Contrastive Learning for Event Extraction

Shunyu Yao
State Key Laboratory of Networking & Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, PR China; Yunnan Key Laboratory of Smart City and Cyberspace Security

Jian Yang
School of Mathematics and Information Technology, Yuxi Normal University, 653100; Yunnan Key Laboratory of Smart City and Cyberspace Security

Xiangqun Lu*
School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, 100876, PR China

Kai Shuang
State Key Laboratory of Networking & Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, PR China

ABSTRACT

Event extraction is an important information extraction task, aiming at extracting event information from text. Each event consists of triggers and arguments with specific roles. Event extraction methods first identify the trigger and classify it into specific types, and then find out the argument and its role. Traditional methods suffer from scarcity of supervised data. Most existing methods use pre-trained language models to solve this problem. Although these models greatly improved event extraction after fine-tuning, they still need a lot of supervised data to achieve the most advanced downstream task performance. In this work, we propose a simple but effective event extraction model CLEE, which leverages contrastive learning to pre-train language models. Our model first uses a pre-trained model to obtain the context representation of the word, and then uses a contrastive target to reduce the distance between triggers and arguments of the same event, and push the distance between triggers and arguments that are not in an event further. Experiments on ACE 2005 dataset show that CLEE achieves significant improvement.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**;

KEYWORDS

Neural network, Natural language processing, Event extraction, Contrastive learning

ACM Reference Format:
Shunyu Yao, Jian Yang, Xiangqun Lu, and Kai Shuang. 2022. Contrastive Learning for Event Extraction. In *2022 The 6th International Conference on Machine Learning and Soft Computing (ICMLSC 2022), January 15–17, 2022, Haikou, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3523150.3523176

*The corresponding author. Email: luxq@bupt.edu.cn

ICMLSC 2022, January 15–17, 2022, Haikou, China

ACM ISBN 978-1-4503-8747-7/22/01…$15.00
https://doi.org/10.1145/3523150.3523176

1 INTRODUCTION

Event extraction aims at extracting event structure from unstructured text, which is an important and challenging task in natural language processing. It includes event detection task to identify triggers and classify the event types, and argument extraction task to identify arguments and classify their roles [1]. Given a sentence, the event extraction model should identify the event types, triggers and arguments that appear in the sentence. Such structured knowledge can benefit many downstream tasks, such as question answering, language understanding and so on. The argument of an event is usually an entity (person, organization, date, etc.), which describes an event together. Each event argument plays a specific role. Traditional supervision methods defines this task as a classification problem, by assigning event triggers to event types from predefined sets. Event extraction is challenging due to the complex structure of events and the semantic gap between text and events.

Traditional event extraction methods follow the supervised learning paradigm to train neural networks in manually labeled datasets [2] [3] [4]. Large-scale pre-training has attracted extensive attention in natural language processing because of its powerful generalization ability and efficient use of large-scale data. Pre-trained language model (PLM) has been proved to be an effective method to improve various natural language processing tasks. Inspired by the success of pre-trained language models, some recent works [5] [6] tried to fine-tune the pre-trained language model (such as BERT [7]) to extract events. Benefiting from the strong general language comprehension ability learned from large-scale unsupervised data, these PLM-based methods achieved the most advanced performance in various public benchmark datasets. Pre-trained language model [7] [8] [9] has achieved excellent performance in text classification [10], named entity recognition [11], and question answering [12].

Benefiting from various effective self-supervised learning objectives, such as mask language model [7], PLMs can effectively capture the grammar and semantics in the text on a limited amount of labeled training data, and generate informative language representation for downstream NLP tasks. These methods work well in many public benchmark datasets, such as ACE 2005 [13] and TAC KBP [14], but they still face problems of data scarcity and limited generalization ability. The existing datasets are limited because annotating event data is particularly expensive and requires extensive human engineering. Therefore, they are not enough to train large-scale neural network models [15]. For given data, the contrast loss tries to reduce the distance between positive pairs. For example, positive pairs can be made by randomly changing the same image (using cropping, flipping and color distortion). Negative pairs can be randomly selected. [16] proved that MLM and NSP are also examples of contrastive learning. In the embedded space, the distance between triggers and arguments of the same event is pulled together, and the distance between triggers and arguments that are not in the same event is pushed away.

