Word-level Commonsense Knowledge Selection for Event Detection

Shuai Yang, Yu Hong*, Shiming He, Qingting Xu, Jianmin Yao
School of Computer Science and Technology, Soochow University, SuZhou, China
{shytostu, tianxianer, smhelpai }@gmail.com, {qtxu0801, jyao}@suda.edu.cn

Abstract

Event Detection (ED) is a task of automatically extracting multi-class trigger words. The understanding of word sense is crucial for ED. In this paper, we utilize context-specific commonsense knowledge to strengthen word sense modeling. Specifically, we leverage a Context-specific Knowledge Selector (CKS) to select the exact commonsense knowledge of words from a large knowledge base, i.e., ConceptNet. Context-specific selection is made in terms of the relevance of knowledge to the living contexts. On this basis, we incorporate the commonsense knowledge into the word-level representations before decoding. ChatGPT is an ideal generative CKS when the prompts are deliberately designed, though it is cost-prohibitive. To avoid the heavy reliance on ChatGPT, we train an offline CKS using the predictions of ChatGPT over a small number of examples (about 9% of all). We experiment on the benchmark ACE-2005 dataset. The test results show that our approach yields substantial improvements compared to the BERT baseline, achieving the F1-score of about 78.3%. All models, source codes and data will be made publicly available.

Keywords: Event Detection, Commonsense Knowledge, ChatGPT

1. Introduction

Event detection (ED for short) is a subtask of Event Extraction. It is required to automatically identify trigger words occurring in sentences, and assign the appropriate event type to each of them (Xie and Tu, 2022). For example, the word "*slaughtered*" in (1) stands for a trigger word that signals the *Attack*-type event.

(1) **Mention**: *The war there, a direct spillover from the 1994 civil war in Rwanda, where government-led militia slaughtered an estimated 800,000 opposition.*[1]
Triggers: *war* (1), *war* (2), *slaughtered*
Type labels: *war* (1)←**Die**, *war* (2)←**Attack**, *slaughtered*←**Attack**

A variety of ED approaches have been studied, ranging from feature-based models (Ahn, 2006; Patwardhan and Riloff, 2009; Yang and Mitchell, 2016; Grishman, 2010) to advanced deep learning methods (Chen et al., 2015; Wang et al., 2019; Devlin et al., 2019; Liu et al., 2022; Li et al., 2021a,b). Recent studies concentrate more on the issues of 1) data sparsity (Lu et al., 2019) as well as 2) knowledge deficiency (Veyseh et al., 2021a). A variety of effective approaches have been proposed to address the issues, which can be roughly divided into the directions of data augmentation (Veyseh et al., 2021a; Gao et al., 2023) and knowledge enrichment (Tong et al., 2020).

Inspired by Tong et al. (2020)'s work, we tend to enhance ED models by knowledge enrichment. We leverage a knowledge base, i.e. ConceptNet (Speer et al., 2017), to obtain commonsense knowledge of words, and incorporate it into the contextualized word-level representations. The goal is to construct interpretable representations by supplementing conceptually-comprehensible information. For example, the item "*war causes death*" in ConceptNet stands for one of 15 pieces of commonsense knowledge of the word "*war*". It helps to improve the interpretability regarding why the earlier mentioned "*war*" in (1) triggers a *Die*-type event, instead of *Attack*-type.

To select relevant concepts to the contexts, we train a Context-specific Knowledge Selector (CKS). It performs multi-classification over all possible commonsense knowledge of a word, conditioned on its sentence-level context. A small number of examples (14,671 words) are taken for training, whose context-specific commonsense knowledge are predicted by ChatGPT[2] (Bahrini et al., 2023).

In our experiments, we use BERT-base-cased (BERT_{base}) (Devlin et al., 2019) to construct CKS, while BERT-large-cased (BERT_{large}) (Devlin et al., 2019) for ED. The test results on ACE-2005 dataset demonstrate the effectiveness of our approach, which obtains a F1-score of about 78.3%.[3]

2. Approach

We treat ED as a word-level classification problem. Given a sentence S={w_1,w_2, ..., w_n}, we classify each w_i in S into the appropriate event class. We consider 33 concrete ACE-2005 event classes (e.g.,

* Corresponding Author.

