Probing into the Root: A Dataset for Reason Extraction of Structural Events from Financial Documents

Pei Chen[1]*, **Kang Liu**[2,3],
Yubo Chen[2,3], **Taifeng Wang**[4], and **Jun Zhao**[2,3]
[1] Texas A&M University, College Station, TX
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
[4] Ant Group, Hangzhou, China
chenpei@tamu.edu, taifeng.wang@antgroup.com
{kliu, yubo.chen, jzhao}@nlpr.ia.ac.cn

Abstract

This paper proposes a new task regarding event reason extraction from document-level texts. Unlike the previous causality detection task, we do not assign target events in the text but only provide structural event descriptions, and such settings accord more with practice scenarios. Moreover, we annotate a large dataset **FinReason** for evaluation, which provides **Reason**s annotation for **Fin**ancial events in company announcements. This task is challenging because the cases of multiple-events, multiple-reasons, and implicit-reasons are included. In total, FinReason contains 8,794 documents, 12,861 financial events and 11,006 reason spans. We also provide the performance of existing canonical methods in event extraction and machine reading comprehension on this task. The results show a 7 percentage point F1 score gap between the best model and human performance, and existing methods are far from resolving this problem.

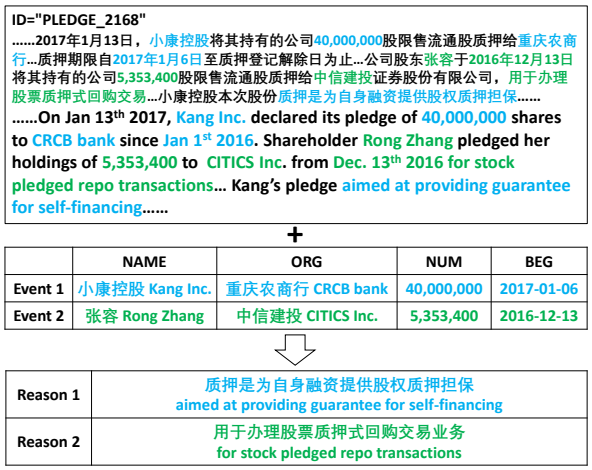

Figure 1: An example of reason extraction for structural events from a document. We need to extract the textual spans from the document as reasons for the given structurally presented events. Here, we need to extract *Reason 1* and *Reason 2* for *Event 1* and *Event 2* respectively.

1 Introduction

Why does the event happen? People are always eager to find the reasons for an event. Automatically extracting the causal explanations of the given events from texts is useful and important for common users and downstream applications. For example, in the financial domain, returning the reasons of a concerned financial event in an Information Retrieval system can free analysts from reading the enormous company announcements and help investors make financial decisions.

Previous work on event causality (Do et al., 2011; Riaz and Girju, 2013; Mirza and Tonelli, 2014; Caselli and Vossen, 2017) mainly focus on the identification of causal relations between two given events that are usually presented as event trigger words. However, in reality, users may only know a particular event happened but without knowing its mention or trigger in the documents, and they just wonder the reasons for it. Therefore, we propose a new task aiming at extracting the causal explanations of the given structurally presented events from document-level texts. Specifically, a ***Structural Event*** defined here is a structural description that contains all necessary roles for an event type. Such a description can completely represent a specific happened event in reality. For example, in Figure 1, the *PLEDGE* event has four predefined roles *NAME*, *ORG*, *NUM*, *BEG* to represent an occurred *PLEDGE* event. Then, our

*Most of the work was done when the first author was a research engineer in the Institute of Automation, CAS.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2042–2048
April 19 - 23, 2021.

task is to extract the reasons of the structural events as textual spans from the document.

To investigate the solution for this challenging task, we construct a large-scale Chinese dataset **FinReason**[1]. Specifically, we automatically collect the formal financial documents with their corresponding structural events the same as Yang et al. (2018). Then, crowd workers are employed to annotate the reasons in the documents for each structural event. In order to guarantee annotation quality and high inter-annotator agreement (IAA), we set several annotation principles and define 3 types of possible causal explanations (*MOTIVATION*, *CAUSE*, *ENABLE*) as ***Reasons*** for the events in company announcements to guide annotators. Finally, there are 8,794 documents, 12,861 collected financial events, and 11,006 reason spans in total. The Cohen's kappa of annotations is 83.87%.

