Position Bias Mitigation: A Knowledge-Aware Graph Model for Emotion Cause Extraction

Hanqi Yan, Lin Gui, Gabriele Pergola, Yulan He
Department of Computer Science, University of Warwick
{hanqi.yan, lin.gui, gabriele.pergola, yulan.he}@warwick.ac.uk

Abstract

The Emotion Cause Extraction (ECE) task aims to identify clauses which contain emotion-evoking information for a particular emotion expressed in text. We observe that a widely-used ECE dataset exhibits a bias that the majority of annotated cause clauses are either directly before their associated emotion clauses or are the emotion clauses themselves. Existing models for ECE tend to explore such relative position information and suffer from the dataset bias. To investigate the degree of reliance of existing ECE models on clause relative positions, we propose a novel strategy to generate adversarial examples in which the relative position information is no longer the indicative feature of cause clauses. We test the performance of existing models on such adversarial examples and observe a significant performance drop. To address the dataset bias, we propose a novel graph-based method to explicitly model the emotion triggering paths by leveraging the commonsense knowledge to enhance the semantic dependencies between a candidate clause and an emotion clause. Experimental results show that our proposed approach performs on par with the existing state-of-the-art methods on the original ECE dataset, and is more robust against adversarial attacks compared to existing models.[1]

1 Introduction

Instead of detecting sentiment polarity from text, recent years have seen a surge of research activities that identify the cause of emotions expressed in text (Gui et al., 2017; Cheng et al., 2017a; Rashkin et al., 2018; Xia and Ding, 2019; Kim and Klinger, 2018; Oberländer and Klinger, 2020). In a typical dataset for *Emotion Cause Extract* (ECE) (Gui et al., 2017), a document consists of multiple clauses, one of which is the emotion clause annotated with a pre-defined emotion class label. In addition, one or more clauses are annotated as the cause clause(s) which expresses triggering factors leading to the emotion expressed in the emotion clause. An emotion extraction model trained on the dataset is expected to classify a given clause as a cause clause or not, given the emotion clause.

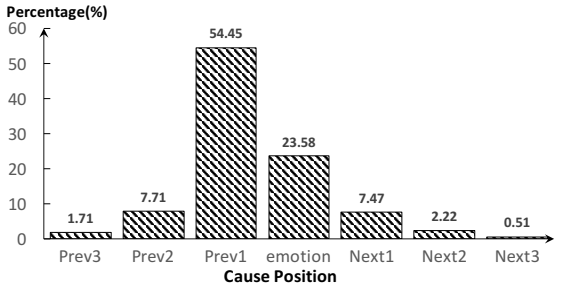

Figure 1: The distribution of positions of cause clauses relative to their corresponding emotion clauses in the ECE dataset (Gui et al., 2016). Nearly 87% of cause clauses are located near the emotion clause (About 55% are immediately preceding the emotion clause, 24% are the emotion clauses themselves and over 7% are immediately after the emotion clause).

However, due to the difficulty in data collection, the ECE datasets were typically constructed by using emotion words as queries to retrieve relevant contexts as candidates for emotion cause annotation, which might lead to a strong positional bias (Ding and Kejriwal, 2020). Figure 1 depicts the distribution of positions of cause clauses relative to the emotion clause in the ECE dataset (Gui et al., 2016). Most cause clauses are either immediately preceding their corresponding emotion clauses or are the emotion clauses themselves. Existing ECE models tend to exploit such relative position information and have achieved good results on emotion cause detection. For example, The Rel-

[1]Our code can be accessed at https://github.com/hanqi-qi/Position-Bias-Mitigation-in-Emotion-Cause-Analysis

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3364–3375
August 1–6, 2021.

ative Position Augmented with Dynamic Global Labels (PAE-DGL) (Ding et al., 2019), RNN-Transformer Hierarchical Network (RTHN) (Xia et al., 2019) and Multi-Attention-based Neural Network (MANN) (Li et al., 2019) all concatenate the relative position embeddings with clause semantic embeddings as the clause representations.

We argue that models utilising clause relative positions would inherently suffer from the dataset bias, and therefore may not generalise well to unseen data when the cause clause is not in proximity to the emotion clause. For example, in a recently released emotion cause dataset, only 25-27% cause clauses are located immediately before the emotion clause (Poria et al., 2020). To investigate the degree of reliance of existing ECE models on clause relative positions, we propose a novel strategy to generate adversarial examples in which the relative position information is no longer the indicative feature of cause clauses. We test the performance of existing models on such adversarial examples and observe a significant performance drop.

To alleviate the position bias problem, we propose to leverage the commonsense knowledge to enhance the semantic dependencies between a candidate clause and the emotion clause. More concretely, we build a clause graph, whose node features are initialised by the clause representations, and has two types of edges i.e., Sequence-Edge (*S-Edge*) and Knowledge-Edge (*K-Edge*). A *S-Edge* links two consecutive clauses to capture the clause neighbourhood information, while a *K-Edge* links a candidate clause with the emotion clause if there exists a knowledge path extracted from the ConceptNet (Speer et al., 2017) between them. We extend Relation-GCNs (Schlichtkrull et al., 2018) to update the graph nodes by gathering information encoded in the two types of edges. Finally, the cause clause is detected by performing node (i.e., clause) classification on the clause graph. In summary, our contributions are three-fold:

- We investigate the bias in the Emotion Cause Extraction (ECE) dataset and propose a novel strategy to generate adversarial examples in which the position of a candidate clause relative to the emotion clause is no longer the indicative feature for cause extraction.
- We develop a new emotion cause extraction approach built on clause graphs in which nodes are clauses and edges linking two nodes capture the neighbourhood information as well as the implicit reasoning paths extracted from a commonsense knowledge base between clauses. Node representations are updated using the extended Relation-GCN.
- Experimental results show that our proposed approach performs on par with the existing state-of-the-art methods on the original ECE dataset, and is more robust when evaluating on the adversarial examples.

