

GENERATING DISENTANGLED ARGUMENTS WITH PROMPTS: A SIMPLE EVENT EXTRACTION FRAMEWORK THAT WORKS

Jinghui Si^{*♣♣♥} Xutan Peng^{*◇} Chen Li^{♣♣} Haotian Xu[♥] Jianxin Li^{♣♣✉}

[♣]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China

[♣]State Key Laboratory of Software Development Environment, Beihang University, China

[♥]Alibaba Group, China [◇]Department of Computer Science, The University of Sheffield, UK

{sijh, lichen, lijx}@act.buaa.edu.cn x.peng@shef.ac.uk albert.xht@alibaba-inc.com

ABSTRACT

Event Extraction bridges the gap between text and event signals. Based on the assumption of trigger-argument dependency, existing approaches have achieved state-of-the-art performance with expert-designed templates or complicated decoding constraints. In this paper, for the first time we introduce the prompt-based learning strategy to the domain of Event Extraction, which empowers the automatic exploitation of label semantics on both input and output sides. To validate the effectiveness of the proposed generative method, we conduct extensive experiments with 11 diverse baselines. Empirical results show that, in terms of F1 score on Argument Extraction, our simple architecture is stronger than any other generative counterpart and even competitive with algorithms that require template engineering. Regarding the measure of recall, it sets new overall records for both Argument and Trigger Extractions. We hereby recommend this framework to the community, with the code publicly available at <https://git.io/GDAP>.

Index Terms— Event Extraction, Argument Extraction, Prompt-based Learning, Constrained Sequence Generation

1. INTRODUCTION

Event Extraction, which aims to extract structured event signals from plain text, is a crucial but challenging Information Extraction task [1, 2, 3]. In the literature, an event is typically defined by a schema, which includes the event type and a set of corresponding roles. Generally speaking, to fill in this schema, an Event Extraction system needs to find *triggers* that suggest an event, and ultimately, to locate the *arguments* that play different roles. Fig. 1 illustrates a real-world example with two events. For the event ‘CONVICT’ which can be triggered by ‘*convicted*’, we need to extract arguments ‘*Toefling*’ and ‘*Copenhagen*’ for roles ‘defendant’ and ‘place’. As for another event ‘ATTACK’ where ‘*assaulting*’ serves as a

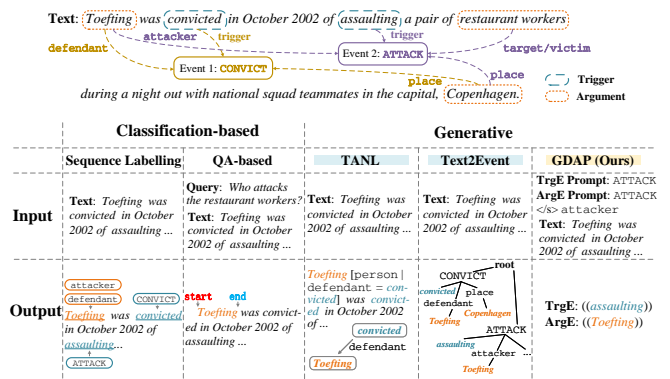


Fig. 1. Upper: A sentence with annotated event records. Lower: The taxonomy of Event Extraction algorithms. We highlight TANL and Text2Event here, as they are most relevant to the proposed method.

trigger, its linked arguments ‘*Toefling*’, ‘*Copenhagen*’, and ‘*restaurant workers*’ should be picked for roles ‘attacker’, ‘place’, and ‘target/victim’, respectively.

Early studies formulate Event Extraction as a token-level classification problem, i.e., to directly locate the triggers and arguments in the text and identify their categories. Many of these works simply adopt sequence labelling techniques based on Neural Networks [4, 5, 6, 7, 8, 9]. However, such methods only capture the internal pattern of input sequences without utilising the knowledge of label semantics. Therefore, another research strand, namely QA-based approaches, emerges [10, 11, 12]. With prepared templates, they first augment the training corpus by generating questions that are respectively targeting event types, triggers, and arguments. Next, the models learn to locate spans in the original sentences as answers, thus explicitly introducing the label knowledge. Nevertheless, the performance of these methods heavily depends on the quality of question templates, while designing them requires high-level expertise and massive human labour.