Moreover, to understand this task's difficulties, we regard this task as an Event Extraction (EE) or a Machine Reading Comprehension (MRC) task. We also try some canonical models, such as BiLSRM+CRF (Ma and Hovy, 2016), and BERT (Devlin et al., 2019) on this task and set benchmarks. Empirical results show that this task is challenging, and there is still an overall gap of 7pp (percentage points) in the F1 score between the best model and human performance.

2 Related Work

Much NLP research has focused on identifying causality relations from text, including knowledge bases (WordNet (Miller, 1998), FrameNet (Baker et al., 1998) and ConceptNet (Speer et al., 2017)), semantic related evaluations (SemEval-2007 task 04 (Girju et al., 2007), COPA (Roemmele et al., 2011), RED (Ikuta et al., 2014)), and event-related systems (Beamer and Girju, 2009; Do et al., 2011; Riaz and Girju, 2013; Hu and Walker, 2017; Caselli and Vossen, 2017). These work tried to identify real-world causality in lexicons or texts from different aspects. However, they have found it is difficult to agree on if a causal relationship exists in reality due to the ambiguity of causality definition. Our dataset mitigates this problem by only identifying contextual causality and do not check with reality.

In addition, plenty of work also only identify context-level causal relationships, such as general causality detection tasks PDTB (Prasad et al., 2007) and BECauSE 2.0 (Dunietz et al., 2017), and emotion causality detection task ECA (Lee et al., 2010). Some work (Radinsky et al., 2012; Mirza and Tonelli, 2014; Zhao et al., 2017) also tries to identify the causal relations between events at the contextual-level. However, our task is different because we focus on extracting the reasons for well-defined structural events, which is more close to practice scenarios.

3 Task Description and Data Collection

Task Description Our task is to extract the corresponding causal explanations for given structural events in a document. The inputs are a document with corresponding structural events described in it. The outputs are the causal text spans for the given events. For a given event in the document, there may be zero, single, or multiple causal explanations that need to be identified.

3.1 Data Collection

Event Type	Doc Count	Event Count	Reason Count	Doc Count w/ reason
Pledge	4,138	5,379	4,714	2,901 (70.11%)
O/U	2,550	4,127	3,565	2,132 (83.61%)
Lawsuit	2,106	3,355	2,727	1,438 (68.28%)
Total	8,794	12,861	11,006	6,471 (73.58%)

Table 1: Statistics of FinReason[2].

To construct this dataset, we first collect a corpus of structural events with their corresponding documents following Yang et al. (2018). The collected documents are constrained to company financial announcements, which are relatively formal documents. Such a setting could improve annotation IAA because of the logical consistency and clarity. In specific, we crawl the public company financial announcements as documents from sohu.com[3] and the structural events from eastmoney.com[4]. Since the documents are not in line with their corresponding structural events, we leverage key event items (see more details in Appendix B) matching to align them. Same as Yang et al. (2018), we assume that if the key event items of a structural event appear in a document, the document mentions the target structural event. This alignment method has a high precision of 94.5% as evaluated by Yang et al. (2018).

[1] http://www.nlpr.ia.ac.cn/cip/~liukang/dataset/finreason1.html

[2] The statistics are calculated after manual cleaning in the second step.

[3] http://q.stock.sohu.com/index.shtml

[4] http://choice.eastmoney.com/

In total, as in Table 5, we align 8,794 documents with corresponding 12,861 structural events of 3 types in financial domain, namely Pledge of Shares *(Pledge)*, Overweight and Underweight of Shares *(O/U)*, Lawsuit and Arbitration *(Lawsuit)*. To the best of our knowledge, this is the largest dataset in the event reason extraction task.

3.2 Event Reason Annotation

Annotation Principles: To construct a corresponding dataset with high IAA, we follow the two principles in the annotation. **First**, we annotate the event reasons according to the contextual expressions. We do not check with reality, even if it is obviously a false statement (*e.g., the stock market falls because of intense sunspot activity*). **Second**, we specifically define 3 types of possible causal explanations as reasons for the financial events in announcements following previous work (Trabasso et al., 1989; Van den Broek, 1990; Dunietz et al., 2015): *MOTIVATION*, *CAUSE*, *ENABLE* (see details in Appendix A). This provides a clear guideline to annotators to decide what to annotate and what not to. Because it is also ambitious to differentiate those reason types from texts (Dunietz et al., 2015), we do not require the annotators to distinguish them but just require them to confirm that the reasons annotated at least belong to one of the 3 types.

