Towards Causal Explanation Detection with Pyramid Salient-Aware Network

Xinyu Zuo[1,2], Yubo Chen[1,2], Kang Liu[1,2], Jun Zhao[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2]School of Artificial Intelligence,
University of Chinese Academy of Sciences, Beijing, 100049, China
{xinyu.zuo,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn

Abstract

Causal explanation analysis (CEA) can assist us to understand the reasons behind daily events, which has been found very helpful for understanding the coherence of messages. In this paper, we focus on *Causal Explanation Detection*, an important subtask of causal explanation analysis, which determines whether a causal explanation exists in one message. We design a **P**yramid **S**alient-**A**ware **N**etwork (PSAN) to detect causal explanations on messages. PSAN can assist in causal explanation detection via capturing the salient semantics of discourses contained in their keywords with a bottom graph-based word-level salient network. Furthermore, PSAN can modify the dominance of discourses via a top attention-based discourse-level salient network to enhance explanatory semantics of messages. The experiments on the commonly used dataset of CEA shows that the PSAN outperforms the state-of-the-art method by 1.8% F1 value on the *Causal Explanation Detection* task.

1 Introduction

Causal explanation detection (CED) aims to detect whether there is a causal explanation in a given message (e.g. a group of sentences). Linguistically, there are coherence relations in messages which explain how the meaning of different textual units can combine to jointly build a discourse meaning for the larger unit. The explanation is an important relation of coherence which refers to the textual unit (e.g. discourse) in a message that expresses explanatory coherent semantics (Jurafsky, 2010). As shown in Figure 1, M1 can be divided into three discourses, and D2 is the explanation that expresses the reason why it is advantageous for the equipment to operate at these temperatures. CED is important for tasks that require an understanding of textual expression (Son et al., 2018). For example, for question answering, the answers of questions are most likely to be in a group of sentences that contains causal explanations (Oh et al., 2013). Furthermore, the summarization of event descriptions can be improved by selecting causally motivated sentences (Hidey and McKeown, 2016). Therefore, CED is a problem worthy of further study.

The existing methods mostly regard this task as a classification problem (Son et al., 2018). At present, there are mainly two kinds of methods, feature-based methods and neural-based methods, for similar semantic understanding tasks in discourse granularity, such as opinion sentiment classification and discourse parsing (Nejat et al., 2017; Jia et al., 2018; Soricut and Marcu, 2003). The feature-based methods can extract the feature of the relation between discourses. However, these methods do not deal well with the implicit instances which lack explicit features. For CED, as shown in Figure 1, D2 lacks explicit features such as *because of*, *due to*, or the features of tenses, which are not friendly for feature-based methods. The methods based on neural network are mainly Tree-LSTM model (Wang et al., 2017) and hierarchical Bi-LSTM model (Son et al., 2018). The Tree-LSTM models learn the relations between words to capture the semantics of discourses more accurately but lack further understanding of the semantics between discourses. The hierarchical Bi-LSTM models can employ sequence structure to implicitly learn the relations between words and discourses. However, previous work shows that compared with Tree-LSTM, Bi-LSTM lacks a direct understanding of the dependency relations between words. Therefore, the method of implicit learning of inter-word relations is not prominent in the tasks related
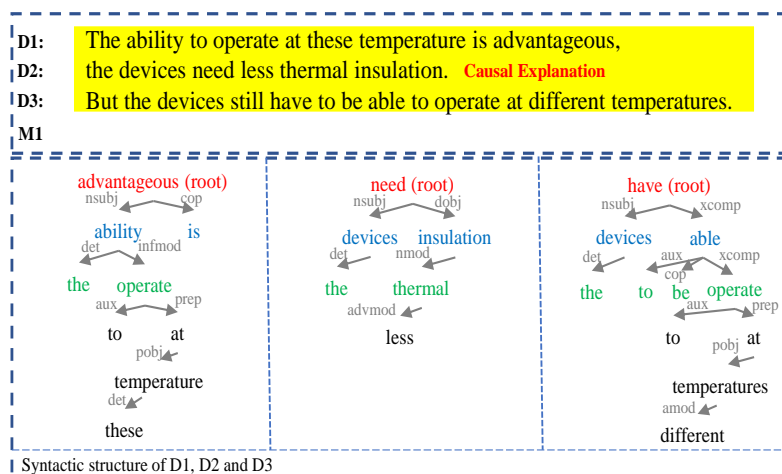
Figure 1: Instance of causal explanation analysis (CEA). The top part is a message which contains its segmented discourses and a causal explanation. The bottom part is the syntactic dependency structures of three discourses divided from M1.

to understanding the semantic relations of messages (Li et al., 2015). Therefore, how to directly learn the relations between words effectively and consider discourse-level correlation to further filter the key information is a valuable point worth studying.