Very recently, instead of following the classification paradigm, a new wave of algorithms frame Event Extraction as a generation task. TANL [13], which is a pipeline

This work is supported by the State Key Laboratory of Software Development Environment (SKLSDE-2020ZX-12) and The NSFC through grants (No.U20B2053 and 61872022). Symbol * indicates equal contribution.

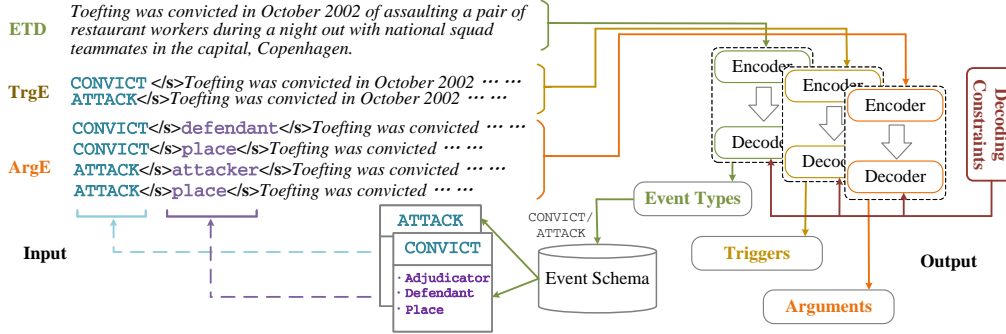


Fig. 2. Architecture overview. **ETD**, **TrgE**, and **ArgE** are the abbreviation of Event Type Detection, Trigger Extraction, and Argument Extraction, respectively. For output examples of TrgE and ArgE, please see the lower right corner of Fig. 1.

powered by the pre-trained T5 [14], learns to *sequentially* ‘translate’ plain text input into sequences where event triggers and arguments are marked, respectively. Another T5-based generative model, namely Text2Event [15], instead attempts an end-to-end manner where the output has a complicated tree-based structure. Thanks to the success of large-scale pre-trained language models, such generative approaches can reduce the manual engineering for the templates to the minimum, hence superior to the aforementioned QA-based methods. Nevertheless, they still exhibit bottlenecks that limit real-world applications. (1) They focus on incorporating the label semantics (as constraints) during decoding but fail to fully exploit such signals (e.g., event types and triggers) on the encoding side. (2) Akin to their classification-based counterparts, generative models assume dependency between Trigger and Argument Extractions, thus implementing these two modules either serially or jointly. However, this long-standing hypothesis is challenged by our observations, e.g., in a simple input sentence ‘*We put the Shah of Iran in power*’, the trigger ‘*put*’ is hardly beneficial to extracting arguments ‘*Shah*’ and ‘*Iran*’ either grammar or semantically. (3) For TANL, a considerable percentage of the generated tokens are task-irrelevant; for Text2Event, the output structure can be too complex to scale.

To alleviate the above issues, in this paper, we propose a novel framework that **generates disentangled arguments with prompts** (GDAP). As its name suggests, GDAP achieves three remarkable algorithmic improvements. (1) To effectively inject knowledge via various label semantics when encoding the input, for the first time we introduce prompt-based learning to the domain of Event Extraction. (2) Unlike *all* existing methods, GDAP disentangles the extraction of triggers and arguments, substantially enhancing the computational parallelism. (3) With both the architecture and the output format hugely simplified, GDAP is easy to be implemented and extended. To empirically verify the effectiveness of these advancements, we conduct extensive experiments on the standard ACE 2005 benchmark [2], where 11 strong baselines (both classification-based and generative) are involved. In the Argument Extraction task, GDAP yields the best F1 score among all generative methods, which is even competi-

tive with state-of-the-art baselines that rely on hand-designed templates. Moreover, GDAP scores the overall highest recall in both Argument and Trigger Extractions, indicating promising applications in commercial scenarios.