Quality Control: Besides the aforementioned 2 principles, we adopt several more rules to control data quality as follows. (1) Each member should find as many reasons for a target event as possible. (2) Each reason annotated should be as short as possible but with complete expressivity. (3) When explicit causal relation terms such as 因为 (*because*), 为了 (*in order to*) are mentioned, they should be included in the annotated reasons. (4) For each reason annotated, the annotator should confirm it by doing a *why* test (Grivaz, 2010), which means the reason should answer the question *why the event happened*.

Then, we employ crowdsourcing to annotate the reasons for each event. Specifically, 9 workers are divided into 3 teams to annotate each event type separately. Each team is trained to acquire the domain knowledge of the target event type so they can figure out the possible reasons for the events. Within each team, 2 members are responsible for annotating the reasons independently, and the 3rd member will be activated to make a judgment when 2 annotators have inconsistent annotations. Because the alignment in the first step may not be perfectly accurate, annotators are also responsible for removing those wrongly aligned cases in the annotation to maintain data quality. Finally, as shown in Table 5, there are totally 8,794 documents, 12,861 collected financial events and 11,006 annotated event reason. And, approximately 73.58% of the documents are annotated with event reason. The Cohen's kappa of IAA is 83.87%.

3.3 Task Challenges

Event Type	Multi-event Doc Count	Multi-reason Doc Count	Implicit-reason Doc Count
Pledge	796 (19.24%)	461 (11.14%)	2,845 (68.75%)
O/U	635 (24.90%)	483 (18.94%)	2,030 (79.61%)
Lawsuit	387 (18.38%)	221 (10.49%)	1,434 (68.09%)
Total	1,818 (20.67%)	1,165 (13.25%)	6,309 (71.74%)

Table 2: Three types of challenges in FinReason.

From the annotation results, we could briefly conclude that extracting the reasons for given structural events in a document is not an easy task. First, a document may mention multiple events like the 2 events in the example of Figure 1. As in Table 5, approximately 20.67% documents mention more than one events. Without event mention assignment, discriminating the corresponding reasons for different events within the same document is difficult. Second, about 13.25% of documents mention multiple reasons for an event. Finding all reasons out is also not easy. Thirdly, 71.74% of the documents mention the reason for the events in an implicit way. There are only 28.26% reasons mentioned with explicit modifiers, like 因为(because),由于(since),原因(cause),为(in order to),目的(aims to), etc. Such implicitly mentioned reasons are harder to be identified because they do not have any syntactic clue and require deep reasoning.

We regard the average performance of the two annotators with respect to the final golden standard in the test set as human performance[5]. We can see in Table 2, the human performance on the test set is in line with intuition. Compared with simple cases (Single-Event, Single-Reason, Explicit-Reason), identifying reasons in multiple-event, multiple-reason, and implicit-reason cases are more challenging.

[5]This setting may overestimate the actual human performance but acceptable as a performance upper bound.

4 Benchmark Settings

4.1 Evaluation Criterion

To evaluate the solution on this task, we follow a similar paradigm of SQuAD 2.0 (Rajpurkar et al., 2018) but also with several differences. In general, we get *precision/recall/f1* scores of every event in the test set and calculate the macro-average of all events as the overall performance. However, there are multiple reason cases in FinReason, and we try to evaluate the ability of multiple reasons identification. As a result, we do not fully follow SQuAD-style evaluation by selecting the best prediction but considering all predictions to avoid the systems cheating by predicting all possible causal expressions in the text. For each case, we compute the scores as follows. 1) When there is no reason annotated for an event, the prediction should be *Null* string so as to get *precision/recall/f1* scores of all 1; otherwise, all scores will be assigned 0. 2) When there is only one reason annotated for an event, we calculate the *precision/recall/f1* scores based on the overlapping strings of prediction and ground truth. 3) When there are multiple reasons for a target event, we first calculate each reason's scores with corresponding predictions as in situation 2 and then calculate the macro-average of all reasons as the final scores for the target event.