[1]The example is adopted from the benchmark ACE-2005 corpus. It can be accessed by looking-up using the ID of CNN_ENG_20030612_173004.10-17-EV0.

[2]https://openai.com/blog/chatgpt

[3]https://catalog.ldc.upenn.edu/LDC2006T06

LREC-COLING 2024, pages 17675–17682

Part-Of-Speech (POS)
FW/ JJ, JJR, JJS/ NN, NNS, NNP, NNPS
RB, RBR, RBS/ VB, VBD, VBG, VBN, VBP,VBZ

Table 1: Content words we considered in the study.

Datasets	**#Sen**	**#Content words**	**#Pro**
Training	7,036	14,671	9.04%
Validation	534	1,230	8.56%

Table 2: Statistics in the training and validation sets for building CKS. **Sen** denotes the number of sentence-level event mentions and **Pro** is the proportion of content words in all.

Attack, *Meet* and *Movement* classes) as well as non-trigger class in our experiments. Hence, multi-class classification is conducted.

2.1. Backbone and Baseline

Our ED model is constructed with an encoder-discriminator framework. BERT$_{large}$ (Devlin et al., 2019) is used as the encoder. It contains 24 transformer (Vaswani et al., 2017) blocks, performing context-aware attentive encoding. The input is formed as "[CLS] S [SEP]". The encoder computes the hidden state h_i with the size of 1024 ($h_i \in \mathbb{R}^{1\times1024}$) for each word w_i in S. The discriminator is constructed by a single-layer Fully-Connected (FC) network with Softmax. It maps h_i into the 34-dimensional probabilistic output embedding that correspond to all the considered event classes. Cross-entropy is used for training.

2.2. Commonsense-aware Classification

We intend to enhance the baseline model by knowledge enrichment. It is easily accomplished by combining the hidden state h_i with the representation k_i of a piece of commonsense knowledge. Accordingly, the discriminator $\mathcal{D}$ performs classification conditioned on both h_i and k_i:

$$\check{y} = \text{Softmax}\,(\mathcal{D}([h_i; k_i], \theta_d)) \quad (1)$$

where, the symbol [h_i;k_i] denotes the concatenation between h_i and k_i, $\check{y}$ is the predicted class label and $\theta_d \in \mathbb{R}^{2048\times34}$ is the parameters of $\mathcal{D}$.

2.3. Knowledge Collection

We collect commonsense knowledge from ConceptNet (Speer et al., 2017), a large knowledge base containing 1.5M nodes along with the in-between semantic relations. Knowledge collection is implemented by two steps, including knowledge retrieval and pattern-based knowledge formation.

Knowledge Retrieval– Given a word w_i, we use it as the head node to retrieve all of the relevant knowledge triples from ConceptNet. Each triple comprises the head node, a tail node and the in-between semantic relation. For example, one of the retrieved triples for the word "*war*" is {*war*|**Head**;*Causes*|**Relation**;*death*|**Tail**}. We consider six relation types in total, including "*CapableOf*", "*Causes*", "*IsA*", "*MannerOf*", "*MotivatedByGoal*" and "*ReceivesAction*".

Knowledge Formation– To facilitate sequence encoding, we convert each knowledge triple into a natural sentence using the fixed pattern. Specifically, we specify head and tail nodes as nominative and accusative, respectively. We embody the relation labels as readable predicates. For example, the relation label "*CapableOf*" is concretely converted into the predicate "*is capable of*". On this basis, we sequentially concatenate the nominative, predicate and accusative to form a sentence. We specify the resultant sentence as the denotation of a knowledge item. For example, the triple {*war*|**Head**;*Causes*|**Relation**;*death*|**Tail**} is formed as the denotation "*War causes death*".