2. METHODOLOGY

As shown in Fig. 2, GDAP possesses three functional modules, namely Event Type Detection, Trigger Extraction, and Argument Extraction. In practice, the high diversity of event types in the schema will lead to a large range of potential trigger and argument selections, making a comprehensive traversal too expensive to afford. Thereupon, all input sentences will first pass through the Event Type Detection module to reduce computational overhead. Based on the predicted event types, GDAP will then process Trigger and Argument Extractions *independently* and *simultaneously*. As discussed in § 1, this is the first attempt of applying such a disentangled design on Event Extraction, to our knowledge. For simplicity, all these three modules hold a similar architecture¹, i.e., an encoder-decoder network based on a pre-trained language model. Please refer to details in the subsequent paragraphs.

Event Type Detection. This module learns to encode a raw sentence and decode its x event types using the Parenthesis Representation [16] as

$$((\mathbf{ET}_1)(\mathbf{ET}_2))\cdots(\mathbf{ET}_i)\cdots(\mathbf{ET}_x)),$$

where \mathbf{ET}_i denotes the i -th event type and is enclosed by special symbols ‘(’ and ‘)’, e.g., the golden output for the sentence in Fig. 1 is ‘((CONVICT)(ATTACK))’. Due to the constraints of this special output format, the conventional decoding algorithms for text generation (e.g., greedy search and beam search), which step-wisely select the token purely based on prediction probability, cannot warrant structural validity here. Inspired by Text2Event, we design a finite-state machine whose states of token production (whether to decode ‘(’, ‘)’, or an event type) is determined by the counts of already generated ‘(’ and ‘)’. Besides, when decoding event types, the subword vocabulary may form tokens that are not within the candidate pool, e.g., ‘TCONTTVIC’ is a false gen-

¹However, they are independently trained without parameter sharing.

eration using subwords ‘CON’, ‘VIC’, and ‘T’. Therefore, we turn to the tree-based constraint decoding algorithm [17, 18], which guarantees the token validness by ensuring the search is only performed within a pre-built subword tree.

Trigger Extraction. We introduce the recipe of prompt-based learning to this module. The input is composed of a sentence **Sent**, one already detected event type **ET_i**, and a special separating token (denoted as \mathcal{T}_{sep} ; in practice we implement it as ‘</s>’, see § 3.1), as

$$\mathbf{ET}_i \mathcal{T}_{sep} \mathbf{Sent}.$$

While previous methods either fail to integrate label semantics during decoding or can only import such information through templates designed by experts, we find that the very simple prompt can effectively instruct GDAP to extract triggers relevant to the semantics of the event type label, in a fully data-driven fashion. Concretely speaking, if **Sent** contains y triggers corresponding to **ET_i**, the expected output is

$$((\mathbf{Trg}_1)(\mathbf{Trg}_2)\dots(\mathbf{Trg}_y)),$$

where **Trg₁** to **Trg_y** all come from the vocabulary of **Sent**. As the format here is similar to that of the Event Type Detection module, at the decoding stage we adopt the same mechanism, i.e., the aforementioned finite-state machine and the tree-based constraint decoding algorithm.

Argument Extraction. Like the Trigger Extraction module, our Argument Extraction module also attends the composition of a prompt and the input sentence:

$$\mathbf{ET}_i \mathcal{T}_{sep} \mathbf{RT}_{ij} \mathcal{T}_{sep} \mathbf{Sent},$$

where **RT_{ij}** is the j -th role type relevant to **ET_i** and can be decided by querying the established event schemae, e.g., role types of event ‘CONVICT’ are ‘defendant’ and ‘place’ (cf. § 1). As for the decoder side, if z arguments are obtained, the Argument Extraction module outputs a sequence whose format is similar to that of the Trigger Extraction module, as

$$((\mathbf{Arg}_1)(\mathbf{Arg}_2)\dots(\mathbf{Arg}_z)),$$

where **Arg₁** to **Arg_z** are also from the vocabulary of **Sent**.