4.2 Baselines

FinReaon is a new task but similar to several existing tasks, such as event extraction (EE) or machine reading comprehension (MRC). So we apply existing canonical methods for those similar tasks on FinReason as benchmarks for future research. The selected baselines are as followed (see more details in Appendix C):

Regular Expressions (RegExp): In this setting, we regard the FinReason task as a causal sentence detection problem and employ some ad-hoc regular expressions to solve it. Specifically, we use five modifiers (因为(because),由于(since),原因(cause),为(in order to),目的(aims to)) as causal clues to detect the sentence as the reasons for an event.

BiLSTM-CRF (BiLSTM): We can take the reasons as one part of the event description and regard the task as an EE task. Similar to Yang et al. (2018), we employ a BiLSTM-CRF (Ma and Hovy, 2016) to predict the start and end positions of each reason. Specifically, We simply get the event participants in the documents via string matching between the documents and the given structural events. Such information is used as features in a BIO tagging format.

BERT-QA: We can take this task as an MRC problem if the structural event is regarded as a query and the target reason as the answer. In particular, we use templates to turn each structural event into a *why-question* and employ BERT-QA (Devlin et al., 2019) model to find the corresponding reasons.

Type	RegExp	BiLSTM	BERT-QA	Human
Pledge	19/21/20	**76/86/81**	**76**/70/73	93/94/93
O/U	20/27/23	**90/94/92**	**90**/89/89	99/99/99
Lawsuit	20/24/22	**73/73/73**	**73**/72/72	74/78/76
All	20/24/22	**80/84/82**	**80**/77/78	89/90/89

Table 3: Performance of baselines and human beings on FinReason (*precision/recall/f1*, %).

Challenges	RegExp	BiLSTM	BERT-QA	Human
Single-event	16/21/18	**86/90/88**	84/81/82	90/92/91
Multi-event	25/28/26	73/**77/75**	**74**/72/73	87/88/87
Single-reason	23/26/24	85/**86/85**	**86**/84/**85**	91/92/91
Multi-reason	8/13/10	**53/81/64**	43/41/42	76/85/80
Explicit-reason	29/32/30	**85/87/86**	**85**/83/84	90/90/90
Implicit-reason	2/4/3	54/**65/59**	**61**/58/**59**	85/90/87

Table 4: Performance for the three challenges (*precision/recall/f1*, %).

Results We split the dataset into *train/dev/test* sets with a ratio of 8:1:1 for experiments. From the results in Table 2, we can see that there is still an average of 7pp (82% vs. 89%) F1 score gap between the best model (BiLSTM) and human performance. Besides, we can see that human performance is relatively low for *Lawsuit*. This is because the reason for a lawsuit usually lies in a whole story between the plaintiff and the defendant, and it is hard to agree on the boundaries of the span. Furthermore, the BiLSTM model generally performs better than BERT-QA. The reason may be that the BiLSTM model knows the positions of event mentions by using the event BIO features, but BERT-QA only uses the structural event as the query. So it may be easier for BiLSTM to locate the correct reasons.

Besides, we also evaluate the 3 challenges on the whole test set. As from Table 2, the average F1 gap between the best model and human for the 3 challenges are 12pp, 16pp, 28pp, respectively, which is much larger than the overall average gap of 7pp. This demonstrates that the challenges are also the bottlenecks of the models to reach comparable

human performance, especially the implicit cases.

5 Conclusion

In this work, we propose a dataset FinReason for a new event causality extraction task. Our experiments show that this task is still challenging for current models. Future work may consider breaking the challenging cases (multiple-events, multiple-reasons, and implicit-reasons) to achieve a more satisfying performance.

Acknowledgments

This work is supported by the National Key RD Program of China under Grant 2018YFB1005100, the National Natural Science Foundation of China (No.61922085, No.61831022, No.61806201), This work is also supported by Beijing Academy of Artificial Intelligence(BAAI). We would like to thank the anonymous reviewers for their valuable feedback.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

Paul Van den Broek. 1990. The causal inference maker: Towards a process model of inference generation in text comprehension. *Comprehension processes in reading*, pages 423–445.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

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

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18.

Cécile Grivaz. 2010. Human judgements on causation in French texts. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 342–351, Saarbrücken, Germany. Association for Computational Linguistics.

Rei Ikuta, Will Styler, Mariah Hamang, Tim O'Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 12–20.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014,*

the 25th International Conference on Computational Linguistics: Technical Papers, pages 2097–2106.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tom Trabasso, Paul Van den Broek, and So Young Suh. 1989. Logical necessity and transitivity of causal relations in stories. *Discourse processes*, 12(1):1–25.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 335–344. ACM.

