the accuracy of the feature prediction. The main idea behind this is that for event feature that is more prone to prediction errors, we need a higher replacement probability to eliminate the noise propagation problem. By using the noisy training method, our model can be more sensitive to noisy features and hence mitigate the error propagation problem.

5.3 Results and Discussion

To probe deeper into the effect of different stages in our pipeline model. We employ our model in three settings: end-to-end, gold triggers, and all gold. In end-to-end, we use predicted triggers and predicted features. In gold trigger setting, given gold triggers we predict event features and then apply event coreference. In all gold setting, we utilize gold triggers and gold features to have a better understanding on coreference stage.

5.3.1 Event Feature Prediction

The feature prediction model formulates the event feature prediction problem as a traditional text classification problem by using special tokens to highlight event triggers. The details can be seen in the appendix. Compared with the joint classification model proposed used by Lai et al. (2021), our model achieves obvious improvement on most event features. According to Lin et al. (2020), the event detection type-F1 score of OneIE on ACE 2005 test set is 74.7. Table 4 shows the accuracy of different event features.

From Table 4, we can see that comparing to CDGM (Lai et al., 2021), our model has a significant advantage on most of the features. We speculate that by introducing special tokens, we are able to maintain more semantic information than simply using the average of the trigger's token embeddings as the classification input.

5.3.2 End-to-end

The end-to-end event coreference result can be seen in Table 5. We use OneIE (Lin et al., 2020) to extract event triggers and their types. We then formulate the event feature classification problem as a traditional text classification problem.

From the Table 5, we can see that our model achieves the state of the art on ACE 2005 dataset. Although Peng et al. (2016) used cross-validation to utilize more training data, our model still shows a great improvement. Also, as a direct comparison, our model outperforms the CDGM (Lai et al., 2021) by more than 2 points on AVG, which shows the effectiveness of our proposed pipeline model.

5.3.3 Gold Triggers and Predicted Features

In order to have a thorough comparison with the CDGM model, we also perform our model on ground-truth triggers and predicted event alone features. The results are shown in Table 6. We achieve an AVG score of 86.63 which significantly improve the CDGM performance by more than 2.5 points. The result sufficiently proves the advantage of our proposed event alone feature classification and multi-similarity coreference model.

5.3.4 All Gold

To have a better focus on the event coreference stage of our model, we conduct our experiments on all gold setting in which we use ground-truth triggers and ground-truth features. Table 7 shows the overall result. In all gold setting, we use two training strategy: noisy training and clean training. In clean training strategy, all event features are not randomly replaced. We can see that by using original clean features, we improve the noisy training

Models		Event Features				
		Type	Polarity	Modality	Genericity	Tense
CDGM(2021)	train	99.9	99.9	99.9	99.9	98.4
	test	**95.3**	**98.8**	88.4	87.2	76.3
Ours	train	99.9	99.9	99.4	99.9	99.7
	test	**95.3**	98.6	**89.2**	**90.3**	**77.6**

Table 4: Results of event feature prediction on ACE-2005

Models		CoNLL	AVG
cross-validation			
SSED(2016)	+Supervised	55.23	52.50
	+MSEP	53.80	51.40
test data			
CDGM(2021) *(all features)*	+Simple	57.55	54.79
	+CDGM	58.99	56.32
	+Simple +Noise	60.43	57.85
	+CDGM +Noise	62.07	59.76
Ours *(all features)*	+Noise	**64.56**	**62.11**

Table 5: End-to-end results on ACE 2005

Models		CoNLL	AVG
PAIREDRL(2020b)		84.65	-
MSEP(2016)		80.37	82.90
CDGM(2021)	+Simple	75.32	74.94
(all features)	+CDGM +Noise	84.76	83.95
Ours *(all features)*	+Noise	**86.78**	**86.63**

Table 6: Gold triggers results on ACE 2005

strategy by near 5.5 points. As a direct comparison, our model outperforms the CDGM without noisy training method and reaches the state-of-the-art of 92.32. Such significant improvements demonstrate the effectiveness of incorporating multiple similarity measures and attention based encoding.