We introduce the contrast goal to strengthen the gap of contextual representation between triggers and arguments in different events, so as to make them easier to distinguish. We propose a contrastive learning framework for event extraction, CLEE, a new framework to improve the ability of PLM to understand triggers and arguments, aiming at better capturing the event information in the text. CLEE consists of a text encoder for learning event semantics. Specifically, in order to learn effective semantic representation of events, we use PLM as a text encoder. Our goal is that the expressions of triggers and arguments in the same event should be close in semantic space, while those with triggers and arguments that are not in the same event should be far away. This is done through contrastive learning, with word pairs in the same event as positive samples and word pairs not in the same event as negative samples. By fine-tuning the pre-trained model on the downstream event extraction dataset, CLEE can protect the traditional supervised event extraction from data scarcity. Experiments show that our model CLEE has achieved good performance compared with the model trained only on supervised datasets.

2 RELATED WORK

2.1 Event extraction

Most existing event extraction works follow the paradigm of supervised learning. Traditional feature-based methods rely on handmade features such as lexical features, syntactic features and external knowledge features to extract events [17] [18] [19] [20] [21]. In recent years, neural models have become the mainstream, which use neural networks to automatically learn effective features, including convolutional neural network [22] [2], recurrent neural network [3], graph convolution network [22] [23]. [2] proposed a neural network that uses convolutional neural network with dynamic multi-pooling. [3] proposed a joint model to get triggers and arguments at the same time using recurrent neural network.

Recently, pre-trained language models show remarkable improvement in some natural language processing tasks through fine-tuning. [24] proposed transformer architecture based on self-attention, which soon became the pillar of many subsequent language models. By pre-training on large-scale corpus, BERT [7] gained the ability to capture a large amount of common knowledge and made remarkable improvements in many tasks. With the recent success of BERT [7], the pre-trained language model has also been used for event extraction [6] [8] [25] [27]. Some efforts make use of the expressive power of pre-training models to effectively capture the general semantic and context-related information of words. Although they achieved success in ACE 2005 dataset [13], these PLM-based works only focused on fine-tuning, not pre-training of event extraction. Despite the success, pre-trained language model need large-scale supervised datasets while fine-tuning. When the labeled data is scarce, a large number of parameters of the model will lead to serious overfitting. Most previous studies were based on ACE 2005, a benchmark dataset annotated by humans. However, manually labeling large-scale training data is expensive, time-consuming and labor-intensive. This paper studies pre-training to make better use of abundant event knowledge in dataset.

2.2 Contrastive learning

Contrastive learning, as a popular unsupervised method, aims to learn representation by comparing positive and negative pairs. The comparison method minimizes the distance between the representations of similar positive samples and maximizes the distance between different negative samples. Contrastive learning has been widely used in computer vision [28] [29] [30] [31]. [32] taking two random transformations (such as cropping, flipping, twisting and rotating) of the same image as positive samples. Recently, a similar method has been adopted in language representation [33] [34], by applying enhanced technologies such as word deletion, reordering and replacement. However, due to its discreteness, data expansion in natural language processing is inherently difficult.

In natural language processing, many existing representational learning works can be regarded as contrastive learning methods, such as Word2Vec [35], BERT [7] and ELECTRA [36]. Contrastive learning has been regarded as an effective method to construct meaningful representations. Contrastive learning focuses on improving the ability of the model to distinguish given data from positive data (data sharing the same label) and negative data (different labels). For example, [37] suggests learning word embedding by taking words near the target word as positive samples and other words as negative samples. Recently, some studies [33] [38] [39] suggested using contrastive learning to train transformer models. However, they usually need data processing technology, such as reverse translation [40], or a priori knowledge about training data (such as order information). In natural language processing, contrastive learning is also widely used to deal with specific tasks, including question answering [41], discourse modeling [42], natural language reasoning [43] and relation extraction [44]. [45] suggests using contrastive learning for image titles, and [36] trained a discriminant model for language representation learning.

