noises are likely to be introduced. For example, when $k = 1$, *need*, *devices* and *insulation* are enough to express the salient semantic of D2 (working at these temperatures need less insulation). Finally, we select the representation of the root word in the final layer as the discourse-level representation which contains the salient semantic.

The graph convolutional network (GCN) (Kipf and Welling, 2016) is a generalization of convolutional neural network (LeCun et al., 1998) for encoding graphs. In detail, given a syntactic-centric graph with v nodes, we utilize an $v \times v$ adjacency matrix $\boldsymbol{A}$, where $A_{ij} = 1$ if there is an edge between node i and node j. In each layer of GCN, for each node, the input is the output $\boldsymbol{h}_i^{k-1}$ of the previous layer (the input of the first layer is the original encoded input words and features) and the output of node i at k-th layer is $\boldsymbol{h}_i^k$, the formula is as following:

$$\boldsymbol{h}_i^k = \sigma\left(\sum_{j=1}^{v} A_{ij} W^k \boldsymbol{h}_j^{k-1} + b^k\right) \tag{4}$$

where W^k is the matrice of linear transformation, b^k is a bias term and σ is a nonlinear function.

However, naively applying the graph convolution operation in Equation (3) could lead to node representations with drastically different magnitudes because the degree of a token varies a lot. This issue may cause the information in h_i^{k-1} is never carried over to h_i^k because nodes never connect to themselves in a dependency graph (Zhang et al., 2018). In order to resolve the issue that the information in h_i^{k-1} may be never carried over to h_i^k due to the disconnection between nodes in a dependency graph, we utilize the method raised by Zhang (2018) which normalizes the activations in the GCN, and adds self-loops to each node in graph:

$$\boldsymbol{h}_i^k = \sigma\left(\sum_{j=1}^{v} \tilde{A}_{ij} W^k \boldsymbol{h}_j^{k-1} / d_i + b^k\right), \tag{5}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I}$ is the $v \times v$ identify matrix and $d_i = \sum_{j=1}^{v} \tilde{A}_{ij}$ is the degree of word i in graph.

Finally, We select the representation $\boldsymbol{h}_{d_{root}}^k$ of the root word in final layer GCN as the salient representation of d-th discourse in message s. For example, as shown in the subplot (b) of Figure 2, we choose the representation of *need* in the final layer as the salient representation of the discourse *"the devices need less thermal insulation"*.

3.3 Top Discourse-Level Salient-Aware Module

How to make better use of the relation between discourse and extract the message-level salient semantic? We modify the dominance of different discourse based on the message-level constraint in terms of explanatory semantic via an attention mechanism. First, we extract the global semantic of message s which contains its causal explanatory tendency. Next, we modify the dominance of different discourse based on global semantic. Finally, we combine the modified representation to obtain the final causal explanatory representation of input message s.

3.3.1 Global Semantic Extraction

Inspired by previous research (Son et al., 2018), the average encoded word representation of all the words in message can represent its overall semantic simply and effectively. We utilize the average pooling on the encoded representation $\boldsymbol{H}_S^{ed}$ of message s to obtain the global representation which contains the global semantic of its causal explanatory tendency. The formula is as following:

$$\boldsymbol{h}_s^{glo} = \sum_{\boldsymbol{h}_s^{ed} \in \boldsymbol{H}_S^{ed}} \boldsymbol{h}_s^{ed} / n, \tag{6}$$

where $\boldsymbol{h}_s^{glo}$ is the global representation of message s via average pooling operation and n is the number of words.