2.4. Knowledge Selector

We collect commonsense knowledge and produce denotations merely for content words, excluding articles, function words, conjunctions, adverbs and named entities. The Parts-Of-Speech (POS) of the considered words are listed in Table 1. We utilize the open NLTK[4] for POS tagging.

Most of the nodes in ConceptNet hold a variety of commonsense knowledge, corresponding to multiple denotations (3.5 items in average). As a result, we obtain a large amount of redundant information from ConceptNet, which appears as unrelated knowledge to the contexts of words in ACE event mentions. Far more than useless, such knowledge is misleading during modeling word senses for ED.

To shield our classifier from unrelated knowledge, we propose a Context-specific Knowledge Selector (CKS). CKS is constructed with BERT$_{base}$ (12 transformer blocks) and a single-layer FC network. We train CKS in the task of multiple-choice Question Answering (QA), where CKS is required to select a sole denotation from the knowledge list $\mathcal{L}_i$ given the context $\mathcal{C}_i$ of w_i. Accordingly, the input question of CKS is formed as "[CLS] *Given* $\mathcal{C}_i$ [SEP] *the sense of* w_i is $\mathcal{L}_i$?". In $\mathcal{L}_i$, the denotations of w_i are concatenated with a special token "[OR]". Computationally, we adopt the hidden state q_i of [CLS] produced by the BERT$_{base}$ encoder, and feed q_i into the FC network for prediction. The output is specified as the index number of the context-related denotation in $\mathcal{L}_i$.

[4] https://www.nltk.org/

Datasets	#Sen	#Word	#Trigger
Training	14,618	249,104	4,258
Validation	817	18,281	493
Testing	629	18,119	438

Table 3: Statistics in the ACE-2005 training, validation and testing sets. **Sen** and **Word** denote the numbers of sentences and words in the corresponding set, respectively. **Trigger** denotes the number of triggers.

To train and develop CKS, we leverage ChatGPT[5] (Bahrini et al., 2023) to construct the training and validation sets. Specifically, we select 7,036 sentences from the ACE-2005 training set and 534 sentences from the ACE-2005 validation set. The statistics of content words in the sets is shown in Table 2. We utilize ChatGPT to predict context-related denotations for the content words in the sets, which are considered as the pseudo ground-truth data for training CKS. To ensure compatibility with CKS, we drive ChatGPT to likewise perform multi-choice selection. Therefore, a series of strict constraints are imposed upon the prompt to ensure reliable predictions. The details of the prompt and instruction can be accessed from the source files in the footnote.[6]

2.5. Supplemental Instruction

Given the selected denotation $\mathring{S}$ of the word w_i, we compute its hidden state k_i using the BERT$_{large}$ encoder. This encoder is shareable for ED and knowledge representation. When computing k_i, the input is formed as "[CLS] $\mathring{S}$ [SEP]". The output embedding of [CLS] is assigned to k_i. If a word fails to obtain a commonsense knowledge, we specify its k_i as the embedding of "NULL".

3. Experimentation

3.1. Experimental settings

Datasets– We utilize the benchmark ACE-2005 corpus for evaluating all the models in the experiments. It comprises 599 documents, 16,064 sentences and 285,504 words. We follow the common practice (Liu et al., 2022; Lu et al., 2021) to divide it into the training, validation and test sets. Statistics in the sets is shown in Table 3.
Evaluation Metrics– We use the Micro-averaged *Precision* (P), *Recall* (R) and F1-score (F1) as evaluation metrics. We report the average performance obtained in the 3-fold cross validation.

[5] https://openai.com/blog/chatgpt
[6] https://github.com/shyorangeshy/Code/blob/master/Prompt.pdf

Model	P	R	F1
Classification-based			
GatedGCN (Lai et al., 2020)	**78.8**	76.3	77.6
ONEIE (Lin et al., 2020a)	69.7	73.5	74.7
SA-GRCN (Liu et al., 2021a)	78.6	77.4	78.0
APEX (Wang et al., 2023b)	-	-	74.9
Generation-based			
BART-GEN (Li et al., 2021b)	69.5	72.8	71.1
TEXT2EVENT (Lu et al., 2021)	67.5	71.2	69.2
GTEE-DYNF (Liu et al., 2022)	63.7	**84.4**	72.6
COFFEE (Zhang et al., 2023)	-	-	75.7
BERT$_{large}$ (Baseline)	74.9	77.8	76.3
WCKS	78.7	77.9	**78.3**

Table 4: Performance (%) comparison in ACE-2005 dataset, where the discriminative and generative approaches are considered.