3 METHODOLOGY

The overall CLEE framework is shown in figure 1. We introduce the input embeddings in section 3.1, the contrastive learning method in section 3.2 and the event extraction method in section 3.3.

3.1 Input embeddings

Given a document with m sentences d = $\{S_1, S_2, \ldots, S_m\}$. Tokens in each sentence S_i are represented as $\{W_{i1}, W_{i2}, \ldots, W_{in}\}$, where n is the length of the sentence. The word embedding is then fed to the encoder to obtain the context representation. Our model leverages BERT as the encoder of sentences, which can effectively grasp universal semantic and contextual information because of its

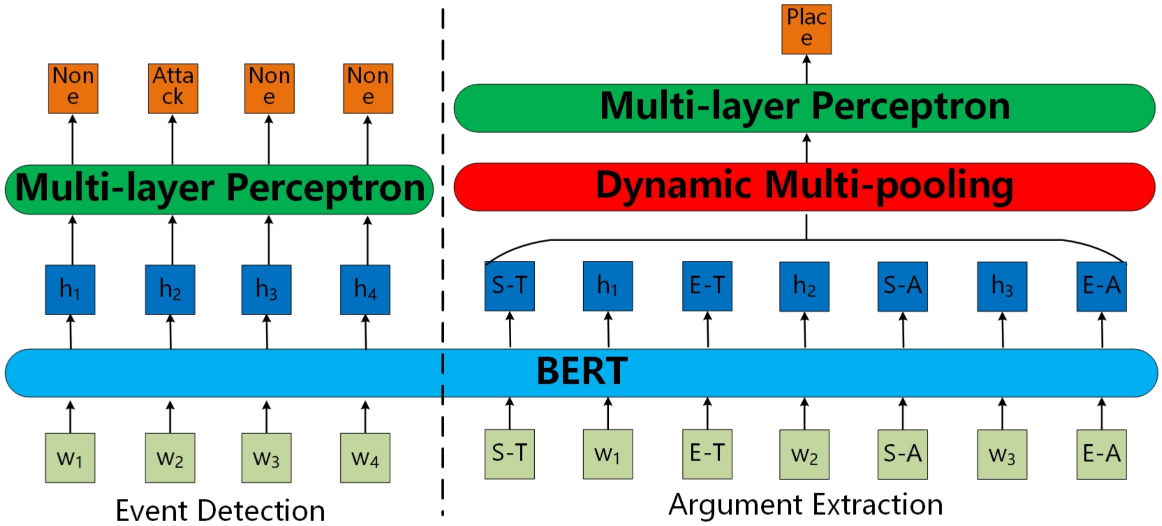

Figure 1: Architecture of the model.

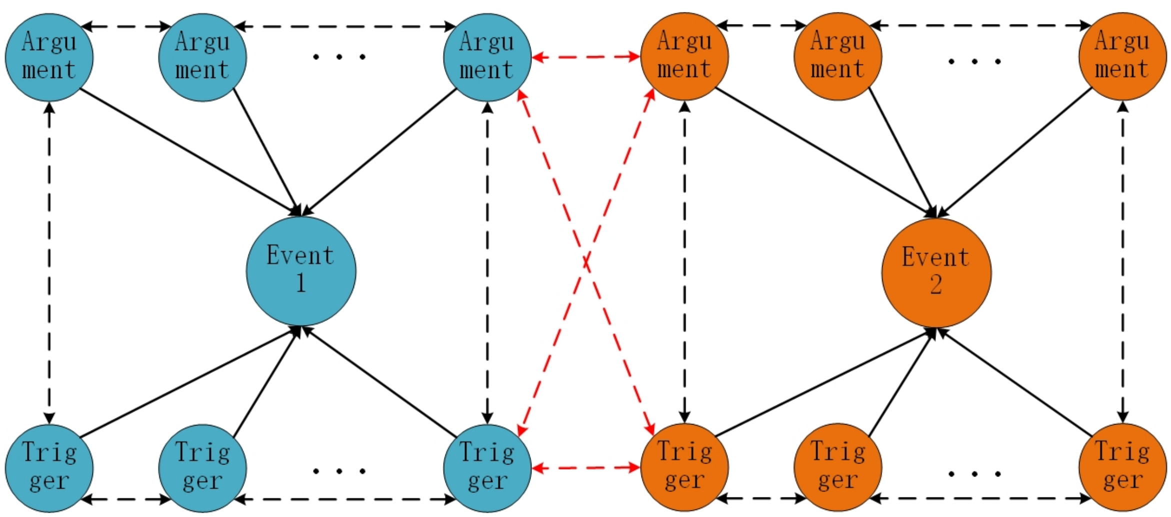

Figure 2: Construction method of contrastive learning dataset.