2 Related Work

The presented work is closely related to two lines of research in emotion cause extraction: position-insensitive and position-aware models.

Position-insensitive Models. A more traditional line of research exploited structural representations of textual units relying on rule-based systems (Lee et al., 2010) or incorporated commonsense knowledge bases (Gao et al., 2015) for emotion cause extraction. Machine learning methods leveraged text features (Gui et al., 2017) and combined them with multi-kernel Support Vector Machine (SVM) (Xu et al., 2017). More recent works developed neural architectures to generate effective semantic features. Cheng et al. (2017b) employed LSTM models, Gui et al. (2017) made use of memory networks, while Li et al. (2018) devised a Convolutional Neural Network (CNN) with a co-attention mechanism. (Chen et al., 2018) used the emotion classification task to enhance cause extraction results.

Position-aware Models. More recent methodologies have started to explicitly leverage the positions of cause clauses with respect to the emotion clause. A common strategy is to concatenate the clause relative position embedding with the candidate clause representation (Ding et al., 2019; Xia et al., 2019; Li et al., 2019). The Relative Position Augmented with Dynamic Global Labels (PAE-DGL) (Ding et al., 2019) reordered clauses based on their distances from the target emotion clause, and propagated the information of surrounding clauses to the others. Xu et al. (2019) used emotion dependent and independent features to rank clauses and identify the cause. The RNN-Transformer Hierarchical Network (RTHN) (Xia et al., 2019) argued there exist relations between clauses in a document and proposed to classify multiple clauses simultaneously. Li et al. (2019) proposed a Multi-Attention-based Neural Network (MANN) to model the interactions between a candidate clause and the emotion clause.

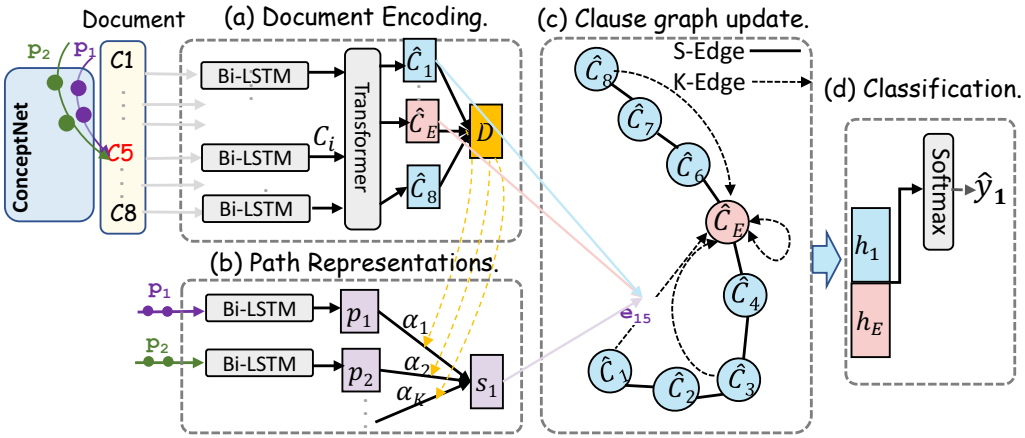

Figure 2: The framework of our proposed KAG. Given an input document consisting of eight clauses ($C1 \cdots C8$), we first extract knowledge paths from ConceptNet between each candidate clause and the emotion clause (§3.1), e.g., two knowledge paths, p_1 and p_2, are extracted between C_1 and the emotion clause C_5. **(a) Document Encoding.** Clauses are fed into a word-level Bi-LSTM and a clause-level Transformer to obtain the clause representations $\hat{\boldsymbol{C}}_i$. The document embedding $\boldsymbol{D}$ is generated by Dot-Attention between the emotion embedding $\hat{\boldsymbol{C}}_E$ and clause embeddings. **(b) Path Representations.** The extracted knowledge paths are fed into Bi-LSTM to derive path representations. Multiple paths between a clause pair are aggregated into $\boldsymbol{s}_i$ based on their attention to the document representation $\boldsymbol{D}$. **(c) Clause Graph Update**. A clause graph is built with the clause representations $\hat{C}_i$ used to initialise the graph nodes. The *K-Edge* weight e_{iE} between a candidate clause $\hat{\boldsymbol{C}}_i$ and the emotion clause $\hat{\boldsymbol{C}}_E$ are measured by their distance along their path $\boldsymbol{s}_i$. **(d) Classification.** Node representation $\mathbf{h}_i$ of a candidate clause C_i is concatenated with the emotion node representation $\mathbf{h}_E$, and then fed to a softmax layer to yield the clause classification result $\hat{\mathbf{y}}_i$.

The generated representations are fed to a CNN layer for emotion cause extraction. The Hierarchical Neural Network (Fan et al., 2019) aimed at narrowing the gap between the prediction distribution p and the true distribution of the cause clause relative positions.

3 Knowledge-Aware Graph (KAG) Model for Emotion Cause Extraction

We first define the Emotion Cause Extraction (ECE) task here. A document $\mathcal{D}$ contains N clauses $\mathcal{D} = \{C_i\}_{i=1}^{N}$, one of which is annotated as an emotion clause C_E with a pre-defined emotion class label, E_w. The ECE task is to identify one or more cause clauses, C_t, $1 \leq t \leq N$, that trigger the emotion expressed in C_E. Note that the emotion clause itself can be a cause clause.