3.3.2 Dominance Modification

We modify the dominance of different discourse based on the global semantic which contains its causal explanatory tendency via an attention mechanism. In detail, after obtaining the global representation $\boldsymbol{h}_s^{glo}$, we modify the salient representation $\boldsymbol{h}_{droot}^{k}$ of discourses d constrained with $\boldsymbol{h}_s^{glo}$. Finally, we obtain final causal representation $\boldsymbol{h}_s^{caul}$ of message s via attention mechanism:

$$\alpha_{ss} = \boldsymbol{h}_s^{glo}\boldsymbol{W}_f(\boldsymbol{h}_s^{glo})^T \tag{7}$$

$$\alpha_{sd} = \boldsymbol{h}_s^{glo}\boldsymbol{W}_f(\boldsymbol{h}_{droot}^{k})^T \tag{8}$$

$$[\alpha_{ss}^{'}, \cdots, \alpha_{sd}^{'}] = softmax([\alpha_{ss}, ..., \alpha_{sd}]) \tag{9}$$

$$\boldsymbol{h}_s^{caul} = \alpha_{ss}^{'}\boldsymbol{h}_s^{glo} + ... + \alpha_{sd}^{'}\boldsymbol{h}_{droot}^{k}, \tag{10}$$

where the $\boldsymbol{W}_f$ is matrice of linear transformation, $\alpha_{ss}^{'}$, $\alpha_{sd}^{'}$ are the attention weight. Finally, we mapping $\boldsymbol{h}_s^{caul}$ into a binary vector and get the output via a softmax operation.

4 Experiment

Dataset We mainly evaluate our model on a unique dataset devoted to causal explanation analysis released by Son (2018). This dataset contains 3,268 messages consist of 1598 positive messages that contain a causal explanation and 1670 negative sentences randomly selected. Annotators annotate which messages contain causal explanations and which text spans are causal explanations (a discourse with a tendency to interpret something). We utilize the same 80% of the dataset for training, 10% for tuning, and 10% for evaluating as Son (2018). Additionally, to further prove the effectiveness of our proposed model, we regard sentences with causal semantic discourse relations in PDTB2 and sentences containing causal span pairs in BECauSE Corpus 2.0 (Dunietz et al., 2017b) as supplemental messages with causal explanations to evaluate our model. In this paper, PDTB-CED and BECauSE-CED are used to represent the two supplementary datasets respectively.

Parameter Settings We set the length of the sentence and discourse as 100 and 30 respectively. We set the batch size as 5 and the dimension of the output in each GCN layer as 50. Additionally, we utilize the 50-dimension word vector pre-trained with Glove. For optimization, we utilize Adam (Kingma and Ba, 2014) with 0.001 learning rate. We set the maximum training epoch as 100 and adopt an early stop strategy based on the performance of the development set. All the results of different compared and ablated models are the average result of five independent experiments.

Compared Models We compare our proposed model with feature-based and neural-based model: (1) **Lin et al. (2014)**: an end-to-end discourse relation parser on PDTB, (2) **Linear SVM**: a linear designed feature based SVM classifier, (3) **RBF SVM**: a complex designed feature based SVM classifier, (4) **Random Forest**: a random forest classifier which relies on designed features, (5) **Son et al. (2018)**: a hierarchical LSTM sequence model which is designed specifically for CEA. (6) **H-BiLSTM + BERT**[3][4]: a fine-tuned language model (BERT) which has been shown to improve the performance in some other classification tasks based on (5), (7) **H-Atten.**: a well-used Bi-LSTM model that captures hierarchical key information based on hierarchical attention mechanism, (8) **Our model**: our proposed pyramid salient-aware network (PSAN). Furthermore, we evaluate the performance of the model (5), (7), and (8) on the supplemental dataset to prove the effectiveness of our proposed model. Additionally, we design different ablation experiments to demonstrate the effectiveness of the bottom word-level salient-aware module (B-WSM), top discourse-level salient-aware module (T-DSM), and the influence of different depths in the syntactic-centric graph.

[3] https://github.com/huggingface/transformers

[4] BERT can not be applied to the feature-based model suitably, so we deploy BERT on the latest neural network model to make the comparison to prove the effectiveness of our proposed model.