Further analysis, why do the relations between words imply the semantics of the message and its discourses? From the view of computational semantics, the meaning of a text is not only the meaning of words but also the relation, order, and aggregation of the words. In other simple words is that the meaning of a text is partially based on its syntactic structure (Jurafsky, 2010). In detail, in CED, the core and subsidiary words of discourses contain their basic semantics. For example, as D1 shown in Figure 1, according to the word order in syntactic structure, we can capture the *ability* of *temperature* is *advantageous*. We can understand the basic semantic of D1 which expresses some kind of *ability* is *advantageous* via root words *advantageous* and its affiliated words. Additionally, why the correlation and key information at the discourse level are so important to capture the causal explanatory semantics of the message? Through observation, the different discourse has a different status for the explanatory semantics of a message. For example, in M1, combined with D1, D2 expresses the explanatory semantics of *why the ability to work at these temperatures is advantageous*, while D3 expresses the semantic of transition. In detail, D1 and D2 are the keys to the explanatory semantics of M1, and if not treated D1, D2, and D3 differently, the transitional semantic of D3 can affect the understanding of the explanatory semantic of M1. Therefore, how to make better use of the information of keywords in the syntactic structure and pay more attention to the discourses that are key to explanatory semantics is a problem to be solved.

To this end, we propose a **P**yramid **S**alient-**A**ware **N**etworks (PSAN) which utilizes keywords on the syntactic structure of each discourse and focuses on the key discourses that are critical to explanatory semantics to detect causal explanation of messages. First, what are the keywords in a syntactic structure? From the perspective of syntactic dependency, the root word is the central element that dominates other words, while it is not be dominated by any of the other words, all of which are subordinate to the root word (Zhang and Xiaojun, 2014). From that, the root and subsidiary words in the dependency structure are the keywords at the syntax level of each discourse. Specifically, we sample 100 positive sentences from training data to illuminate whether the keywords obtained through the syntactic dependency contain the causal explanatory semantics. And we find that the causal explanatory semantics of more than 80% sentences be captured by keywords in dependency structure[1]. Therefore, we extract the root word and its surrounding words on the syntactic dependency of each discourse as its keywords.

---

[1]Five Ph.D. students majoring in NLP judge whether sentences could be identified as which containing causal explanatory semantics by the root word and its surrounding words in syntactic dependency, and the agreement consistency is 0.8

Next, we need to consider how to make better use of the information of keywords contained in the syntactic structure. To pay more attention to keywords, the common way is using attention mechanisms to increase the attention weight of them. However, this implicitly learned attention is not very interpretable. Inspired by previous researches (Vashishth et al., 2019; Bastings et al., 2017), we propose a bottom graph-based word-level salient network which merges the syntactic dependency to capture the salient semantics of discourses contained in their keywords. Finally, how to consider the correlation at the discourse level and pay more attention to the discourses that are key to the explanatory semantics? Inspired by previous work (Li et al., 2016), we propose a top attention-based discourse-level salient network to focus on the key discourses in terms of explanatory semantics.

In summary, the contributions of this paper are as follows:

- We design a **P**yramid **S**alient-**A**ware **N**etwork (PSAN) to detect causal explanations of messages which can effectively learn the pivotal relations between keywords at word level and further filter the key information at discourse level in terms of explanatory semantics.

- PSAN can assist in causal explanation detection via capturing the salient semantics of discourses contained in their keywords with a bottom graph-based word-level salient network. Furthermore, PSAN can modify the dominance of discourses via a top attention-based discourse-level salient network to enhance explanatory semantics of messages.

- Experimental results on the open-accessed commonly used datasets show that our model achieves the best performance. Our experiments also prove the effectiveness of each module.