We propose a Knowledge-Aware Graph (KAG) model as shown in Figure 2, which incorporates knowledge paths extracted from ConceptNet for emotion cause extraction. More concretely, for each document, a graph is first constructed by representing each clause in the document as a node. The edge linking two nodes captures the sequential relation between neighbouring clauses (called the *Sequence Edge* or *S-Edge*). In addition, to better capture the semantic relation between a candidate clause and the emotion clause, we identify keywords in the candidate clause which can reach the annotated emotion class label by following the knowledge paths in the ConceptNet. The extracted knowledge paths from ConceptNet are used to enrich the relationship between the candidate clause and the emotion clause and are inserted into the clause graph as the *Knowledge Edge* or *K-Edge*. We argue that by adding the *K-Edges*, we can better model the semantic relations between a candidate clause and the emotion clause, regardless of their relative positional distance.

In what follows, we will first describe how to extract knowledge paths from ConceptNet, then present the incorporation of the knowledge paths into context modelling, and finally discuss the use of Graphical Convolutional Network (GCN) for learning node (or clause) representations and the prediction of the cause clause based on the learned node representations.

3.1 Knowledge Path Extraction from ConceptNet

ConceptNet is a commonsense knowledge graph, which represents entities as nodes and relationship between them as edges. To explore the causal re-

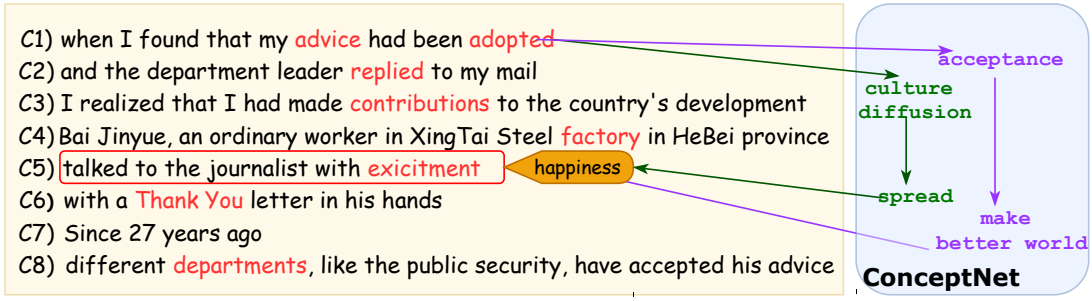

Figure 3: A document consisting of 8 clauses in the ECE dataset with extracted knowledge paths from the ConceptNet. Words in red are identified keywords. '*happiness*' is the emotion label of the emotion clause $C5$. For better visualization, we only display two extracted knowledge paths between '*adopt*' and '*happiness*' in the ConceptNet.

lation between a candidate clause and the emotion clause, we propose to extract cause-related paths linking a word in the candidate clause with the annotated emotion word or the emotion class label, E_w, in the emotion clause. More concretely, for a candidate clause, we first perform word segmentation using the Chinese segmentation tool, Jieba[2], and then extract the top three keywords ranked by Text-Rank[3]. Based on the findings in (Fan et al., 2019) that sentiment descriptions can be relevant to the emotion cause, we also include adjectives in the keywords set.

We regard each keyword in a candidate clause as a *head entity*, e_h, and the emotion word or the emotion class label in the emotion clause as the *tail entity*, e_t. Similar to (Lin et al., 2019), we apply networkx[4] to perform a depth-first search on the ConceptNet to identify the paths which start from e_h and end at e_t, and only keep the paths which contain less than two intermediate entities. This is because shorter paths are more likely to offer reliable reasoning evidence (Xiong et al., 2017). Since not all relations in ConceptNet are related to or indicative of causal relations, we further remove the paths which contain any of these four relations: '*antonym*', '*distinct from*', '*not desires*', and '*not capable of*'. Finally, we order paths by their lengths in an ascending order and choose the top K paths as the result for each candidate-emotion clause pair[5].

An example is shown in Figure 3. The 5-th clause is annotated as the emotion clause and the emotion class label is '*happiness*'. For the keyword, '*adopted*', in the first clause, we show two example paths extracted from ConceptNet, each of which links the word '*adopted*' with '*happiness*'. One such a path is "*adopted* $-$`related_to`$\rightarrow$ *acceptance* $-$`has_subevent`$\rightarrow$ *make better world* $-$`causes`$\rightarrow$ *happiness*".

3.2 Knowledge-Aware Graph (KAG) Model

As shown in Figure 2, there are four components in our model: a document encoding module, a context-aware path representation learning module, a GCN-based graph representation updating module, and finally a softmax layer for cause clause classification.

Initial Clause/Document Representation Learning For each clause C_i, we derive its representation, $\boldsymbol{C}_i$, by using a Bi-LSTM operating on its constituent word vectors, where each word vector $\boldsymbol{w}_i \in \mathbb{R}^d$ is obtained via an embedding layer. To capture the sequential relationship (*S-Edges*) between neighbouring clauses in a document, we feed the clause sequence into a transformer architecture. Similar to the original transformer incorporating the position embedding with the word embedding, we utilise the clause position information to enrich the clause representation. Here, the position embedding $\mathbf{o}_i$ of each clause is concatenated with its representation $\mathbf{C}_i$ generated by Bi-LSTM.

$$\hat{\boldsymbol{C}}_i = \text{Transformer}(\boldsymbol{C}_i \,||\, \boldsymbol{o}_i) \quad (1)$$

We consider different ways for encoding position embeddings using either relative or absolute clause positions and explore their differences in the experiments section. In addition, we will also show the results without using position embeddings at all.