4.1 Main Results

Model	F1	F1	F1
	Facebook	**PDTB-CED**	**BEcuasE-CED**
Lin et al. (2014)	63.8	-	-
Linear SVM (Son et al., 2018)	79.1	-	-
RBF SVM (Son et al., 2018)	77.7	-	-
Random Forest (Son et al., 2018)	77.1	-	-
Son et al. (2018)	75.8	63.6	69.6
H-Atten.	80.9	70.6	76.5
H-BiLSTM + BERT	85.0	-	-
Our model	**86.8**	**76.6**	**81.7**

Table 1: Comparisons of the state-of-the-art methods on causal explanation detection.

Table 1 shows the comparison results on the Facebook dataset and two supplementary datasets. From the results, we have the following observations.

(1) Comparing with the current best feature-based and neural-based models on CED: **Lin et al. (2014)**, **Linear SVM** and **Son et al. (2018)**, **our model** improves the performance by 23.0, 7.7 and 11.0 points on F1, respectively. It illustrates that the pyramid salient-aware network (PSAN) can effectively extract and incorporate the word-level key relation and discourse-level key information in terms of explanatory semantics to detect causal explanation. Furthermore, comparing with the well-used hierarchical key information captured model (**H-Atten.**), **our model** improves the performance by 5.9 points on F1. This confirms the statement in section 1 that directly employing the relation between words with syntactic structure is more effective than the implicit learning.

(2) Comparing the **Son et al. (2018)** with pre-trained language model (**H-BiLSTM+BERT**), there is 9.2 points improvement on F1. It illustrates that the pre-trained language model (LM) can capture some causal explanatory semantics with the large-scale corpus. Furthermore, **our model** can further improve performance by 1.8 points compared with **H-BiLSTM+BERT**. We believe the reason is that the LM is pre-trained with large-scale regular sentences that do not contain causal semantics only, which is not specifically suitable for CED compared to the proposed model for explanatory semantic. Furthermore, the performance of **H-Atten.** is better than **Son et al. (2018)** which indicates focusing on salient keywords and key discourses helps understand explanatory semantics.

(3) It is worth noting that, regardless of our proposed model, comparing the **Linear SVM** with **Son et al. (2018)**, the simple feature classifier is better than the simple deep learning model for CED on the Facebook dataset. However, when combining the syntactic-centric features with deep learning, we could achieve a significant improvement. In other words, our model can effectively combine the *interpretable information* of the feature-based model with the *deep understanding* of the deep learning model.

(4) To further prove the effectiveness of the proposed model, we evaluate **our model** on supplemental messages with causal semantics in other datasets (PDTB-CED and BEcausE-CED). As shown in Table 1, the results show that the proposed model performs significantly better than the **Son et al. (2018)** and **H-Atten.** on the other two datasets[5]. It further demonstrates the effectiveness of our proposed model.

(5) Moreover, **our model** is twice as fast as the **Son et al. (2018)** during training because of the computation of self-attention and GCN is parallel. It illustrates that our model can consume less time and achieve significant improvement in causal explanation detection. Moreover, compared with the feature-based models, the neural-based models rely less on artificial design features.

Dataset	Facebook		PDTB-CED		BEcausE-CED	
Model	**F1**	**∇**	**F1**	**∇**	**F1**	**∇**
our model	**86.8**	-	**76.6**	-	**81.7**	-
w/o B-WSM + root	80.1	-6.7	69.9	-6.7	75.8	-5.9
w/o B-WSM + ave	84.7	-2.1	74.4	-2.2	79.8	-1.9

Table 2: Effectiveness of B-WSM. (**w/o** B-WSM denotes the models without B-WSM. + denotes repalcing the B-WSM with the module after +. **root** denotes using the encoded representation of the root word in each discourse to represent it. **ave** denotes using the average encoded representation of words in discourse to represent it.)