powerful representation ability. Through the encoder, we can get the context-aware embedding h_i of the sentence S_i.

$$\{h_{i1}, h_{i2}, \ldots, h_{in}\} = BERT(w_{i1}, w_{i2}, \ldots, w_{in})$$

Pre-trained language models (such as BERT) usually use WordPiece technology [46] [47] to tokenize words to reduce the size of the vocabulary so that a word can be divided into multiple word pieces. For example, the word "loving" can be segmented into two word pieces "lov" and "ing". Therefore, we use the average operation to get a fixed-size feature vector. Assuming that the hidden layer states of sub words corresponding to the target word w_t are from h_i to h_j, we average these hidden states. C is the contextual feature of the target word w_t, which is calculated as follows:

$$C = \frac{1}{j - i + 1} \sum_{k=i}^{j} h_k$$

In order to encode sentences with perception of trigger and argument, we added extra-special tokens "S-T", "E-T", "S-A", "E-A" to mark the positions of trigger and argument, and they are placed at the beginning and end of trigger and argument respectively.

3.2 Contrastive learning

Contrastive learning aims to learn effective representation by pulling semantically similar neighbors together and pushing non-neighbors away [48]. Contrastive learning distinguishes whether the relationship between two words is semantically close. We have developed a new contrastive learning framework to make full use of data. The goal of our framework CLEE is that triggers and arguments in the same event should be as close as possible in the hidden semantic space, while triggers and arguments that are not in the same event should be as far as possible.

Introducing more supervisory knowledge will be beneficial to event extraction, but it is label-intensive. Contrastive learning can make more in-depth use of labeling information in datasets. Labeling requires the manual efforts of experts, which is time-consuming and laborious. Although the latest development of pre-training models reduces the labeling workload, they still need a lot of labeled data to avoid overfitting. We propose a data generation strategy which uses supervised datasets to construct positive and negative sample pairs. A pre-training dataset can be built based on the existing labeled dataset without additional annotation by experts.

The construction methods of positive sample pairs are as follows: (1) Trigger-trigger pair of the same event. (2) Trigger-argument pair of the same event. (3) Argument-argument pair of the same event. The construction methods of negative sample pairs are as follows: (1) Trigger-trigger pair of different events. (2) Trigger-argument pair of different events. (3) Argument-argument pair of different events. These rules are shown in figure 2. It taught pre-training language models to understand an event by considering the relationship between triggers and arguments through distinguishing positive and negative sample pairs. We follow [30] using cross entropy loss for negative samples. Let $\mathrm{h_i}$ and $\mathrm{h^+}$ be the representations of positive samples. The training objectives of positive sample pairs are as follows:

$$L = -\log \frac{exp(sim\left(h_i, h^+\right)/\tau)}{\Sigma_{h\epsilon\{h^+,h^-\}}\exp(\frac{sim(h_i,h)}{\tau})}$$

Where τ is the temperature hyperparameter, and $\mathrm{sim(h_1, h_2)}$ is the cosine similarity $\frac{h_1^T h_2}{h_1 \cdot h_2}$.

In order to inherit language understanding ability of BERT and avoid catastrophic forgetting, we also add masked language model (MLM) task to our framework. The masked language model pre-training task randomly masks some tokens in the sentence and lets the model predict the masked tokens, which trains the model to grasp the rich semantic information. We train MLM and contrastive learning tasks at the same time.

3.3 Event extraction

We fine-tune our pre-trained CLEE and set the original RoBERTa without our pre-training of event semantics as an important baseline. In order to do ablation study, we evaluated a variant of CLEE on dataset: w/o S model uses original RoBERTa without event semantic pre-training.