Model	P	R	F1
Data augmentation			
HNN (Ferguson et al., 2018)	**84.6**	64.9	73.4
DMBERT (Wang et al., 2019)	77.9	72.5	75.1
GAIL-ELMo (Zhang et al., 2019)	74.8	69.4	72.0
EDE (Li et al., 2022)	76.2	76.7	76.5
GPTEDOT (Veyseh et al., 2021a)	82.3	76.3	**79.2**
MTF (Gao et al., 2023)	-	-	69.6
DMCED (Chen et al., 2017)	75.7	66.0	70.5
SFT$_{AMR}$ (Xu et al., 2023)	-	-	75.0
DAEE (Wang et al., 2023a)	75.1	76.6	75.8
Knowledge enrichment			
GD (Nguyen and Grishman, 2018)	77.9	68.8	73.1
DEEB-RNN3 (Zhao et al., 2018)	72.3	75.8	74.0
ETEED (Ji et al., 2019)	74.1	**78.2**	76.1
EKD (Tong et al., 2020)	79.1	78.0	78.6
WCKS	78.7	77.9	78.3

Table 5: Performance (%) comparison over the ACE-2005 dataset, where data augmentation and knowledge enrichment are considered.

Hyperparameter Settings– For encoding the sentences in ACE-2005 corpus, we limit the maximum length of the input sentence to 128. Truncation and padding are used. During encoding the commonsense-knowledge denotations, we set the maximum length to 16. For CKS, the maximum length of the multi-choice question is set to 512. The batch size is set to 8, the number of epochs is set to 16, and the learning rate is set to 1e-5.

3.2. Results and Analysis

We compare with BERT$_{large}$ baseline and State-of-The-Art (SoTA) ED models. The performance is shown in Tables 4 and 5, where the previous studies are divided into the discriminative and generative models, as well as the ones using data augmentation and knowledge enrichment.

It can be observed that CKS yields a substantial improvement compared to the baseline, increasing the F1-score to 78.3% with the growth rate of 2%. This benefits from the higher precision, which is increased with a growth rate of 3.8% when a

Model	P	R	F1
BERT$_{large}$ (Baseline)	74.9	77.8	76.3
+RANDOM	75.7	78.1	76.8
+CKS	**78.7**	77.9	**78.3**
ChatGPT	77.8	**78.1**	77.9

Table 6: Verifyinging the effectiveness of CKS.

comparable recall is obtained. The test results demonstrate that CKS contributes to the detection and modeling of exact word senses for ED. Within the previous arts of extractive ED (denoted as *Classification-based* in Table 4), APEX (Wang et al., 2023b) is most similar to our approach. APEX expands the input sentence with prompts that depict all kinds of event types. By contrast, we expand the input with sample-specific relevant commonsense knowledge. The test results show that our approach outperforms APEX.

Besides, our model (BERT$_{large}$+CKS)[7] achieves a comparable performance with EKD (Tong et al., 2020) (78.3% versus 78.6% at $F1$-score). Though, our model is vest-pocket like a gadget. It only uses 14,671 denotations for knowledge enrichment during training, which are adopted merely at the stage of training the single FC layer for classification. By contrast, EKD trains the unabridged ED model using 733,848 automatically-annotated sentences that contain 2.65M triggers, where the powerful knowledge distillation is used. Nevertheless, EKD is more applicable to an open-domain ED scenario than our model. It is because EKD is perfectly generalized, dispensing with new knowledge during testing. Our model needs to provide context-related knowledge in real time during testing.

We fail to surpass the strong GPTEDOT (Veyseh et al., 2021a). It leverages GPT-2 to generate the same amount of in-domain annotated data with the original training set, and uses the augmented data to train a sophisticated BERT-based ED model within a multi-task distillation framework.