[2]https://github.com/fxsjy/jieba

[3]We have also experimented with other keyword extraction strategies, such as extracting words with higher TFIDF values or keeping all words after removing the stop words. But we did not observe improved emotion cause detection results.

[4]http://networkx.github.io/

[5]We set K to 15, which is the median of the number of paths between all the candidate-emotion clause pairs in our dataset.

Since the aim of our task is to identify the cause clause given an emotion clause, we capture the dependencies between each candidate clause and the emotion clause. Therefore, in the document context modelling, we consider the emotion clause $\hat{\boldsymbol{C}}_E$, generated in a similar way as $\hat{\boldsymbol{C}}_i$, as the query vector, and the candidate clause representation $\hat{\boldsymbol{C}}_i$ as both the key and value vectors, in order to derive the document representation, $\mathbf{D} \in \mathbb{R}^d$.

Context-Aware Path Representation In Section 3.1, we have chosen a maximum of K paths $\{p_t\}_{t=1}^K$ linking each candidate C_i with the emotion clause. However, not every path correlates equally to the document context. Taking the document shown in Figure 3 as an example, the purple knowledge path is more closely related to the document context compared to the green path. As such, we should assign a higher weight to the purple path than the green one. We propose to use the document-level representation $\boldsymbol{D}$ obtained above as the query vector, and a knowledge path as both key and value vectors, in order to calculate the similarity between the knowledge path and the document context. For each pair of a candidate clause C_i and the emotion clause, we then aggregate the K knowledge paths to derive the context-aware path representation $\boldsymbol{s}_i \in \mathbb{R}^d$ below:

$$\boldsymbol{s}_i = \sum_{t=1}^{K} \alpha_t \boldsymbol{p}_t \quad \alpha_t = \text{softmax}(\frac{\boldsymbol{D}^T \boldsymbol{p}_t}{\sum_{j=1}^{K} \boldsymbol{D}^T \boldsymbol{p}_j}) \tag{2}$$

where $\boldsymbol{D}$ is the document representation, $\boldsymbol{p}_t$ is the path representation obtained from Bi-LSTM on a path expressed as an entity-relation word sequence.

Update of Clause Representations by GCN After constructing a clause graph such as the one shown in Figure 2(c), we update the clause/node representations via *S-Edges* and *K-Edges*. Only clauses with valid knowledge paths to the emotion clause are connected with the emotion clause node.

After initialising the node (or clause) in the clause graph with $\hat{\boldsymbol{C}}_i$ and the extracted knowledge path with $\boldsymbol{s}_i$, we update clause representation using an extended version of GCN, i.e. Relation-GCNs (aka. R-GCNs) (Schlichtkrull et al., 2018), which is designed for information aggregation over multiple different edges:

$$\boldsymbol{h}_i^{\ell+1} = \sigma\Big(\sum_{r \in \mathcal{R}_{N_i}} \sum_{j \in N_i} \frac{1}{c_{i,r}} \boldsymbol{W}_r^{\ell} \boldsymbol{h}_j^{\ell} + \boldsymbol{W}_0^{\ell} \boldsymbol{h}_i^{\ell} \Big) \tag{3}$$

where $\boldsymbol{W}_r^{\ell} \boldsymbol{h}_j^{\ell}$ is the linear transformed information from the neighbouring node j with relation r at the ℓ-th layer, $\boldsymbol{W}_r^{\ell} \in \mathbb{R}^{d \times d}$ is relation-specific, $\mathcal{N}_i$ is the set of neighbouring nodes of the i-th node, $\mathcal{R}_{N_j}$ is the set of distinct edges linking the current node and its neighbouring nodes.

When aggregating the neighbouring nodes information along the *K-Edge*, we leverage the path representation $\boldsymbol{s}_i$ to measure the node importance. This idea is inspired by the translation-based models in graph embedding methods (Bordes et al., 2013). Here, if a clause pair contains a possible reasoning process described by the *K-Edge*, then $\hat{\boldsymbol{h}}_E \approx \hat{\boldsymbol{h}}_i + \boldsymbol{s}_i$ holds. Otherwise, $\hat{\boldsymbol{h}}_i + \boldsymbol{s}_i$ should be far away from the emotion clause representation $\hat{\boldsymbol{h}}_E$.[6] Therefore, we measure the importance of graph nodes according to the similarity between $(\boldsymbol{h}_i + \boldsymbol{s}_i)$ and $\boldsymbol{h}_E$. Here, we use the scaled Dot-Attention to calculate the similarity e_{iE} and obtain the updated node representation $\boldsymbol{z}_i$.

$$\boldsymbol{z}_i = \text{softmax}(\boldsymbol{e}_E)\boldsymbol{h}_E^{\ell} \quad e_{iE} = \frac{(\boldsymbol{h}_i + \boldsymbol{s}_i)^T \boldsymbol{h}_E}{\sqrt{d}} (i \neq E) \tag{4}$$

where $\boldsymbol{e}_E$ is $\{e_{iE}\}_{i=1}^{N-1}$. d is the dimension of graph node representations, and $\mathcal{N}^{r_k}$ is a set of neighbours by the *K-Edge*.

Then, we combine the information encoded in *S-Edge* with $\boldsymbol{z}_i$ as in Eq. 3, and perform a non-linear transformation to update the graph node representation $\boldsymbol{h}_i^{\ell+1}$:

$$\boldsymbol{h}_i^{\ell+1} = \sigma\big(\boldsymbol{z}_i^{\ell} + \sum_{j \in N_i^{rs}} (\boldsymbol{W}_j \boldsymbol{h}_j)\big) \tag{5}$$

where N_i^{rs} is a set of i-th neighbours connected by the *S-Edges*.

