An Event Coreference Resolution Method Based on Multimodal

Hui Xu

guangjun@shu.edu.cn

School of Artificial Intelligence Shanghai University Shanghai, China

Abstract: It is crucial for many applications of Natural Language Processing (NLP) to recognize correfering events and entities that describe the same events in the real world. Despite this being an important task, the current research has mainly focused on text with very little attention towards other information sources, such as images. Consequently, we propose a neural architecture based on multimodal for Event Coreference Resolution (ECR), combining text features, image features, and text event graph features to identify coreference events. The experimental results demonstrate an average F1 score of 61.2%, which is an improvement on the previous event coreference model using a single text modality.

Keywords: event coreference resolution, multimodal event, multimodal information fusion

1. INTRODUCTION

With the rapid development of the Internet, news articles have become increasingly multimodal. For instance, a news story about thousands of people being left homeless by heavy rains in Brazil's Sao Paulo state, may include a traditional text article, a journalist's interview video, and images of collapsed houses. This variety of multimodal information helps provide a multi-faceted understanding of the news and allows users to focus more effectively on its content. ECR is the process of determining whether two words or phrases mentioned in the article describe the same real event in the real world; it is also the first step to such an understanding for machines.

Similarly, the recognition of text events that refer to the same entity or event is a key Natural Language Processing (NLP) task. Consider the two news headlines below, for example: 1) 2021 Nobel laureates in Literature goes to Abdulrazak Gurnah. 2) Abdulrazak Gurnah is awarded the Nobel laureates in Literature. Though people can easily and accurately discern the coreference relations between Abdulrazak Gurnah and the Nobel Laureates in Literature when reading these headlines, this task has been demonstrated to be quite challenging for machines. The state-of-the-art coreference resolution models not only accept textual input, but also require up to 2000 annotated articles to be properly trained [1].

Although this task is of great importance, research on coreference resolution has predominantly focused on the text content and given relatively little attention to the image information featured in news. Even though the correlation between the image and text modalities may be relatively

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334292

weak in news, they offer different perspectives on the same event. Most of the information necessary to resolve coreference resolution can be obtained from the text, however, the image information can be utilized as an additional supplement to the meaning of the event's trigger word. For example, without additional information, it is unclear whether “meet” refers to a meeting of two people, or a meeting of a broader group. However, if it is supplemented with images, we can understand the connection between the trigger words.

Although some studies use clustering method [2] for ECR, most work converts the ECR task into an event-pair classification task [3,4,5], i.e., determining whether there is a coreferent relationship between two events. Thus, encoding events into uniform vector space is an essential step in this task. Earlier work often used manual methods to manually select features [6]. In recent studies, trigger words, event parameters and other information [7,8] are encoded, and neural networks are used to identify event arguments. Finally, after the success of BERT model in the field of NLP, the way of event coding is changed to BERT, and the model can get better results.

Therefore, in this paper, we propose a method of coreference event resolution using image information. The experimental results show that the effect can be improved by introducing image modality information into ECR.

2.METHOID

2.1 Dataset

We construct a dataset for ECR, inspired by the image M2E2 dataset [8], which consists of 1000 manually annotated sentences, each paired with an image released by authentic news outlets such as Voice of American (VoA), British Broadcasting Company (BBC) and Reuters. The coreference labels are annotated by two annotators with two rounds of adjudication following ACE 2005 guideline. We randomly divided the training set and test set into 3:1. Four metrics of coreference resolution are used to evaluate models: the MUC, B3, CEAFe and AVG-F1 (average of above three metrics).

2.2 Model Architecture

The model we proposed mainly consists of three parts: multimodal feature extraction, feature fusion and coreferenced event prediction. The model structure is shown in figure 1.