Models		CoNLL	AVG
CDGM(2021)	+Simple	85.75	85.40
(all features)	+CDGM	87.90	88.30
	+CDGM +Noise	85.40	85.38
Ours	+Noise	87.10	86.83
(all features)	-	**91.97**	**92.32**

Table 7: All gold results on ACE 2005

6 Further Exploration

6.1 Gated or Not?

According to Lai et al. (2021), they propose Context Dependent Gated Module (CDGM) to obtain the event features. CDGM selectively distill input event features using a gating mechanism. Therefore, to have a thorough comparison between CDGM and our proposed feature filtering layer, we introduce the CDGM method into our model. More specifically, after calculating the multi-similarity based feature pair representations, we employ CDGM instead of feature filtering layer to clean the event feature pair. We conduct experiments on all three settings and Table 8 shows the comparison result.

Settings	AVG		Δ
	CDGM	*Ours*	
End-to-end	61.38	62.11	+0.73
Gold triggers	85.72	86.63	+0.91
All gold	92.05	92.32	+0.27

Table 8: Results of using gated strategy and simple strategy with different settings on ACE 2005

From Table 8, we can see that feature filtering layer outperforms the CDGM in all three settings. The largest improvement is made in gold trigger setting which reaches 0.91 points. Whereas, in all gold scenario, CDGM reaches a similar result to our model with a relatively small margin of 0.27 points. We speculate that in all gold setting, the ground-truth event features are already clean and hence filtering features exerts limited influence on the overall performance. When using ground truth triggers, all event features are predicted and therefore containing errors. In this case, using feature filtering layer can evolve our model to its greatest potential. In conclusion, the result shows that, compared to CDGM, simply using a feed forward neural network can reserve more information and hence achieves obvious improvement.

6.2 Attention or Not?

To have a better analysis on our proposed attention based encoder, we conduct ablation experiments. In the experiments, all models simply concatenate noisy feature pair representation together without any filtering operation (e.g., CDGM; feature filtering layer). We denote this concatenate strategy as *simple*. To be more specific, We treat the CDGM model (Lai et al., 2021) without using cleaned features as the baseline which treats event representations as the average of their triggers' token embedding. To have a fair comparison, we also use noisy feature representations by stripping the feature filtering layer and multi-similarity measures off our model and simply keep the attention based encoder. The overall results are shown in Table 9.

Models		AVG
baseline		
CDGM(2021)	+ Average + Simple	85.40
Ours	+ Attention + Simple	**90.45**

Table 9: Results of using attention based encoder with all gold settings on ACE 2005

All models use the same event linguistic features. From Table 9, we can conclude that our model yields an absolute improvement of 5.05 points in AVG score. The great performance sufficiently proves the superiority of our attention based encoder. By introducing learnable weights to calculate the word-level and event-level representations, we believe that it can help our model focus on more critical tokens instead of simply averaging them. In a nutshell, our model acquires the ability to grab the core information for the given event through using an attention based encoder and hence achieves a better result.

6.3 Multi-similarity or Not?

We also conduct ablation experiments for the purpose of understanding the impact of different similarity measures. In the experiment, we deprive our model of multiple similarity measures pair encoder and similarity fusion layer and return to the original method which simply uses element-wise similarity to capture the relevance between two events. We treat the deprived model as the baseline. We then introducing the similarity fusion layer to generate more similarity-aware event pair representations. On the basis of fusion layer, we respectively add cosine similarity and linear similarity to better analyse the different influence they exert on the overall results. Finally ,we train our full model on multiple similarity measures. The detailed results can be seen in Table 10.