3.3. Effectivenss of CKS

We verify the effectiveness of CKS by comparing it to two models, including RANDOM and ChatGPT. RANDOM randomly selects one of the retrieved knowledge items, and combines it into the final representation for classification. It is conducted during both training and testing. ChatGPT is directly used to identify and classify triggers without additional learning, where the prompt for the multi-choice QA is used (see § 2.4). We show the performance in Table 6. It can be observed that RANDOM obtains an insignificant improvement. ChatGPT obtains a slightly lower performance than CKS.

[7] https://github.com/shyorangeshy/Code/blob/master/

4. Related Work

The earlier studies explore various feature engineering methods (Ahn, 2006; Gupta and Ji, 2009; Grishman, 2010; Patwardhan and Riloff, 2009). Statistical, pragmatic and syntactic features are used to identify triggers. To obtain general models, neural network based deep learning approaches are soon brought into the area of ED (Nguyen and Grishman, 2015; Chen et al., 2015; Duan et al., 2017; Lin et al., 2018; Chen et al., 2018; Wang et al., 2019; Lai et al., 2020; Li et al., 2021a), where CNN, LSTM and GCN as well as transformer-based Pretrained Language Models (PLMs) like BERT are used as backbones in the encoder-discriminator framework. Recently, other ED frameworks are proposed, including template-free generation (Zhang et al., 2023), conditional generation (Liu et al., 2022; Li et al., 2021b; Hsu et al., 2022) and mimetic translation (Paolini et al., 2021) frameworks.

Recent studies reveal that ED models are insufficiently trained because of data sparsity and knowledge deficiency. The resultant drawbacks occur when fine-tuning data-hungry PLMs for ED, as well as detecting the unseen/sparsely labeled triggers (Lu et al., 2019; Veyseh et al., 2021a). To address the issues, the previous work expands the training data by bootstrapping (Ferguson et al., 2018; Zhang et al., 2019; Cao et al., 2019) and distant supervision (Chen et al., 2017; Wang et al., 2019; Li et al., 2022). Besides, local knowledge is used for enhancement, such as trigger-event co-occurrence (Nguyen and Grishman, 2018), attentive document-level clues (Zhao et al., 2018), entity types (Ji et al., 2019; He et al., 2022) and event-type prompts (Wang et al., 2023b).

Most recently, a variety of novel data augmentation and knowledge enrichment approaches have been proposed. Veyseh et al. (2021a) fine-tune GPT-2 (Radford et al., 2019) to generate a new ED training set, where Teacher-Student knowledge distillation is used to avoid noises. Gao et al. (2023) develop a Mask-then-Fill data augmentation approach. It masks out dispensable fragments, and produces additional ED instances by regenerating the masked fragments using T5 (Raffel et al., 2020). Wang et al. (2023a) use a structure-to-text generation model for augmentation, where reinforcement learning is used for denoising. Xu et al. (2023) develop a framework of self-training with feedback, where the binary feedback (reliable/unreliable) of self-labeled external data is determined according to a compatibility score. Tong et al. (2020) construct a large set of open-domain trigger knowledge conditioned on sense-event correspondence, where WordNet (Fellbaum, 1998) is used. Similarly, knowledge distillation is conducted during remodeling over the knowledge base.

5. Conclusion

We utilize commonsense knowledge to enhance the word sense modeling for event detection. Experiments show that our approach yields substantial improvements compared to BERT$_{large}$. Besides, our model outperforms the previous work when data augmentation isn't used, and achieves comparable performance with the arts using data augmentation and knowledge enrichment. In the future, we will connect our model with a commonsense knowledge generator. It will be not only used to construct a generalized knowledge-based ED model, but solve the out-of-vocabulary problem.

6. Acknowledgement

We want to express gratitude to all anonymous reviewers and area chairs for their valuable time and insightful suggestions. This work is supported by National Science Foundation of China (62376182, 62076174).