Cause Clause Detection Finally, we concatenate the candidate clause node h_i and the emotion node representation h_e generated by the graph, and apply a softmax function to yield the predictive class distribution $\hat{y}_i$.

$$\hat{y}_i = \text{softmax}\big(\boldsymbol{W}(\boldsymbol{h}_i^L \,||\, \boldsymbol{h}_E^L) + b\big), \tag{6}$$

4 Experiments

We conduct a thorough experimental assessment of the proposed approach against several state-of-the-art models[7].

[6] Here, we do not consider the cases when the candidate clause is the emotion clause (i.e., $\hat{\boldsymbol{h}}_i = \hat{\boldsymbol{h}}_E$), as the similarity between $\hat{\boldsymbol{h}}_E + \boldsymbol{s}_i$ and $\hat{\boldsymbol{h}}_E$ will be much larger than the other pairs.

[7] Training and hyper-parameter details can be found in Appendix A.

	Methods	P (%)	R (%)	F1 (%)
W/O Pos	RB	67.47	42.87	52.43
	EMOCause	26.72	71.30	38.87
	Ngrams+SVM	42.00	43.75	42.85
	Multi-Kernel	65.88	69.27	67.52
	CNN	62.15	59.44	60.76
	CANN	77.21	68.91	72.66
	Memnet	70.76	68.38	69.55
W. Pos	HCS	73.88	71.54	72.69
	MANN	78.43	75.87	77.06
	LambdaMART	77.20	74.99	76.08
	PAE-DGL	76.19	69.08	72.42
	RTHN	76.97	**76.62**	76.77
Our	KAG	**79.12**	75.81	**77.43**
	: w/o R-GCNs	73.68	72.76	73.14
	: w/o K-Edge	75.67	72.63	74.12
	: w/o S-Edge	76.34	75.46	75.88

Table 1: Results of different models on the ECE dataset. Our model achieves the best Precision and F1 score.

Dataset and Evaluation Metrics The evaluation dataset (Gui et al., 2016) consists of 2,105 documents from SINA city news. As the dataset size is not large, we perform 10-fold cross-validation and report results on three standard metrics, i.e. Precision (P), Recall (R), and F1-Measure, all evaluated at the clause level.

Baselines We compare our model with the position-insensitive and position-aware baselines: **RB** (Lee et al., 2010) and **EMOCause** (Russo et al., 2011) are rules-based methods. **Multi-Kernel** (Gui et al., 2016) and **Ngrams+SVM** (Xu et al., 2017) leverage **S**upport **V**ector **M**achines via different textual feature to train emotion cause classifiers. **CNN** (Kim, 2014) and **CANN** (Li et al., 2018) are vanilla or attention-enhanced approaches. **Memnet** (Gui et al., 2017) uses a deep memory network to re-frame ECE as a question-answering task. Position-aware models use the relative position embedding to enhance the semantic features. **HCS** (Yu et al., 2019) uses separate hierarchical and attention module to obtain context and information. Besides that, **PAE-DGL** (Ding et al., 2019) and **RTHN** (Xia et al., 2019) use similar **G**lobal **P**rediction **E**mbedding (*GPE*) to twist the clauses' first-round predictions. **MANN** (Li et al., 2019) performs multi-head attention in CNN to jointly encode the emotion and candidate clauses. **LambdaMART** (Xu et al., 2019) uses the relative position, word-embedding similarity and topic similarity as emotion-related feature to extract cause.

4.1 Main Results

Table 1 shows the cause clause classification results on the ECE dataset. Two rule-based methods have poor performances, possibly due to their pre-defined rules. Multi-Kernel performs better than the vanilla SVM, being able to leverage more contextual information. Across the other three groups, the precision scores are higher than recall scores, and it is probably due to the unbalanced number of cause clauses (18.36%) and non-cause clauses (81.64%), leading the models to predict a clause as non-cause more often.

Models in the position-aware group perform better than those in the other groups, indicating the importance of position information. Our proposed model outperforms all the other models except RHNN in which its recall score is slightly lower. We have also performed ablation studies by removing either *K-Edge* or *S-Edge*, or both of them (w/o *R-GCNs*). The results show that removing the R-GCNs leads to a drop of nearly 4.3% in F1. Also, both the *K-Edge* and *S-Edge* contributes to emotion cause extraction. As contextual modelling has considered the position information, the removal of *S-Edge* leads to a smaller drop compared to the removal of *K-Edge*.

4.2 Impact of Encoding Clause Position Information

In order to examine the impact of using the clause position information in different models, we replace the relative position information of the candidate clause with absolute positions. In the extreme case, we remove the position information from the models. The results are shown in Figure 4. It can be observed that the best results are achieved using relative positions for all models. Replacing relative positions using either absolution positions or no position at all results in a significant performance drop. In particular, MANN and PAE-DGL have over 50-54% drop in F1. The performance degradation is less significant for RTHN, partly due to its use of the Transformer architecture for context modeling. Nevertheless, we have observed a decrease in F1 score in the range of 20-35%. Our proposed model is less sensitive to the relative positions of candidate clauses. Its robust performance partly attributes to the use of (1) hierarchical contextual modeling via the Transformer structure, and (2) the *K-Egde* which helps explore causal links via commonsense knowledge regardless of a clause's

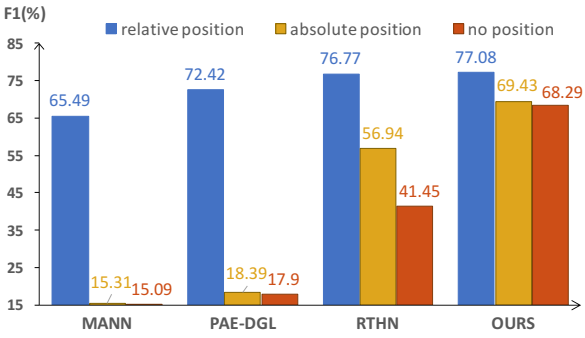

Figure 4: Emotion cause extraction when using relative, absolute or no clause positional information. Our model demonstrates most stable performance without the relative position information.

relative position.