We argue that apart from the enhanced encoder that can absorb valuable label semantics, the decoder of GDAP also achieves outstanding advancements beyond existing generative Event Extraction algorithms. On the one hand, although a large partition of words in **Sent** are irrelevant to Event Extraction, they are still included by TANL. In contrast, the output of GDAP only contains the extracted targets (triggers or arguments) without redundancy, which significantly improves data efficiency. On the other hand, while the tree-based decoding format of Text2Event is very complex and thus hard to scale, the generating format of GDAP is a simple list-style sequence and can therefore be easily extended to other tasks. We leave exploring this direction as an important future work.

Negative Sampling. When training the modules for Trigger and Argument Extractions, we introduce a simple yet effective negative sampling mechanism that makes our model more fault-tolerant. To be exact, for each **Sent**, we randomly select N event types that have not appeared. The model should learn *not* to extract triggers or arguments when such negative

samples appear in the prompt; instead, it should only generate an empty sequence, i.e., ‘(())’. It is worth noting that while increasing N contributes to the extraction robustness, it can lead to a significant training time boost as the number of training samples grows by approximately $N + 1$ times.

3. EXPERIMENTS

3.1. Setup

Dataset. The English partition in the ACE 2005 benchmark [2] is the *de facto* standard of Event Extraction tests. It has 599 documents annotated by 33 different event types. We adopt the popular splits released by [19], where there are respectively 17172, 923, and 832 sentences for training, validating, and testing. We also perform the preprocessing steps using the script of [19].

Baselines. To evaluate the Event Extraction efficacy of GDAP, we consider 11 strong baselines from a wide range, including (1) *methods based on sequence labelling*: LSTM-based **dbRNN** [6], RNN/GCN-based **JMEE** [7], BiGRU-based **Joint3EE** [20], BERT-based **DYIE++** [19], ELMO-based **GAIL** [21]; (2) *QA-based methods*: element-centred **BERT_QA** [11], multi-turn **MQAEE** [12], style-transfer-inspired **RCEE.ER** [10]; (3) *generative methods*: **TANL** [13] and **Text2Event** [15] (recall § 1 for detailed introductions). For fair comparisons, all baselines (including JMEE and RCEE.ER) and our method do not utilise golden entities as they are unlikely to be available in real-world settings.

Configurations of GDAP. Parallel to the generative baselines (TANL and Text2Event), GDAP adopts the pre-trained T5 as the backbone for each module, with both base (**T5-B**) and large (**T5-L**) versions tested. To align with the original implementation of T5, we choose ‘</s>’ as the separating token \mathcal{T}_{sep} . We leverage golden event type labels when composing prompts during training. Throughout all experiments, for cost-performance tradeoff, we set N in negative sampling at 4 and 2 for Trigger and Argument Extractions, respectively. Identical to Text2Event, we fix the random seed at 421. The learning rate is set at $5e-5$. We utilise label smoothing [22] and AdamW [23], and try the number of epochs within {20, 25, 30} to optimise validating scores.

Metrics. Following past studies [4, 7], we report the precision (**P**), recall (**R**), and F1 score (**F1**) of Trigger and Argument Extractions. Note that the output is marked as correct only when both text spans and predicted labels match with the ground-truth reference. In most industrial scenarios, arguments are the end product of an event extraction system, hence we attach greater importance to Argument Extraction than Trigger Extraction in this paper.

3.2. Result and Analysis

The main results of our experiments are listed in Tab. 1. As mentioned in § 3.1, we first focus on the Argument Extraction tests, where the F1 score measures the overall performance of

Table 1. Results of Event Extraction tests. Baseline performance is adapted from the original publications (NB: TANL has not attempted T5-L). **Bold** and underlined numbers are the best results for all and generative models, respectively.