Models	AVG	Δ
-wo fusion layer		
Baseline(element-wise similarity)	*90.50*	-
-w fusion layer		
Element-wise similarity	90.92	+0.42
Cosine Similarity	89.20	-1.30
Linear Similarity	90.72	+0.22
Element-wise + Cosine Similarity	91.07	+0.57
Element-wise + Linear Similarity	91.73	+1.23
Multi-similarity	92.32	+1.82

Table 10: Results of using different similarity measures with all gold setting on ACE 2005

We can see clearly that our model outperforms the baseline model significantly by 1.82 points on the AVG score. When combined with element-wise similarity, cosine similarity achieves a relatively small improvement compared to linear similarity and even suffers a minor decrease when used alone. We speculate that the main reason is that cosine similarity is more sensitive to the angle between two vectors instead of their lengths. Whereas, with the help of trainable weights, linear similarity can better capture the relevance between events by learning different perspectives of vectors. Nevertheless, our model still shows a great improvement when using all three similarity measures indicating that different similarity measures specialist in different aspects. As a result, we believe that using multiple similarity measures can help the model better capture the similarity between two events from more diverse and comprehensive aspects.

6.4 Noisy or Not?

To better analyze the usefulness of the noisy training method under different circumstances, we compare the noisy trained models and clean trained models under all three settings. The overall results are shown in Table 11.

Settings	AVG		Δ
	Clean	*Noisy*	
End-to-end	57.00	62.11	+5.11
Gold triggers	78.87	86.63	+7.76
All gold	92.32	86.83	-5.49

Table 11: Results of using noisy features and clean features with different settings on ACE 2005

We denote the models which replace the event feature labels randomly as *noisy* and *clean* for models which do not. Compared to the end-to-end setting, the model under the gold trigger setting achieves a much more apparent improvement when using the noisy training method. We presume the main reason is that since the event trigger detection is far from perfect, the predicted event features would be more accurate when given the ground-truth triggers. With fewer error propagation problems, our model is able to uncover the deeper connections between event features. Similar to Lai et al. (2021), when using the all gold setting, however, our model suffers a decline of 5.49 points on the AVG score. Since all event features are clean in the all gold setting, our model do not need noisy training to tackle error event features and on the contrary, simply using clean features can already boost the performance to the most degree.

6.5 Performance on Different Length

By introducing attention based encoder and multiple measures pair encoder, we believe that we can accredit the great improvement of our model to better performances in long documents. Compared to CDGM, our attention based encoder can better construct word-level representation which can be less influenced by the distance and multiple similarity measures pair encoder further addresses the long-document problem.

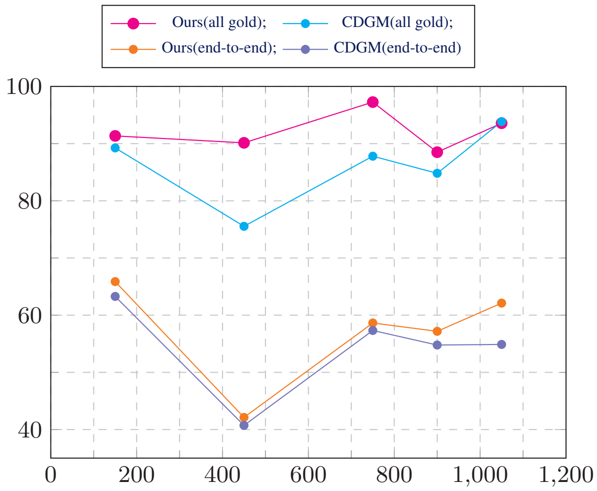


Figure 4: Performance of different document length in ACE 2005 test set with end-to-end and all gold settings.

To verify our hypothesis, we evaluate our model and the CDGM model respectively on the ACE 2005 test set for different document length ranges and the results are shown in Figure 4. According to the curves in the figure, in the all gold setting, it is obvious that our model do not suffer a significant decline as the document lengths grow. Although the CDGM model achieves a similar result with our model when the document lengths range between 1050 and 1199, our model still shows a great advantage over most document length ranges. Especially, when document length range grows from 300 to 900, the CDGM suffers a drastic decline while our model's performance remains at a high level. In the end-to-end setting, our model shows a more obvious advantage in the long documents over the CDGM model indicating that our method can effectively cope with the long documents in different settings.