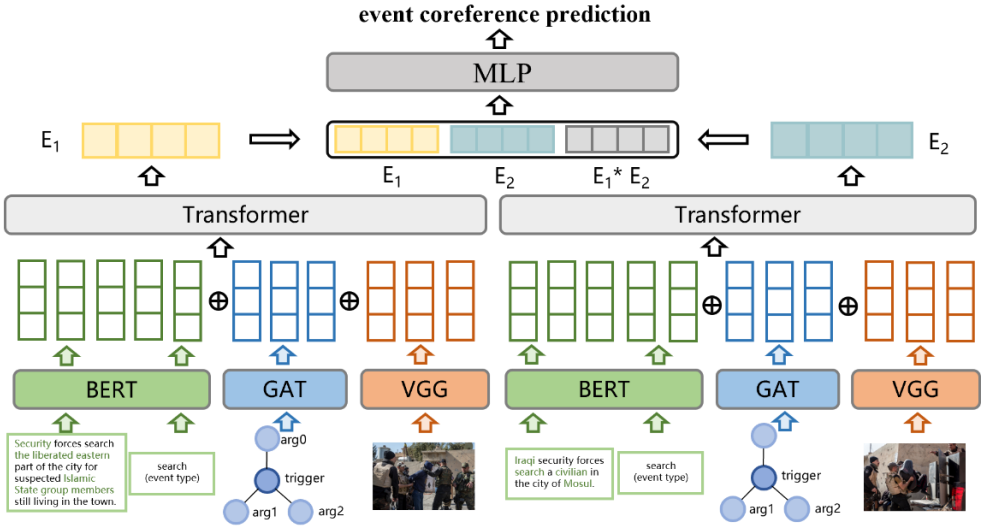

Figure 1 Basic framework of the proposed method

1) multimodal feature extraction

The multimodal features to be extracted mainly include text features, text event graph features and image features. First, the pre-trained language model BERT is used to generate feature vectors for input event trigger words, event argument elements and event types respectively, and then the text feature was obtained by concatenation above three features as T. Then, it is considered that the events parsed by coreference have structural similarity to the textual event graph composed of event trigger words and event arguments. Therefore, Graph Attention Networks(GAT) is used to extract the text event graph's structure feature and generate feature vectors with structure feature information. Then, the feature vectors are spliced to obtain the vector G of the text event graph. Finally, the image of the event pairs with coreference relationship is similar. Therefore, we use the VGG model to extract the image feature information as I.

2) Feature fusion

Based on obtained multimodal features, Transformer is adopted as an encoder to realize the fusion of text features, text event graph features and image features. By concatenation three features as (E') and input into Transformer, the output fused feature is E:

$$E = Transformer(E') \tag{1}$$

3) coreference resolution prediction

The two events multimodal features(E) and the product of the multimodal feature are taken as inputs:

$$F = E_1 + E_2 + E_1 \times E_2 \tag{2}$$

We use Multi-Layer perceptron(MLP) with two hidden layers as models to compute features and we select the highest score as the coreference event of the event. The MLP model consists

of three main parts: input layer, hidden layer and output layer. Each layer of the MLP model is fully connected, and two multi-mode event features can be fully calculated, and the final output is:

$$R = g\left(b^2 + W^3\left(sigmoid\left(\left(b^2 + W^2\left(sigmoid(b^1 + W^1 * x)\right)\right)\right)\right)\right) \quad (3)$$

Where g is softmax function, $W^1, W^2, W^3 W^4$ are weights between layers, b^1, b^2, b^3, b^4 are bias values.

2.3 Results

In order to prove the effectiveness of proposed method, the model is compared with two methods:

1)Lu&Ng2021[9]: event coreference is trained with five tasks: trigger detection, entity coreference, realis detection, anaphoricity determination, and argument extraction. This method designs penalty functions to guide learning of this model.

2)Xu&Li2022[10]: modal employs a multiple tensor to capture event embeddings at document, sentence, and topic levels.

The experimental results in table 1 show that, thanks to the integration of image modal information, text messages extracted by BERT, and the extraction of event graph structure features by GAT, our proposed model achieves a satisfactory extraction effect. In this paper, the incorporation of image modal information has been referenced. Compared to the baseline model, it experiences a remarkable improvement on AVG-F1, which indicates that the inclusion of image information into the event coreference parsing task can provide additional supplementary information for the events in the text and thereby improves the degree of differentiation between the event pair and other entities, thus enabling an accurate judgment of whether two events point to the same event in the real world.

Table 1 Experimental results

Modal	MUC(%)	B^3(%)	CEAFe(%)	AVG-F1(%)
Lu&Ng2021	45.2	54.7	53.8	48.0
Xu&Li2022	46.2	57.4	56.9	51.2
ours	**50.8**	**61.2**	**60.9**	**57.6**

4. Conclusions

In this paper, we first extract text event features, event graph structure features, and image features corresponding to the text event and utilize transformer encoder to combine above features to get a multimodal embedded representation of the event. Finally, using the above embedding, we employ MLP to determine whether there is a coreferent relationship between two events. In future work, we will continue to study how to utilize fine-grained objects in images to reduce noise.

References

[1] Lu, Y., Lin, H., Tang, J., Han, X., & Sun, L. (2022). End-to-end neural event coreference resolution. Artificial Intelligence, 303, 103632.

[2] Chen, Z., & Ji, H. (2009, August). Graph-based event coreference resolution. In Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4) (pp. 54-57).

[3] Krause, S., Xu, F., Uszkoreit, H., & Weissenborn, D. (2016, August). Event linking with sentential features from convolutional neural networks. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (pp. 239-249).

[4] Huang, Y. J., Lu, J., Kurohashi, S., & Ng, V. (2019, June). Improving event coreference resolution by learning argument compatibility from unlabeled data. 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) (pp. 785-795).

[5] Lu, J., & Ng, V. (2021, May). Span-based event coreference resolution. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13489-13497).

[6] Cybulska, A., & Vossen, P. (2015). " Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates. Int. J. Comput. Linguistics Appl., 6(2), 11-27.

[7] Lu, J., & Ng, V. (2017, July). Joint learning for event coreference resolution. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 90-101).

[8] Li, M., Zareian, A., Zeng, Q., Whitehead, S., Lu, D., Ji, H., & Chang, S. F. (2020). Cross-media structured common space for multimedia event extraction. arXiv preprint arXiv:2005.02472.

[9] Lu, J., & Ng, V. (2021, June). Constrained multi-task learning for event coreference resolution. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4504-4514).

[10] Xu, S., Li, P., & Zhu, Q. (2022, December). Improving Event Coreference Resolution Using Document-level and Topic-level Information. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 6765-6775).