(%)	Trigger			Argument ☆		
	P	R	F1	P	R	F1
<i>Classification-based</i>						
dbRNN	-	-	69.6	-	-	50.1
JMEE	-	-	-	-	-	50.4
Joint3EE	-	-	69.8	52.1	52.1	52.1
DYGIE++	-	-	69.7	-	-	48.8
GAIL	74.8	69.4	72.0	61.6	45.7	52.4
BERT_QA	71.1	73.7	72.4	56.8	50.2	53.3
MQAEE	-	-	71.7	-	-	53.4
RCEE_ER	-	-	-	-	-	58.7
<i>Generative</i>						
TANL (T5-B)	-	-	68.4	-	-	47.6
Text2Event (T5-B)	67.5	71.2	69.2	46.7	53.4	49.8
Text2Event (T5-L)	<u>69.6</u>	<u>74.4</u>	<u>71.9</u>	<u>52.5</u>	55.2	53.8
GDAP (T5-B)	<u>66.1</u>	75.3	70.4	47.3	59.1	52.6
GDAP (T5-L)	65.6	74.7	69.9	48.0	61.6	<u>54.0</u>

precision and recall. In this dimension, GDAP (T5-L) hits the highest among all generative methods. Its T5-B variant, although yields a slightly lower result, still outperforms TANL and Text2Event when they adopt pre-trained language models at the same scale. When we expand the scope to baselines of all kinds, GDAP (T5-L) ranks 2nd among the 13 approaches benchmarked. Despite it downperforms the state-of-the-art RCEE_ER, we argue that while the latter is a QA-based algorithm that needs templates carefully designed by experts for strong *a priori*, GDAP is fully data-driven and maximally reduces human labour, which is, by all means, more accessible.

To understand the model behaviours in better detail, we additionally report the precision and recall, both of which are missing in many baseline studies. We observe that GDAP (both the T5-B and T5-L versions) achieves record-breaking recall in Argument Extraction. To be concrete, GDAP (T5-L) exceeds the previous state-of-the-art method, Text2Event (which is also a generative model based on T5-L), by a huge margin of 6.4%. This gain is particularly valuable for commercial applications that are intolerant towards signal omissions. On the other side of the coin, we find that the precision of GDAP is relatively weak, though it is still higher than baselines such as Text2Event (T5-B). One possible cause is that, errors via if incorrectly detected event types may propagate to the downstream extraction modules (cf. § 2). We aim to dive deeper into this phenomenon in the upcoming research.

Although the Trigger Extraction results are less important in practice, we still investigate them for further insights. In terms of F1 score, we show that whilst GDAP does not stand out, it still yields performance that is on par with or even better than more complex baselines. As for the results of recall and precision, GDAP (both with T5-B and T5-L) again shoots the best recall among all tests approaches but fails to obtain

Table 2. Results of ablation studies on Argument Extraction. For reference, we duplicate the scores of GDAP in the base setup (cf. Tab. 1), which are highlighted in colour.

(%)	P	R	F1
GDAP (T5-B)	47.3	59.1	52.6
GDAP (T5-L)	48.0	61.6	54.0
<i>+ Golden event types</i>			
RCEE_ER	69.6	68.4	69.0
GDAP (T5-B)	68.6	69.8	69.2
GDAP (T5-L)	69.0	74.2	71.5
<i>- Test samples w/o events</i>			
GDAP (T5-B)	57.0	59.1	58.1
GDAP (T5-L)	58.9	61.6	59.7
<i>- Negative sampling</i>			
GDAP (T5-B)	45.2	56.1	50.1
GDAP (T5-L)	45.4	62.5	52.6

high precision. One interesting finding is, the T5-B version of GDAP, whose scale is smaller, performs better than its T5-L counterpart in all metrics of Trigger Extraction. We will try to uncover the reasons in the future.

We additionally conduct three ablation studies on Argument Extraction, with results exhibited in Tab. 2. To begin with, we provide golden event type annotations to RCEE_ER and GDAP as external signals during inference. It is not surprising that the performance of tested models rises in all aspects. However, we note that contrary to the F1 score comparison in Tab. 1, both the T5-B and T5-L versions of GDAP now outperform RCEE_ER. This justifies our aforesaid assumption that the state-of-the-art RCEE_ER does benefit a lot from manually introduced *a priori*, whereas GDAP may be less precise due to errors in Event Type Detection.