2 Related Works

Causal Semantic Detection: Recently, causality detection which detects specific causes and effects and the relations between them has received more attention, such as the researches proposed by Li (Li and Mao, 2019), Zhang (Zhang et al., 2017), Bekoulis (Bekoulis et al., 2018), Do (Do et al., 2011), Riaz (Riaz and Girju, 2014), Dunietz (Dunietz et al., 2017a) and Sharp (Sharp et al., 2016). Specifically, to extract the causal explanation semantics from the messages in a general level, some researches capture the causal semantics in messages from the perspective of discourse structure, such as capturing counterfactual conditionals from a social message with the PDTB discourse relation parsing (Son et al., 2017), a pre-trained model with Rhetorical Structure Theory Discourse Treebank (RSTDT) for exploiting discourse structures on movie reviews (Ji and Smith, 2017), and a two-step interactive hierarchical Bi-LSTM framework (Xia and Ding, 2019) to extract emotion-cause pair in messages. Furthermore, Son (2018) defines the causal explanation analysis task (CEA) to extract causal explanatory semantics in messages and annotates a dataset for other downstream tasks. In this paper, we focus on causal explanation detection (CED) which is the fundamental and important subtask of CEA.

Syntactic Dependency with Graph Network: Syntactic dependency is a vital linguistic feature for natural language processing (NLP). There are some researches employ syntactic dependency such as retrieving question answering passage assisted with syntactic dependency (Cui et al., 2005), mining opinion with syntactic dependency (Wu et al., 2009) and so on. For tasks related to causal semantics extraction from relevant texts, dependency syntactic information may evoke causal relations between discourse units in text (Gao et al., 2019). And recently, there are some researches (Marcheggiani and Titov, 2017; Zhang et al., 2018) convert the syntactic dependency into a graph with graph convolutional network (GCN) (Kipf and Welling, 2016) to effectively capture the syntactic dependency semantics between words in context, such as a semantic role model with GCN (Marcheggiani and Titov, 2017), a GCN-based model assisted with a syntactic dependency to improving relation extraction (Zhang et al., 2018). In this paper, we capture the salient explanatory semantics based on the syntactic-centric graph.

Self-attention Mechanism: Self-attention has been introduced to machine translation by Vaswani (Vaswani et al., 2017) for capturing global dependencies between input and output and achieved state-of-the-art performance. For language understanding tasks, Shen (Shen et al., 2018) exploits self-attention to learn long-range dependencies. Tan (Tan et al., 2018) applies self-attention for semantic role labeling
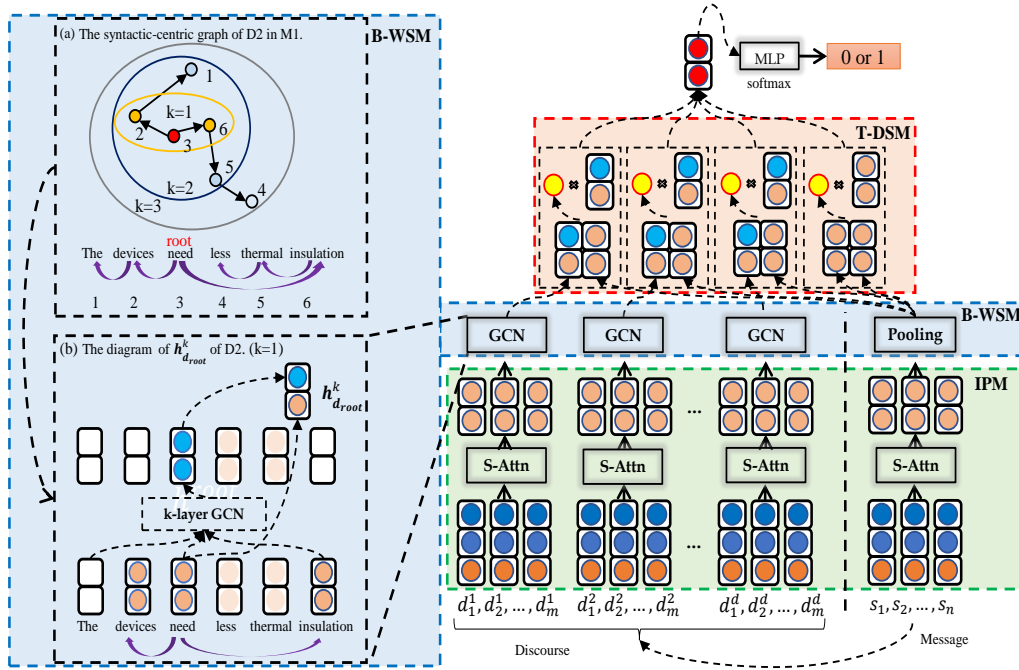
Figure 2: The structure of the pyramid salient-aware network (PSAN). The left side is the detail of the bottom word-level salient-aware module (B-WSM), the top of right side is the top discourse-level salient-aware module (T-DSM) and the bottom of right side is the input processing module (IPM).