Event trigger extraction. Event trigger extraction task is considered as token classification task, similar to named entity recognition. Trigger classifier is a simple multi-layer perceptron (MLP) layer with a hidden layer to classify triggers into 34 categories (33 events and none).

Event argument extraction. Since the training of downstream tasks is not the focus of our task, we use the dynamic multi-pool mechanism [5] to obtain the representation of features in the argument extraction stage and fine-tune in the supervised dataset. Candidate arguments are selected from entities in sentences. Each candidate argument will be paired with the trigger in argument role classification task. We use trigger and candidate argument to divide the sentence into three parts, and then make maximum pooling and splicing to get the final feature representation, and finally feed it into the classifier for classification. The classifier will classify each trigger-entity pair into one of 36 classes (35 argumentation roles, and "none" for entities without links to candidate triggers). Parameters are updated by minimizing the cross-entropy loss.

$$L = \frac{1}{n}\sum_{i=1}^{n} y_i log p_i$$

4 EXPERIMENT

Our experiment is designed to verify the effectiveness of our proposed framework CLEE.

4.1 Dataset

In order to verify the effectiveness of our model, we conducted an experiment on the widely used ACE2005 [13] dataset to evaluate our method. ACE 2005 contains 599 English documents with 8 event types, 33 subtypes and 35 argument roles. Following previous works [2] [19] [21] [49] [50] [51], we use the same data split with 40 newswire articles for the test set, 30 other documents for the development set and 529 remaining documents for the training set.

4.2 Evaluation metrics

Performance of event extraction is evaluated by the performance of two subtasks: event detection (ED) and event argument extraction (AE). We use the following criteria to evaluate the correctness of these two subtasks: (1) A trigger prediction is correct only if its span and type match with the labels. (2) An argument prediction is correct only if its span and all roles it plays match with the labels.

We report the accuracy (P), recall (R) and F1 scores as evaluation results.

Precision: the proportion of correctly predicted events in total predicted events.

Recall: the proportion of correctly predicted events in total gold events of the dataset.

F1-measure: $\frac{2 \times P \times R}{(P+R)}$

4.3 Hyperparameter settings

For the text encoder, we use the same model architecture as RoBERTa, which includes 24 layers, 1024 hidden dimensions and 16 attention heads. We start our event semantic pre-training from the released checkpoint. The learning rate of using warm up is in the first 10% steps. Learning rate is 5e-5, weight decay is 1e-5, batch size is 32, temperature is 1e-2 and dropout is 0.3. We use the Adam optimizer and set the training epoch to 10.

4.4 Overall performance

We start our pre-training from a well-trained PLM to get general language comprehension ability. CLEE has a good ability to distinguish triggers and arguments. We compare our method with the

Table 1: Overall performance on dataset.

	Trigger Identification			Trigger Classification			Trigger Identification			Argument Role		
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DMCNN	**80.4**	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5
JRNN	68.5	75.7	71.9	66.0	73.0	69.3	61.4	64.2	62.8	54.2	56.7	55.4
JMEE	80.2	72.1	75.9	**76.3**	71.3	73.7	**71.4**	65.6	68.4	**66.8**	54.9	60.3
CLEE w/o S	75.8	77.9	76.8	73.4	75.3	74.3	68.0	70.4	69.2	59.8	62.0	60.9
CLEE	76.9	**78.5**	**77.7**	74.7	**76.1**	**75.4**	68.9	**70.9**	**69.9**	60.7	**62.7**	**61.7**

following state-of-the-art models: (1) DMCNN, which used dynamic multi-pooling CNN [2], (2) JRNN, which is based on RNN [3], (3) JMEE, which is based on GNN [52]. Experiments have proved the effectiveness of our CLEE system. Compared with several benchmark models including the current state-of-the-art methods, it has a stable and significant improvement. Table 1 shows the comparison between different methods.

5 CONCLUSION

In this paper, we propose a contrastive learning framework CLEE for event extraction, a general framework for PLM to improve the understanding of triggers and arguments in events by using the abundant event knowledge in datasets through contrastive learning. We have proved the effectiveness of our method in downstream tasks. Experimental results on a standard dataset show that CLEE outperforms baseline. It means that CLEE helps PLM to better understand the relevance between triggers and arguments in events.