7. References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. Chatgpt: Applications, opportunities, and threats. *CoRR*, abs/2304.09103.

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 261–270. Association for Computational Linguistics.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 409–419. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 167–176. The Association for Computer Linguistics.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1267–1276. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 352–361. Asian Federation of Natural Language Processing.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

James Ferguson, Colin Lockard, Daniel S. Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 359–364. Association for Computational Linguistics.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. 2023. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. *CoRR*, abs/2301.02427.

Ralph Grishman. 2010. The impact of task and corpus on event extraction systems. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, pages 2928–2931. European Language Resources Association.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 369–372. The Association for Computer Linguistics.

Shiming He, Yu Hong, Zhongqiu Li, Jianmin Yao, and Guodong Zhou. 2022. Event detection with cross-sentence graph convolutional networks. In *34th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2022, Macao, China, October 31 - November 2, 2022*, pages 334–341. IEEE.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1890–1908. Association for Computational Linguistics.

Yuze Ji, Youfang Lin, Jianwei Gao, and Huaiyu Wan. 2019. Exploiting the entity type sequence to benefit event detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 613–623. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5405–5411. Association for Computational Linguistics.

Rui Li, Wenlin Zhao, Cheng Yang, and Sen Su. 2021a. Treasures outside contexts: Improving event detection via global statistics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2625–2635. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021b. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 894–908. Association for Computational Linguistics.

Zhongqiu Li, Yu Hong, Jie Wang, Shiming He, Jianmin Yao, and Guodong Zhou. 2022. Unregulated chinese-to-english data expansion does NOT work for neural event detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2633–2638. International Committee on Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018. Nugget proposal networks for chinese event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1565–1574. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020a. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7999–8009. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020b. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7999–8009. Association for Computational Linguistics.

Anan Liu, Ning Xu, and Haozhe Liu. 2021a. Self-attention graph residual convolutional networks for event detection with dependency relations. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 302–311. Association for Computational Linguistics.

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5216–5228. Association for Computational Linguistics.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021b. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6418–6425. AAAI Press.

Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2019. Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4366–4376.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2795–2806. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Yixin Cao, and Aixin Sun. 2023. Few-shot event detection: An empirical study and a unified view. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11211–11236. Association for Computational Linguistics.

Thien Huu 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 and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 365–371. The Association for Computer Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5900–5907. AAAI Press.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Siddharth Patwardhan and Ellen Riloff. 2009. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 151–160. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *CoRR*, abs/2304.08979.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5887–5897. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Amir Pouran Ben Veyseh, Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021a.

Unleash GPT-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6271–6282. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Trung Ngo, Bonan Min, and Thien Huu Nguyen. 2021b. Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5403–5413. Association for Computational Linguistics.

Bo Wang, Heyan Huang, Xiaochi Wei, Ge Shi, Xiao Liu, Chong Feng, Tong Zhou, Shuaiqiang Wang, and Dawei Yin. 2023a. Boosting event extraction with denoised structure-to-text augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11267–11281. Association for Computational Linguistics.

Sijia Wang, Mo Yu, and Lifu Huang. 2023b. The art of prompting: Event detection based on type specific prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1286–1299. Association for Computational Linguistics.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 998–1008. Association for Computational Linguistics.

Zhipeng Xie and Yumin Tu. 2022. A graph convolutional network with adaptive graph generation and channel selection for event detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11522–11529. AAAI Press.

Zhiyang Xu, Jay Yoon Lee, and Lifu Huang. 2023. Learning from a friend: Improving event extraction via self-training with feedback from abstract meaning representation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10421–10437. Association for Computational Linguistics.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 289–299. The Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, and Dong Wang. 2023. Zero- and few-shot event detection via prompt-based meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7928–7943. Association for Computational Linguistics.

Meiru Zhang, Yixuan Su, Zaiqiao Meng, Zihao Fu, and Nigel Collier. 2023. COFFEE: A contrastive oracle-free framework for event extraction. *CoRR*, abs/2303.14452.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intell.*, 1(2):99–120.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 414–419. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1237–1252. Association for Computational Linguistics.