7 Conclusion

In this work, we propose a multiple measures learning model for event coreference resolution for the first time, as far as we know. By incorporating multiple event features and similarity measures, we are capable of calculating antecedent scores more comprehensively from different aspects. We unleash the potential of employing multiple similarity measures and filtering event features through various ablation experiments and proved that our method can effectively address the long sentence dependency problem. Our model achieves an evident and significant improvement on the ACE 2005 benchmark compared to current state-of-the-art models. In the future work, we aim at exploring the potential of incorporating multiple similarity metrics into different tasks.

8 Limitation

The findings of this study have to be seen in light of some limitations. Compared to CDGM (Lai et al., 2021), introducing multiple metrics to capture flexible cross-sentence event relationships would cause extra computation cost and slightly slow our training. The training parameters of different models can be seen in Table 12. We deprive our model of multiple similarity metrics and denote it as the baseline model. We can see that our model increases 7% parameters compared to CDGM and 0.009% compared to the baseline model.

Models	#Training Parameters
Baseline	110.51M
CDGM(2021)	108.57M
Ours	110.52M

Table 12: The number of training parameters of different models (# denotes numbers)

References

Jun Araki and Teruko Mitamura. 2015. Joint event trigger identification and event coreference resolution with structured perceptron. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2074–2080, Lisbon, Portugal. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948, San Diego, California. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3491–3499, Online. Association for Computational Linguistics.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2401–2411. Association for Computational Linguistics.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6730–6737. AAAI Press.

Zuchao Li, Kevin Parnow, and Hai Zhao. 2022. Incorporating rich syntax information in grammatical error correction. *Inf. Process. Manag.*, 59(3):102891.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021. Syntax role for neural semantic role labeling. *Comput. Linguistics*, 47(3):529–574.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2846–2857, Online. Association for Computational Linguistics.

Zhengzhong Liu, Jun Araki, Teruko Mitamura, and Eduard H. Hovy. 2016. CMU-LTI at KBP 2016 event nugget track. In *Proceedings of the 2016 Text Analysis Conference, TAC 2016, Gaithersburg, Maryland, USA, November 14-15, 2016*. NIST.

Jing Lu and Vincent Ng. 2017a. Joint learning for event coreference resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–101, Vancouver, Canada. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2017b. Learning antecedent structures for event coreference resolution. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 113–118. IEEE.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5479–5486. ijcai.org.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. 2022. End-to-end neural event coreference resolution. *Artif. Intell.*, 303:103632.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Austin, Texas. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Nikolaos Stylianou and Ioannis P. Vlahavas. 2021. A neural entity coreference resolution review. *Expert Syst. Appl.*, 168:114466.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Yifei Yang, Zuchao Li, and Hai Zhao. 2022. Nested named entity recognition as corpus aware holistic structure parsing. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2472–2482. International Committee on Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020a. Paired representation learning for event and entity coreference. *arXiv preprint arXiv:2010.12808*.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020b. Paired representation learning for event and entity coreference. *CoRR*, abs/2010.12808.

A Appendix

A.1 Full Experiment Results

We report the all four metrics of our experiment results in Table 13.

A.2 Ablation Study

In order to show that adding extra feature filtering FFNN layer is indeed effective instead of simply using additional parameters to help the model fit, we deprive our model of the feature filtering layer and increase the number of training parameters to the same size as before (110.52M parameters). We found that the final AVG score is 90.05 which is 2.05 lower than the model with the feature filtering layer. The result shows that adding extra FFNN is necessary for alleviating the errors. In fact, cleaning the features means distilling reliable signals from noisy features by mapping and activation, and the noisy training method is used to handle the error propagation problem. We believe that by using individual FFNN for each feature, our model can filter reliable signals from noisy features and hence generate more accurate feature representations.