To further demonstrate how event type errors affect model performance, from the test set we remove sentences that are not linked to any event. This adjustment lowers the chance of GDAP being *misled* to predict wrong event types. As expected, the precision of GDAP instantly jumps by around 10%. Lastly, we downgrade our proposed framework by omitting the negative sampling step. Although the overall impact on recall is not substantial, we see a precision drop for both T5-B and T5-L variants, which highlights the usefulness of our negative sampling technique.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel GDAP model that attempts prompt-based learning in the Event Extraction domain for the first time. This simple method also innovatively decouples the generation of triggers and arguments, which is proved to be effective in comprehensive experiments with 11 diverse baselines. In the future, we will continue investigating our empirical observations discussed in § 3.2. Moreover, we plan to explore model weight sharing across different modules, improve the performance (especially the precision) of the GDAP framework, and transfer it to more applications.

5. REFERENCES

- [1] Xiangyu Xi, Wei Ye, Tong Zhang, Quanxiu Wang, Shikun Zhang, Huixing Jiang, and Wei Wu, “Improving event detection by exploiting label hierarchy,” in *Proceedings of the ICASSP*, 2021.
- [2] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda, *ACE 2005 Multilingual Training Corpus*, LDC corpora. Linguistic Data Consortium, 2005.
- [3] Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao, “Extracting events and their relations from texts: A survey on recent research progress and challenges,” *AI Open*, 2020.
- [4] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao, “Event extraction via dynamic multi-pooling convolutional neural networks,” in *Proceedings of the ACL-IJCNLP*, 2015.
- [5] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman, “Joint event extraction via recurrent neural networks,” in *Proceedings of the HLT-NAACL*, 2016.
- [6] Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui, “Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction,” in *Proceedings of the AAAI*, 2018.
- [7] Xiao Liu, Zhunchen Luo, and Heyan Huang, “Jointly multiple events extraction via attention-based graph information aggregation,” in *Proceedings of the EMNLP*, 2018.
- [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 ACL*, 2019.
- [9] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu, “A joint neural model for information extraction with global features,” in *Proceedings of the ACL*, 2020.
- [10] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu, “Event extraction as machine reading comprehension,” in *Proceedings of the EMNLP*, 2020.
- [11] Xinya Du and Claire Cardie, “Event extraction by answering (almost) natural questions,” in *Proceedings of the EMNLP*, 2020.
- [12] Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu, “Event extraction as multi-turn question answering,” in *Findings of the EMNLP*, 2020.
- [13] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto, “Structured prediction as translation between augmented natural languages,” in *Proceedings of the ICLR*, 2021.
- [14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, 2020.
- [15] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen, “Text2event: Controllable sequence-to-structure generation for end-to-end event extraction,” in *Proceedings of the ACL-IJCNLP*, 2021.
- [16] J. Ian Munro and Venkatesh Raman, “Succinct representation of balanced parentheses and static trees,” *SIAM Journal on Computing*, 2001.
- [17] Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu, “Parallel sentence mining by constrained decoding,” in *Proceedings of the ACL*, 2020.
- [18] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni, “Autoregressive entity retrieval,” in *Proceedings of the ICLR*, 2021.
- [19] David Wadden, Ulme Wennberg, Yi Luan, and Hananeh Hajishirzi, “Entity, relation, and event extraction with contextualized span representations,” in *Proceedings of the EMNLP-IJCNLP*, 2019.
- [20] Trung Minh Nguyen and Thien Huu Nguyen, “One for all: Neural joint modeling of entities and events,” in *Proceedings of the AAAI*, 2019.
- [21] Tongtao Zhang, Heng Ji, and Avirup Sil, “Joint entity and event extraction with generative adversarial imitation learning,” *Data Intelligence*, 2019.
- [22] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton, “When does label smoothing help?,” in *Proceedings of the NeurIPS*, 2019.
- [23] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *Proceedings of the ICLR*, 2019.