4.2 Effectiveness of Bottom Word-Level Salient-Aware Module (B-WSM)

Table 2 tries to show the effectiveness of the salient information contained in the keywords of each discourse captured via the proposed B-WSM for causal explanation detection (3.2). The results illustrate B-WSM can effectively capture the salient information which contains the most causal explanatory semantics. It is worth noting that when using the average encoded-word representation to represent each discourse (**w/o B-WSM + ave**), the model also achieves acceptable performance. This confirms the conclusion from Son (2018) that the average word representation at word level contains certain causal explanatory semantic. Furthermore, only the root word of each discourse also contains some causal semantics (**w/o B-WSM + root**) which proves the effectiveness of capturing salient information via syntactic dependency from the keywords.

4.3 Effectiveness of Top Discourse-Level Salient-Aware Module (T-DSM)

Dataset	Facebook		PDTB-CED		BEcausE-CED	
Model	**F1**	**∇**	**F1**	**∇**	**F1**	**∇**
our model	**86.8**	-	**76.6**	-	**81.7**	-
w/o T-DSM + seq D	83.8	-3.0	72.9	-3.7	78.1	-3.6
w/o T-DSM + ave S/D	84.0	-2.8	73.5	-3.1	77.8	-3.9

Table 3: Effectiveness of T-DSM. (**w/o** T-DSM denotes models without T-DSM. + denotes replacing the T-DSM with the module after +. **seq D** denotes mapping the representation of discourses via a sequence LSTM to represent the whole message. **ave S/D** denotes using the average encoded representation of words in message and its discourses to represent the whole message.)

Table 3 tries to show the effectiveness of the salient information of the key discourses modified and incorporated via T-DSM for causal explanation detection (3.3). The results compared with **w/o T-DSM + seq D** illustrate our T-DSM can effectively modify the dominance of different discourses based on the global semantic constraint via an attention mechanism to enhance the causal explanatory semantic. Specifically, the results of **w/o T-DSM + ave S/D** show that both discourse-level representation and global representation contain efficient causal explanatory semantics, which further proves the effectiveness of the proposed T-DSM.

4.4 Comparisons of Different Depths of Syntactic-Centric Semantic

To demonstrate the influence of the causal explanatory semantics contained in the syntactic-centric graph with different depths, we further compare the performance of our proposed model with a different number of GCN layers. As shown in Figure 3, when the number of GCN layers is 2, the most efficient syntactic-centric information can be captured for causal explanation detection.

[5] We obtain the performance with the publicly released code by Son et al. (2018). The supplementary datasets are not specifically suitable for this task, and the architectural details of designed feature-based models are not public, so we only compare the performance of the latest model to prove the effectiveness of our proposed model.

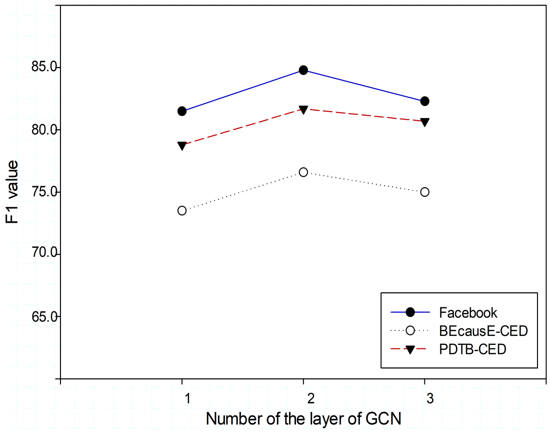
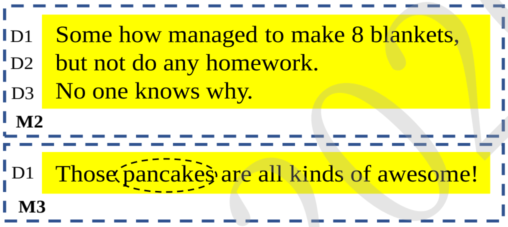


Figure 3: Comparisons of different number of GCN layers.