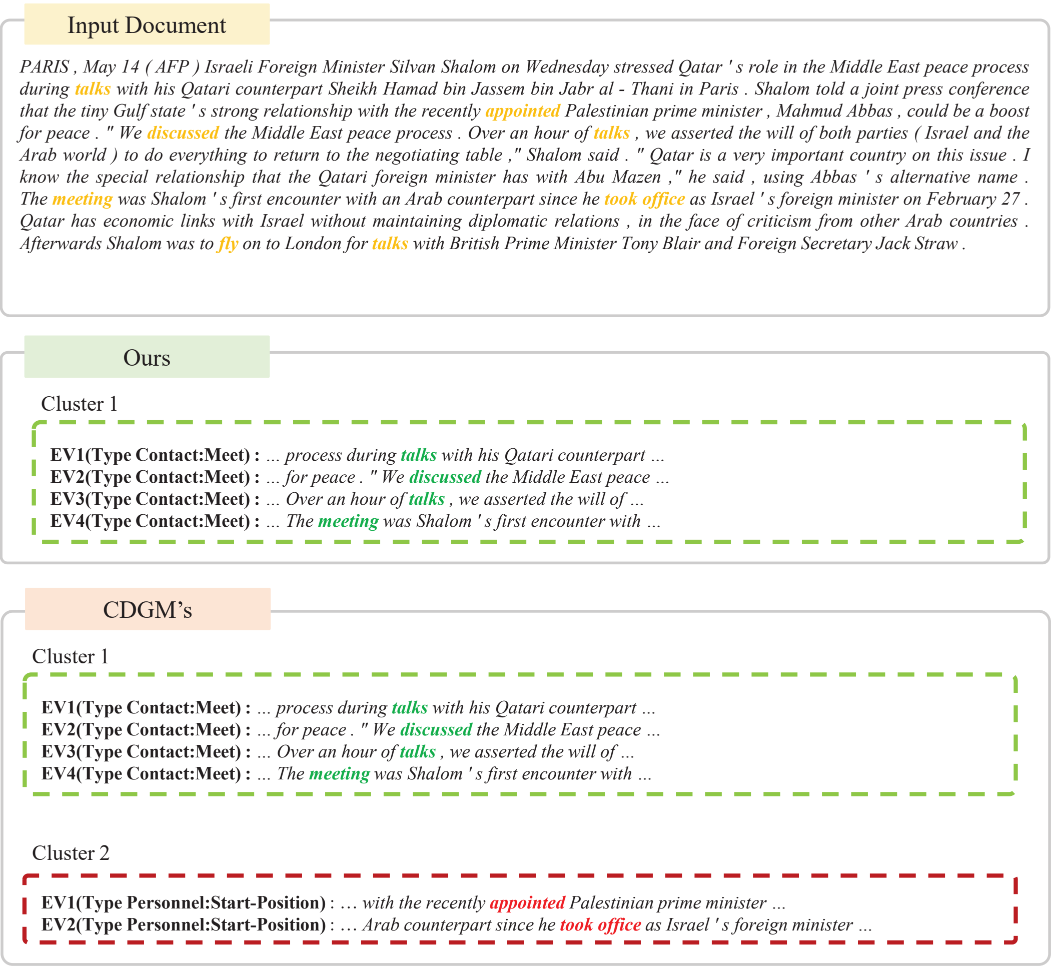
Secondly, to prove that calculating the similarity for the event-only representation and event feature representation separately is more effective than concatenating the original event-only representation with its feature representation and have a single similarity measures encoding step. We conduct an experiment where the representations are simply concatenated and go through a single similarity measurement. The results showed that the model will suffer a performance drop of 2.03 which sufficiently verified the importance of calculating the similarity measures separately.

A.3 Case Study

To facilitate a more illustrative comparison between our model and CDGM, a case study for the ACE 2005 dataset can be seen in Figure 5. We can see that compared to our model, CDGM additionally clusters "appointed" and "took office" into an event cluster. However, despite these two events having the same event type, they refer to two different persons being appointed as different ministers, and hence they are two different event. By incorporating different similarity measures and event features, we are able to analyse the coreference relation from different aspects, leading to a more accurate result.