REFERENCES

[1] David Ahn. 2006. The stages of event extraction. In Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pages 1–8. Association for Computational Linguistics.

[2] 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, pages 167– 176. Association for Computational Linguistics.

[3] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In Proceedings of NAACL, pages 300–309.

[4] Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection.

[5] Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun,and Peng Li. 2019a. Adversarial Training for Weakly Supervised Event Detection. In Proceedings of NAACL-HLT, pages 998–1008.

[6] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In Proceedings of EMNLP-IJCNLP, pages 5784–5789.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

[8] Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. Exploring pre-trained language models for event extraction and generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5284–5294, 2019

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP1. Association for Computational Linguistics.

[11] Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Languageindependent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.

[12] Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4911–4921. Association for Computational Linguistics.

[13] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. Linguistic Data Consortium, 57.

[14] Mitamura, Teruko, Zhengzhong Liu, and Eduard H. Hovy. Overview of TAC KBP 2015 Event Nugget Track. TAC. 2015.

[15] Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In Proceedings of EMNLP, pages 1652–1671.

[16] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2020. A Mutual Information Maximization Perspective of Language Representation Learning. In Proceedings of ICLR.

[17] Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In Proceedings of ACL-08: HLT, pages 254–262. Association for Computational Linguistics.

[18] Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 369–372. Association for Computational Linguistics.

[19] Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 789–797.

[20] Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1127–1136. Association for Computational Linguistics

[21] Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 73–82.

[22] Huu Thien Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pages 365–371. Association for Computational Linguistics.

[23] Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event Detection: Gate Diversity and Syntactic Importance Scores for Graph Convolution Neural Networks. In Proceedings of EMNLP, pages 5405–5411.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar,Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.2017.At-tention is All you Need.In Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fer-gus, S. Vishwanathan, and R. Garnett (Eds.).5998-6008.

[25] Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019b. HMEAE: Hierarchical Modular Event Argument Extraction. In Proceedings of EMNLP-IJCNLP, pages 5777–5783.

[26] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada.

[27] Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juaizi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles.

[28] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Nonparametric Instance Discrimination. In Proceedings of CVPR, pages 3733–3742.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In Proceedings of NIPS.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of ICML, pages 1597–1607.

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In CVPR.

[32] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), volume 27.

[33] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

[34] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. arXiv preprint arXiv:2102.08473.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In Proceedings of ICLR.

[36] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of ICLR.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. NeurIPS.

[38] Hongchao Fang and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. arXiv preprint arXiv:2005.12766.

[39] John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. arXiv preprint arXiv:2006.03659.

[40] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In ACL.

[41] Yi-Ting Yeh and Yun-Nung Chen. 2019. QAInfomax: Learning Robust Question Answering System by Mutual Information Maximization. In Proceedings of EMNLP-IJCNLP, pages 3370–3375.

[42] Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models. In Proceedings of ACL, pages 4859–4870.

[43] Wanyun Cui, Guangyu Zheng, and Wei Wang. 2020. Unsupervised Natural Language Inference via Decoupled Multimodal Contrastive Learning. In Proceedings of EMNLP, pages 5511–5520.

[44] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In Proceedings of EMNLP, pages 3661–3672.

[45] Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 898–907.

[46] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, *et al.* 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. OpenAI blog, 1(8):9.

[48] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of CVPR, volume 2, pages 1735–1742.

[49] Shulin Liu, Kang Liu, Shizhu He, and Jun Zhao. 2016b. A probabilistic soft logic based approach to exploiting latent and global information in event classification. In Proceedings of the thirtieth AAAI Conference on Artificail Intelligence, pages 2993–2999. Association for Computational Linguistics.

[50] Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1789– 1798, Vancouver, Canada. Association for Computational Linguistics.

[51] Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. In Proceedings of AAAI, pages 4865– 4872.

[52] Xiao Liu, Zhunchen Luo, and Heyan Huang. Jointly multiple events extraction via attention-based graph information aggregation. arXiv preprint arXiv:1809.09078 (2018).