4.3 Performance under Adversarial Samples

In recent years, there have been growing interests in understanding vulnerabilities of NLP systems (Goodfellow et al., 2015; Ebrahimi et al., 2017; Wallace et al., 2019; Jin et al., 2020). Adversarial examples explore regions where the model performs poorly, which could help understanding and improving the model. Our purpose here is to evaluate if KAG is vulnerable as existing ECE models when the cause clauses are not in proximity to the emotion clause.Therefore, we propose a principled way to generate adversarial samples such that the relative position is no longer an indicative feature for the ECE task.

Generation of adversarial examples We generate adversarial examples to trick ECE models, which relies on swapping two clauses C_{r_1} and C_{r_2}, where r_1 denotes the position of the most likely cause clause, while r_2 denotes the position of the least likely cause clause.

We identify r_1 by locating the most likely cause clause based on its relative position with respect to the emotion clause in a document. As illustrated in Figure 1, over half of the cause clauses are immediately before the emotion clause in the dataset. We assume that the position of a cause clause can be modelled by a Gaussian distribution and estimate the mean and variance directly from the data, which are, $\{\mu, \sigma^2\} = \{-1, 0.5445\}$. The position index r_1 can then be sampled from the Gaussian distribution. As the sampled value is continuous, we round the value to its nearest integer:

$$r_1 \leftarrow \lfloor g \rceil, \quad g \backsim \text{Gaussian}(\mu, \sigma^2). \tag{7}$$

To locate the least likely cause clause, we propose to choose the value for r_2 according to the attention score between a candidate clause and the emotion clause. Our intuition is that if the emotion clause has a lower score attended to a candidate clause, then it is less likely to be the cause clause. We use an existing emotion cause extraction model to generate contextual representations and use the Dot-Attention (Luong et al., 2015) to measure the similarity between each candidate clause and the emotion clause. We then select the index i which gives the lowest attention score and assign it to r_2:

$$r_2 = \arg\min_i \{\lambda_i\}_{i=1}^{N}, \quad \lambda_i = \text{Dot-Att.}(\hat{\mathbf{C}}_i, \hat{\boldsymbol{C}}_E), \tag{8}$$

where $\hat{\mathbf{C}}_i$ is the representation of the i-th candidate clause, $\hat{\mathbf{C}}_E$ is the representation of the emotion clause, and N denotes a total of N clauses in a document.

Here, we use existing ECE models as different discriminators to generate different adversarial samples.[8] The desirable adversarial samples will fool the discriminator to predict the inverse label. We use leave-one-model-out to evaluate the performance of ECE models. In particular, one model is used as a *Discriminator* for generating adversarial samples which are subsequently used to evaluate the performance of other models.

Results The results are shown in Table 2. The attacked ECE models are merely trained on the original dataset. The generated adversarial examples are used as the test set only. We can observe a significant performance drop of 23-32% for the existing ECE models, some of which even perform worse than the earlier rule-based methods, showing their sensitivity to the positional bias in the dataset. We also observe the performance degradation of our proposed KAG. But its performance drop is less significant compared to other models. The results verify the effectiveness of capturing the semantic dependencies between a candidate clause and the emotion clause via contextual and commonsense knowledge encoding.

4.4 Case Study and Error Analysis

To understand how KAG aggregate information based on different paths, we randomly choose two examples to visualise the attention distributions (Eq. 4) on different graph nodes (i.e., clauses)

[8]The adversarial sample generation is independent from their training process.

Discriminator	Attacked ECE models			
	PAEDGL	MANN	RTHN	KAG
PAEDGL	49.62 ↓31.76%	48.92 ↓28.6%	59.73 ↓ 22.20%	64.98 ↓ **16.08**%
MANN	51.82 ↓28.45%	47.24 ↓31.27%	60.13 ↓21.65%	66.32 ↓**14.35**%
RTHN	48.63 ↓32.85%	49.63 ↓ 27.64%	57.78 ↓ 24.74%	63.47 ↓**18.03**%
KAG	48.52 ↓ 33.00%	48.24 ↓29.67%	59.53 ↓22.46%	62.39 ↓**19.42**%
Ave. Drop(%)	↓31.51%	↓29.29%	↓22.62%	↓**16.97%**

Table 2: F1 score and relative drop (marked with ↓) of different ECE models on adversarial samples. The listed four ECE models are attacked by the adversarial samples generated from the respective discriminator. Our model shows the minimal drop rate comparing to other listed ECE models across all sets of adversarial samples.

in Figure 5.[9] These attention weights show the '*distance*' between a candidate clause and the emotion clause during the reasoning process. The cause clauses are underlined, and keywords are in bold. C_i in brackets indicate the relative clause position to the emotion clause (which is denoted as C_0).

Ex.1 The **crime** that ten people were **killed** shocked the whole country (C_{-4}). This was due to personal grievances (C_{-3}). Qiu had **arguments** with the management staff (C_{-2}), and thought the Taoist temple host had **molested** his wife (C_{-1}). He became **angry** (C_0), and killed the host and destroyed the temple (C_1).

In Ex.1, the emotion word is '*angry*', the knowledge path identified by our model from ConceptNet is, "*arguments* → *fight* →*angry*" for Clause C_{-2}, and "*molest* → *irritate* →*exasperate*→*angry*" for Clause C_{-1}. Our model assigns the same attention weight to the clauses C_{-2}, C_{-1} and the emotion clause, as shown in Figure 5. This shows that both paths are equally weighted by our model. Due to the *K-Edge* attention weights, our model can correctly identify both C_{-2} and C_{-1} clauses as the cause clauses.