Settings	Models	MUC	B^3	CEAF$_e$	BLANC	CoNLL	AVG
End-to-end	SSED + Supervised(2016)	47.10	59.90	58.70	44.40	55.23	52.53
	CDGM(2021)	-	-	-	-	62.07	59.76
	Ours	59.83	68.17	65.67	54.76	64.56	62.11
Gold Triggers	MSEP(2016)	68.00	92.90	87.40	83.20	82.77	82.88
	PAIREDRL(2020b)	76.10	90.70	87.20	-	84.65	-
	CDGM(2021)	-	-	-	-	84.76	83.95
	Ours	79.18	92.37	88.79	86.18	86.78	86.63
All Gold	CDGM(2021)	-	-	-	-	85.40	85.38
	Ours	87.44	95.48	92.99	93.38	91.97	92.32

Table 13: Full experiment results with all four metrics



Input Document

PARIS , May 14 (AFP) Israeli Foreign Minister Silvan Shalom on Wednesday stressed Qatar ' s role in the Middle East peace process during ***talks*** *with his Qatari counterpart Sheikh Hamad bin Jassem bin Jabr al - Thani in Paris . Shalom told a joint press conference that the tiny Gulf state ' s strong relationship with the recently* ***appointed*** *Palestinian prime minister , Mahmud Abbas , could be a boost for peace . " We* ***discussed*** *the Middle East peace process . Over an hour of* ***talks*** *, we asserted the will of both parties (Israel and the Arab world) to do everything to return to the negotiating table ," Shalom said . " Qatar is a very important country on this issue . I know the special relationship that the Qatari foreign minister has with Abu Mazen ," he said , using Abbas ' s alternative name . The* ***meeting*** *was Shalom ' s first encounter with an Arab counterpart since he* ***took office*** *as Israel ' s foreign minister on February 27 . Qatar has economic links with Israel without maintaining diplomatic relations , in the face of criticism from other Arab countries . Afterwards Shalom was to* ***fly*** *on to London for* ***talks*** *with British Prime Minister Tony Blair and Foreign Secretary Jack Straw .*

Ours

Cluster 1

EV1(Type Contact:Meet) : *... process during* ***talks*** *with his Qatari counterpart ...*
EV2(Type Contact:Meet) : *... for peace . " We* ***discussed*** *the Middle East peace ...*
EV3(Type Contact:Meet) : *... Over an hour of* ***talks*** *, we asserted the will of ...*
EV4(Type Contact:Meet) : *... The* ***meeting*** *was Shalom ' s first encounter with ...*

CDGM's

Cluster 1

EV1(Type Contact:Meet) : *... process during* ***talks*** *with his Qatari counterpart ...*
EV2(Type Contact:Meet) : *... for peace . " We* ***discussed*** *the Middle East peace ...*
EV3(Type Contact:Meet) : *... Over an hour of* ***talks*** *, we asserted the will of ...*
EV4(Type Contact:Meet) : *... The* ***meeting*** *was Shalom ' s first encounter with ...*

Cluster 2

EV1(Type Personnel:Start-Position) : *… with the recently* ***appointed*** *Palestinian prime minister ...*
EV2(Type Personnel:Start-Position) : *… Arab counterpart since he* ***took office*** *as Israel ' s foreign minister ...*

Figure 5: Case study for ACE 2005 dataset (The event triggers for input document are highlighted in yellow)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7 limitation

☑ A2. Did you discuss any potential risks of your work?
Section 7 limitation

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract Section 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 Experiment

☑ B1. Did you cite the creators of artifacts you used?
Section 4 Experiment

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Due to space limitation, we do not include these.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4 Experiment

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Due to space limitation, we do not include these.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 Experiment

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4 Experiment

C ☑ Did you run computational experiments?

Section 4 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 7 Limitation

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 Model Setup

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 Experiments Appendix B Full Experiment Results we report the average results of 3 different runs

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.2 Model Setup

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.