task and achieves state-of-the-art results. In this paper, we utilize a multi-head self-attention encoder to capture the representation of words.

3 Methodology

The architecture of our proposed model is illustrated in Figure 2. In this paper, the Pyramid Salient-Aware Network (PSAN) primarily involves the following three components: (i) **input processing module (IPM)**, which processes and encodes the input message and its discourses via self-attention module; (ii) **bottom word-level salient-aware module (B-WSM)**, which captures the salient semantics of discourses contained in their keywords based on the syntactic-centric graph; (iii) **top discourse-level salient-aware module (T-DSM)**, which modifies the dominance of different discourse based on the message-level constraint in terms of explanatory semantic via an attention mechanism, and obtain the final causal explanatory representation of input message m.

3.1 Input Processing Module

In this component, we split the input message m into discourses d. Specially, we utilize the self-attention encoder to encode input messages and their corresponding discourses.

3.1.1 Discourse Extraction

As shown in Figure 1, we split the message into discourses with the same segmentation methods as Son (2018) based on semantic coherence. In detail, first, we regard (','), ('.'), ('!'), ('?') tags and periods as discourse makers. Next, we also extract the discourse connectives set from PDTB2 as discourse makers. Specifically, we remove some simple connectives (e.g. I like running **and** basketball) from extracted discourse marks. Finally, we divide messages into discourses by the discourse makers.

3.1.2 Embedding Layer

For the input message $s = \{s_1, ..., s_n\}$ and discourse $d = \{d_1^d, ..., d_m^d\}$ separated from s, we lookup embedding vector of each word s_n (d_m^d) as $\boldsymbol{s_n}$ $(\boldsymbol{d_m^d})$ from the pre-trained embedding. Finally, we obtain

the word representation sequence $s = \{s_1, ..., s_n\}$ of message s and $d = \{d_1^d, ..., d_m^d\}$ of discourse d corresponding to s.

3.1.3 Word Encoding

Inspired by the application of self-attention to multiple tasks (Tan et al., 2018; Cao et al., 2018), we exploit multi-head self-attention encoder to encode input words. The scaled dot-product attention can be described as follows:

$$(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{\boldsymbol{d}}}\right)\boldsymbol{V} \tag{1}$$

where $\boldsymbol{Q} \in \mathbb{R}^{N \times 2dim_h}$, $\boldsymbol{K} \in \mathbb{R}^{N \times 2dim_h}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times 2dim_h}$ are query matrices, keys matrices and value matrices, respectively. In our setting, $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{s}$ for encoding sentence, and $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{d}$ for encoding discourse.

Multi-head attention first projects the queries, keys, and values h times by using different linear projections. The results of attention are concatenated and once again projected to get the final representation. The formulas are as following:

$$head_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \tag{2}$$

$$\mathbf{H}' = (head_i \oplus \ldots \oplus head_h)\mathbf{W}_o \tag{3}$$

where, $\mathbf{W}_i^Q \in \mathbb{R}^{2dim_h \times dim_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{2dim_h \times dim_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{2dim_h \times dim_k}$ and $\mathbf{W}_o \in \mathbb{R}^{2dim_h \times 2dim_h}$ are projection parameters and $dim_k = 2dim_h/h$. And the output is the encoded message $\boldsymbol{H}_S^{ed} = \{\boldsymbol{h}_{s_1}^{ed}, ..., \boldsymbol{h}_{s_n}^{ed}\}$ and discourse $\boldsymbol{H}_{D^d}^{ed} = \{\boldsymbol{h}_{d_1^d}^{ed}, ..., \boldsymbol{h}_{d_m^d}^{ed}\}$.