4.5 Error Analysis

As shown in Figure 4, we find the two main difficulties in this task:



Some how managed to make 8 blankets,
but not do any homework.
No one knows why.

Those pancakes are all kinds of awesome!

Figure 4: Predictions of the proposed model.

(1) **Emotional tendency** The same expression can convey different semantic under different emotional tendencies, especially in this kind of colloquial expressions. As M2 shown in Figure 4, *make 8 blankets* expresses *anger* over *not do any homework*, and our model wrongly predicts the *make 8 blankets* is the reason for *not do any homework*.

(2) **Excessive semantic parsing** Excessive parsing of causal intent by the model will lead to identifying messages that do not contain causal explanations as containing. As shown in Figure 4, M3 means pancakes are awesome, but the model overstates the reason for *awesome* is a pancake.

5 Conclusion

In this paper, we devise a pyramid salient-aware network (PSAN) to detect causal explanations in messages. PSAN can effectively learn the key relation between words at the word level and further filter out the key information at the discourse level in terms of explanatory semantics. Specifically, we propose a bottom word-level salient-aware module to capture the salient semantics of discourses contained in their keywords based on a the syntactic-centric graph. We also propose a top discourse-level salient-aware module to modify the dominance of different discourses in terms of global explanatory semantic constraint via an attention mechanism. Experimental results on the open-accessed commonly used datasets show that our model achieves the best performance.

Acknowledgements

This work is supported by the Natural Key RD Program of China (No.2018YFB1005100), the National Natural Science Foundation of China (No.61533018, No.61922085, No.61976211, No.61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301), CCF-Tencent Open Research Fund and independent research project of National Laboratory of Pattern Recognition.

References

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark, September. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. *arXiv preprint arXiv:1808.06876*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium, October-November. Association for Computational Linguistics.

Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *ACM SIGIR*, pages 400–407. ACM.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017a. Automatically tagging constructions of causation and their slot-fillers. *TACL*, 5:117–133.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017b. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain, April. Association for Computational Linguistics.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *North American Chapter of the Association for Computational Linguistics*, pages 1808–1817, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany, August. Association for Computational Linguistics.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829*.

Yanyan Jia, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao. 2018. Modeling discourse cohesion for discourse parsing via memory network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 438–443, Melbourne, Australia, July. Association for Computational Linguistics.

Daniel Jurafsky. 2010. *Speech and Language Processing: An Introduction to Natural Language*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523.

Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard Hovy. 2015. When are tree structures necessary for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2304–2314, Lisbon, Portugal, September. Association for Computational Linguistics.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas, November. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Empirical Methods in Natural Language Processing*, pages 1506–1515, sep.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298, Saarbrücken, Germany, August. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra-and inter-sentential causal relations. In *Association for Computational Linguistics*, volume 1, pages 1733–1743.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *SIGDIAL*, pages 161–170.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. *arXiv preprint arXiv:1609.08097*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Association for Computational Linguistics*, pages 654–658.

Youngseo Son, Nipun Bayas, and H Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Empirical Methods in Natural Language Processing*.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *AAAI*.

Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. 2019. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3308–3318, Florence, Italy, July. 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. *ArXiv*, abs/1706.03762.

Yizhong Wang, Sujian Li, Jingfeng Yang, Xu Sun, and Houfeng Wang. 2017. Tag-enhanced tree-structured neural networks for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 496–505, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Empirical Methods in Natural Language Processing*, pages 1533–1541, Singapore, August. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Association for Computational Linguistics*, pages 1003–1012, Florence, Italy, July. Association for Computational Linguistics.

Zhang and Xiaojun. 2014. Chengqing zong: Statistical natural language processing (second edition). *Machine Translation*, 28(2):155–158.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Empirical Methods in Natural Language Processing*, pages 35–45.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Empirical Methods in Natural Language Processing*, pages 2205–2215.