Ex.2 The LongBao Primary school locates between the two villages (C_{-2}). Some **unemployed** people always cut through the school to take a shortcut (C_{-1}). Liu Yurong **worried** that it would affect children's study (C_0). When he did not have teaching duties (C_1), he stood **guard** outside the school gate (C_2).

In Ex.2, the path identified by our model from ConceptNet for Clause (C_{-1}) is "*unemployment* → *situation* → *trouble/danger*→ *worried*". It has been assigned the largest attention weight as shown in Figure 5. Note that the path identified is spurious since the emotion of '*worried*' is triggered by '*unemployment*' in the ConceptNet, while in the original text, '*worried*' is caused by the event, '*Unemployed people cut through the school*'. This shows that simply using keywords or entities searching for knowledge paths from commonsense knowledge bases may lead to spurious knowledge extracted. We will leave the extraction of event-driven commonsense knowledge as future work.

[9] More cases can be found in the Appendix.

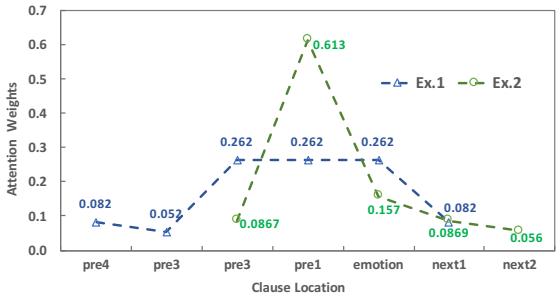

Figure 5: Attention weights among different graph nodes/clauses on **Ex.1** and **Ex.2**.

5 Conclusion and Future Work

In this paper, we examine the positional bias in the annotated ECE dataset and investigate the degree of reliance of the clause position information in existing ECE models. We design a novel approach for generating adversarial samples. Moreover, we propose a graph-based model to enhance the semantic dependencies between a candidate clause and a given emotion clause by extracting relevant knowledge paths from ConceptNet. The experimental results show that our proposed method achieves comparative performance to the state-of-the-art methods, and is more robust against adversarial attacks. Our current model extracts knowledge paths linking two keywords identified in two separate clauses. In the future, we will exploit how to incorporate the event-level commonsense knowledge to improve the performance of emotion cause extraction.

Acknowledgements

This work was funded by the EPSRC (grant no. EP/T017112/1, EP/V048597/1). HY receives the PhD scholarship funded jointly by the University of Warwick and the Chinese Scholarship Council. YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1). We thank Yizhen Jia and

Daoye Zhu for their valuable work on earlier code framework of this paper. We also thank the anonymous reviewers for their valuable comments.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26, NIPS13*, pages 2787–2795.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651, Brussels, Belgium. Association for Computational Linguistics.

Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017a. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17:1–19.

Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017b. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transaction Asian Low-Res. for Lang. Inf. Process.*, 17(1).

Jiayuan Ding and Mayank Kejriwal. 2020. An experimental study of the effects of position bias on emotion causeextraction. *CoRR*, abs/2007.15066.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6343–6350.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. 2019. A knowledge regularized hierarchical approach for emotion cause analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5614–5624, Hong Kong, China. Association for Computational Linguistics.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517 – 4528.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach for emotion cause extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1593–1602.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1639–1649.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

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, Los Angeles, CA. Association for Computational Linguistics.

Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205 – 218.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4752–4757.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 2829–2839.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Laura Oberländer and Roman Klinger. 2020. Sequence labeling vs. clause classification for english emotion stimulus detection. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics (*SEM 2020)*, Barcelona, Spain. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2020. Recognizing emotion cause in conversations. *arXiv preprint arXiv:2012.11820*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 153–160.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne vanden Berg, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*.

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

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1003–1012.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A rnn-transformer hierarchical network for emotion cause extraction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5285–5291.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark. ACL.

B. Xu, H. Lin, Y. Lin, Y. Diao, L. Yang, and K. Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.

Ruifeng Xu, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui. 2017. An ensemble approach for emotion cause detection with event extraction and multikernel svms. *Tsinghua Science and Technology*, 22(6):646–659.

Xinyi Yu, Wenge Rong, Zhuo Zhang, Yuanxin Ouyang, and Zhang Xiong. 2019. Multiple level hierarchical network-based clause selection for emotion cause extraction. *IEEE Access*, 7:9071–9079.

A Model Architecture

In this section, we describe the details of the four main components in our model: *contextual modelling*, *knowledge path encoding*, *clause graph update* and *cause clause classification*.

The dataset has 2,105 documents. The maximum number of clauses in a document is 75 and the maximum number of words per clause is 45. So we first pad the input documents into a matrix $\mathbf{I}$ with the shape of $[2105, 75, 45]$.

A.1 Contextual Modelling

a. **token → clause** We first apply a 1-layer Bi-LSTM of 100 hidden units to obtain word embeddings, $w \in \mathbb{R}^{200}$. We then use two linear transformation layers (hidden units are [200,200],[200,1]) to map the original w to a scalar attention score α, then perform a weighted aggregation to generate the clause representation $\hat{\mathbf{C}}_i \in \mathbb{R}^{200}$.

b. **clause → document** We feed the clause representations into a Transformer. It has 3 stacked blocks, with the multi-head number set to 5, and the dimension of *key*, *value*, *query* is all set to 200. The query vector is the emotion clause representation $\hat{\mathbf{C}}_E \in \mathbb{R}^{200}$, the key and value representations are candidate clause representations, also with 200 dimensions. Finally, the updated clause representations are aggregated via Dot-Attention to generate the document representation $\boldsymbol{D} \in \mathbb{R}^{200}$.