Appendices

Appendix A: Event Causality Types

We define the Event Causality Types referring to some previous work. Trabasso et al. (1989) and Van den Broek (1990) defined 4 types of narrative causality relations between events.

- *PHYSICAL*: Event A physically causes event B to happen.
- *MOTIVATIONAL*: Event A happens with B as a motivation.
- *PSYCHOLOGICAL*: Event A brings about emotions expressed in event B.
- *ENABLING*: Event A creates a state or condition for B to happen.

Dunietz et al. (2015) also defined 4 types of causal languages in texts:

- *CONSEQUENCE*: The cause naturally leads to the effect via some chain of events.
- *MOTIVATION*: Some agent perceives the cause, and therefore consciously thinks, feels, or chooses something.
- *PURPOSE*: An agent chooses the effect out of a desire to make the cause true.
- *INFERENCE*: Present the cause as evidence or justification for the effect.

However, we refer to those previous definitions, but we only define 3 types of causal explanations as **Reasons** according to our specific task and application domain. These causality types are the most common reasons for the financial events described in company announcements.

- *MOTIVATION*: The event happens with the explanation as a motivation or purpose. *e.g., He pledged the stocks aiming at providing a guarantee for self-financing.*
- *CAUSE*: The cause in the explanation naturally leads to the occurring of the event. *e.g., He sued the company because of loan disputes.*
- *ENABLE*: The explanation creates a condition or state for the event's occurrence. *e.g., He reduced his shares according to the contract.*

The reasons annotated should at least belong to one of the 3 types. However, we do not require the annotators to distinguish specific types because it is ambitious to differentiate them (Dunietz et al., 2015).

Appendix B: Event-Document Alignment

Following Yang et al. (2018), we assume that if the key items of a structured event appear in a document, then the document mentions the event. We first group those events and documents with the same Announcement Date (DATE[6]) and then use the following key items to align them:

- *Pledge*: Number of Shares (NUM), Name of Shareholder (NAME), Pledge Institution (ORG).

[6]The event item abbreviations are used in the corpus.

- *O/U*: Number of Shares (NUM), Name of Shareholder (NAME), Name of Shares (STOCK).
- *Lawsuit*: Plaintiff (OBG), Defendant (NAME), Court Name (ORG).

Appendix C: Model Settings

We build our baselines based on the open sources of the BiLSTM[7] and BERT-QA[8]. We do not do hyperparameter search and mainly use the default settings. Common parameters are in the Table 5:

Models	Batch Size	Learning Rate	Max Epochs	Optimizer
BiLSTM	20	$1e^{-3}$	30	Adam
BERT-QA	16	$5e^{-5}$	30	BERTAdam

Table 5: Parameters Settings.

Specifically, for BiLSTM, we use the 100-dim character embeddings[9] trained with on Chinese WiKi corpus as initial word features. For structural event embeddings, we first label back the event items to corresponding documents by string matching with BIO schema, then use the 100-dim random vectors as initial BIO features. Besides, the hidden dimension and dropout rate of the LSTM are set as 100 and 0.5, respectively. For BERT-QA, we use the *bert-base-chinese* version and choose the maximum sequence length as 512, the document stride as 128, the max query length as 64, and the max answer length as 30. We train them on two Nvidia GEFORCE GTX 1080Ti GUPs.

Moreover, we need the models to deal with the multiple-event cases. For the RegExp method, we cannot distinguish different events, so we just regard all the extracted reasons as explanations for all the events indiscriminately. For BiLSTM, we create different samples for different events from the same documents to ensure one sample just have one document with at most one event. The BERT-QA regards events as queries so it can naturally adapt to it.

Even for one event, there may be multiple reasons for it, and our models need adaptation. The RegExp is the same as before. The BiLSTM can label multiple pairs of start and end for textual spans, so it naturally adapts to multiple-reason cases. The BERT-QA can return the top k answers as reasons from the documents, and in practice, we get the best result when setting k as 1.

[7] https://github.com/zjy-ucas/ChineseNER.

[8] https://github.com/huggingface/transformers/tree/master/examples/\question-answering.

[9] https://github.com/zjy-ucas/ChineseNER/blob/master/wiki_100.utf8