3.2 Bottom Word-Level Salient-Aware Module

In this component, we aim to capture the salient semantics of discourses contained in their keywords based on syntactic-centric graphs. For each discourse, first, it extracts the syntactic dependency and constructs the syntactic-centric graph. Second, it collects the keywords and their inter-relations to capture the discourse-level salient semantic based on the syntactic-centric graph.

3.2.1 Syntactic-Centric Graph Construction

We construct a syntactic-centric graph of each discourse based on syntactic dependency to assist in capturing the semantics of discourses. We utilize Stanford CoreNLP tool[2] to extract the syntactic dependency of each discourse and convert them into syntactic-centric graphs. Specifically, in the syntactic-centric graph, the nodes represent words, and the edges represent whether there is a dependency relation between two words or not. As shown in the subplot (a) of Figure 2, *need* is the root word in the syntactic dependency of *"the devices need less thermal insulation"* (D2 in S1), and words which are syntactically dependent on each other are connected with solid lines.

3.2.2 Keywords Collection and Salient Semantic Extraction

For each discourse, we collect the keywords based on the syntactic-centric graph and capture the salient semantic based on the syntactic-centric graph from its keywords. Firstly, as illustrated in section 1, we combine the root word and the affiliated words that connected with the root word in k hops as the keywords. For example, as shown in Figure 2, when $k = 1$, the keywords are {*need, devices and insulation*}, and the keywords are {*need, devices, insulation, the and thermal*} when $k = 2$. Secondly, inspired by previous works, we utilize k-layer graph convolutional network (GCN) (Kipf and Welling, 2016) to encode the k hops connected keywords based on the syntactic-centric graph. For example, when $k = 1$, we encode 1-hop keywords with 1-layer GCN to capture the salient semantic. Specifically, we can capture different degrees of salient semantics by changing the value of k. However, it is not the larger the value of k, the deeper the salient semantics are captured. Conversely, the larger the k, the more

---

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 h_s^{glo}, we modify the salient representation $h_{d_{root}}^k$ of discourses d constrained with h_s^{glo}. Finally, we obtain final causal representation h_s^{caul} of message s via attention mechanism:

$$\alpha_{ss} = h_s^{glo} W_f (h_s^{glo})^T \tag{7}$$

$$\alpha_{sd} = h_s^{glo} W_f (h_{d_{root}}^k)^T \tag{8}$$

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

$$h_s^{caul} = \alpha'_{ss} h_s^{glo} + ... + \alpha'_{sd} h_{d_{root}}^k, \tag{10}$$

where the W_f is matrice of linear transformation, $\alpha'_{ss}, \alpha'_{sd}$ are the attention weight. Finally, we mapping 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**[34]: 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 Facebook	F1 PDTB-CED	F1 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.
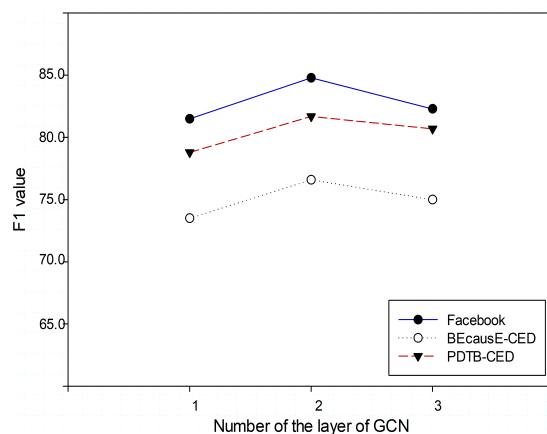
Figure 3: Comparisons of different number of GCN layers.

4.5 Error Analysis

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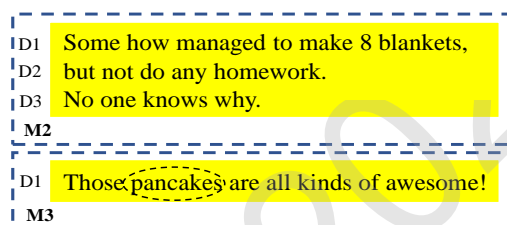

Figure 4: Predictions of the proposed model.

(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

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.