A.2 Knowledge Path Encoding

For each candidate clause and the emotion clause, we extract knowledge paths from ConceptNet and only select K paths. The values of K is set to 15, since the median of the number of paths between a candidate clause and the emotion clause is 15 in our dataset.

We use the same Bi-LSTM described in Section A.1 to encode each knowledge path and generate the K number of path representations $\{\boldsymbol{p}_{it}\}_{t=1}^{K}$ between the i-th clause and the emotion clause. Then, the document representation $\boldsymbol{D}$ is applied as the query to attend to each path in $\{\boldsymbol{p}_{it}\}$ to generate the final context-aware path representation $\boldsymbol{s}_i \in \mathbb{R}^{200}$.

A.3 Clause Graph Update

The graph nodes are initialised by clause presentations, with the feature dimension 200. To calculate the attention weights e_{iE} in R-GCNs, We use the non-linearly transformed $\boldsymbol{h}_i + \boldsymbol{s}_i$ as the query, the non-linearly transformed $\boldsymbol{h}_E$ as the value and key. The non-linear functions are independent Selu layers.

A.4 Cause Clause Classification

The MLP with [400,1] hidden units takes the concatenation of each candidate node $\{\boldsymbol{h}_i^L\}_{i=1}^{N}$ and the emotion node representation $\boldsymbol{h}_E^L$ to predict the logit, after which, a softmax layer is applied to predict the probability of the cause clause.

B Training Details for KAG

We randomly split the datasets into 9:1 (train/test). For each split, we run 50 iterations to get the best model on the validation set, which takes an average time of around 23 minutes per split, when conducted on a NVIDIA GTX 1080Ti. For each split, we test the model on the test set at the end of each iteration and keep the best resulting F1 of the split. The number of model parameters is 1,133,002.

Hyper-parameter Search We use the grid search to find the best parameters for our model on the validation data, and report in the following the hyper-parameter values providing the best performance.

- The word embeddings used to initialise the Bi-LSTM is provided by NLPCC[10]. It was pre-trained on a 1.1 million Chinese Weibo corpora following the Word2Vec algorithm. The word embedding dimension is set to 200.
- The position embedding dimension is set to 50, randomly initialised with the uniform distribution (-0.1,0.1).
- The number of Transformer blocks is 2 and the number of graph layers is 3.
- To regularise against over-fitting, we employ dropout (0.5 in the encoder, 0.2 in the graph layer).
- The network is trained using the the Adam optimiser with a mini-batch size 64 and a learning rate $\eta = 0.005$. The parameters of our model are initialised with Glorot initialisation.

C Error Analysis

We perform error analysis to identify the limitations of the proposed model. In the following examples (Ex.1 and Ex.2), the cause clauses are in bold, our predictions are underlined.

[10]https://github.com/NUSTM/RTHN/tree/master/data

Ex.1 Some kind people said (C_{-6}), if Wu Xiaoli could find available kidneys (C_{-5}), they would like to donate for her surgery (C_{-4}). 4000RMB donation had been sent to Xiaoli (C_{-3}), Qiu Hua said (C_{-2}). The child's desire to survival shocked us (C_{-1}). **The family's companion was touching (C_0)**. Wish kind people will be ready to give a helping hand (C_1). Help the family in difficulty (C_2).

In the first example **Ex.1**, our model identifies the keyword ***survival*** in C_{-1} and extracts several paths from '*survival*' to '*touching*'. However, the main event in clause C_{-1} concerns ***desire*** rather than ***survival***. Our current model detects the emotion reasoning process from ConceptNet based on keywords identified in text, and inevitably introduces spurious knowledge paths to model learning.

Ex.2 I have only one daughter (C_0), and a granddaughter of 8 year-old (C_{-10}). I would like to convey these memory to her (C_{-9}). Last Spring Festival (C_{-8}), I gave the DVD away to my granddaughter (C_{-7}). I hope she can inherit my memory (C_{-6}). Thus (C_{-5}), I feel like that my ages become eternity (C_{-4}). Sun Qing said (C_{-3}). His father is a sensitive and has great passion for his life (C_{-2}). **He did so (C_{-1})**. Making me feel touched (C_0). His daughter said (C_1).

In the **Ex** 2, our model detected the ***passion*** as a keyword and extracted knowledge paths between the clause C_{-2} and the emotion clause. However, it ignores the semantic dependency between the clause C_{-1} and the emotion clause. It is therefore more desirable to consider semantic dependencies or discourse relations between clauses/sentences for emotion reasoning path extraction from external commonsense knowledge sources.

D Human Evaluation on the Generated Adversarial Samples

The way adversarial examples generated changes the order of the original document clauses. Therefore, we would like to find out if such clause reordering changes the original semantic meaning and if these adversarial samples can be used to evaluate on the same emotion cause labels.

We randomly selected 100 adversarial examples and ask two independent annotators to manually annotate emotion cause clauses based on the same annotation scheme of the ECE dataset. Compared to the original annotations, Annotator 1 achieved 0.954 agreement with the cohen's kappa value of 0.79, while Annotator 2 achieved 0.938 agreement with the cohen's kappa value of 0.72. This aligns with our intuition that an emotion expressed in text is triggered by a certain event, rather than determined by relative clause positions. A good ECE model should be able to learn a correlation between an event and its associated emotion. This also motivates our proposal of a knowledge-aware model which leverages commonsense knowledge to explicitly capture event-emotion relationships.