From Coarse to Fine: A Distillation Method for Fine-Grained Emotion-Causal Span Pair Extraction in Conversation

Xinhao Chen[1,2], Chong Yang[2], Changzhi Sun[3], Man Lan[1,4*], Aimin Zhou[1]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, P.R. China
[2]AntGroup, Shanghai, P.R. China
[3]Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai, P.R. China
[4]Shanghai Institute of AI for Education, East China Normal University, Shanghai, P.R. China
xinhaochen@stu.ecnu.edu.cn, yangchong.yang@antgroup.com,
changzhisun@stu.ecnu.edu.cn, {mlan,amzhou}@cs.ecnu.edu.cn

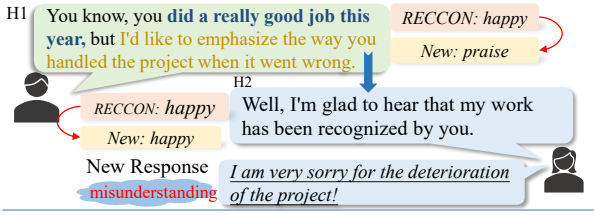
Abstract

We study the problem of extracting emotions and the causes behind these emotions in conversations. Existing methods either tackle them separately or jointly model them at the coarse-grained level of emotions (fewer emotion categories) and causes (utterance-level causes). In this work, we aim to jointly extract more fine-grained emotions and causes. We construct a fine-grained dataset FG-RECCON, includes 16 fine-grained emotion categories and span-level causes. To further improve the fine-grained extraction performance, we propose to utilize the casual discourse knowledge in a knowledge distillation way. Specifically, the teacher model learns to predict causal connective words between utterances, and then guides the student model in identifying both the fine-grained emotion labels and causal spans. Experimental results demonstrate that our distillation method achieves the state-of-the-art performance on both RECCON and FG-RECCON dataset.

Introduction

The task of **E**motion-**C**ausal **S**pan **P**air **E**xtraction (**ECSPE**) in conversations aims to recognize the emotions expressed by speakers in a dialogue and identify the causal spans (i.e., emotion cause) for non-neutral utterances. As illustrated in Figure 1(I), the speaker demonstrates an emotion (denoted as *happy*) in *H2* in the dialogue, the cause of which is highlighted in *H1*. The **ECSPE** task is essential for many downstream tasks, such as empathy generation (Kim and Kim 2021) and emotional support (Liu et al. 2021b).

Existing works present various frameworks for emotion recognition (Shen et al. 2021; Ghosal et al. 2019; Zhu et al. 2021) and cause reasoning (Poria et al. 2021), most of which can be grouped into two categories according to the task objectives, as shown in Fig 1: (a) **Cause identification only**: such as RECCON (Poria et al. 2021) and KEC (Li et al. 2022). These approachs required the construction of an emotion recognition module in a pipeline for real-world scenarios, which suffers from error propagation. (b) **Joint identification of emotions and causes**: such as



Methods	CG-Emotion	FG-Emotion	Causal Utterance	Causal Span
RECCON (Poria et al. 2022), KEC (Li et al. 2022)	—	—	√	—
MGSAG (Bao et al. 2022), UECA (Zheng et al. 2022), ECPE-2D (Ding et al. 2022)	√	—	√	—
GoEmotions (Dem et al. 2022)	√	√	—	—
Ours	√	√	√	√

Figure 1: Illustration of the **ECSPE** task.including a conversation example from the RECCON, and a comparison of the focus of our work with that of past studies.

MGSAG (Bao et al. 2022), UECA (Zheng et al. 2022) and ECPE-2D (Ding, Xia, and Yu 2020), etc. These approachs performed coarse-grained **E**motion-**C**ause **P**air **E**xtraction (**ECPE**), which have faced two major challenges. On the one hand, previous works only focused on the six basic emotions defined by Emkan (Ekman 1992). Utterances with nuanced emotions beyond these six categories defy straightforward classification. For example in Figure 1, the genuine emotion for *H1* is *praise* but it has to be labeled as *happy*. Although GoEmotions (Demszky et al. 2020) have proposed fine-grained datasets for comments, the lack of a conversation reasoning dataset with fine-grained emotions hinders the progress. On the other hand, most of previous work (Li et al. 2022; Zheng et al. 2022) mainly focused on utterance-level cause recognition. However, informal dialogue utterances may contain multiple clauses with different emotions. Not all clauses within the utterance contribute equally to the reasoning of the emotion, and some may even lead to confusion and misunderstanding as shown in Figure 1(I)(Gao et al. 2021). Therefore, identifying causal spans within utterances in dialogues is essential. Overall, there is a progres-

*Man Lan is the corresponding author

sive relationship between **ECSPE** and **ECPE** tasks, so the key focus is improving from coarse-grained to fine-grained in both emotion and causal span recognition.

Based on the above challenges, existing dialogue benchmark dataset RECCON (Poria et al. 2021) for emotion cause identification is inadequate. RECCON only contains 6 basic emotion labels (*Surprise, Anger, Disgust, Happiness, Sadness* and *Fear*), which is not capable to reveal the emotional details of the interlocutor in the conversation. And, the data distribution is extremely imbalanced, where 47% of the utterances are labelled as *neutral*, and 38.3% of the utterances are labelled as *happy*. To better evaluate the performance on **ECSPE** task, we make additional modifications and annotations to the RECCON dataset and propose a new Fine-Grained RECCON (FG-RECCON) dataset. Inspired by (Zhou, Lan, and Wu 2018), we use 16 emotion labels commonly appearing in daily dialogues, such as expanding *happiness* into granular labels like *gratitude, like, praise* etc. After annotation, 23.21% of neutral utterances are labelled with fine-grained and accurate emotions, and $1,217$ new pairs of emotion and causal spans are added.

To address the problem of analyzing emotional causes from the coarse to the fine, we propose a novel **K**nowledge **D**istillation method to identify fine-grained **E**motion-**C**ausal **S**pan pairs (**KD-ECS**). Knowledge distillation (Hinton, Vinyals, and Dean 2015) enhances the performance of newer models by transferring insights from a comprehensive teacher model to a specialized student model. Past methods (Li et al. 2022) which identify utterance-level causes in RECCON, are capable of leveraging a large volume of existing data to acquire abundant coarse-grained knowledge. By knowledge distillation, we leverage the soft outputs (probability distributions) from a teacher model trained on ECPE task to guide the student model in filtering non-causal utterance pairs, thereby refining the search space for causal spans in the dialogue. Previous work (Pitler and Nenkova 2009) has shown that there is a strong correlation between connectives and discourse relations. Inspired by (Zhou et al. 2022), we first train a teacher model to identify causal relationships between utterances, which uses the connective word prediction method with utterance-level emotion causal relations as supervision. Given the consistency between utterance-level and span-level causal relations, the student model can improve its fine-grained emotion recognition and causal span identification when guided by the inter-utterance causal relation knowledge from the teacher model. Both models are prompt-learning-based and fine-tuned on **p**retrained **l**anguage **m**odels (**PLMs**) to gain better performance. Additionally, as only the student model is used to make inference in distillation methods, we can try large language models with different size in the teacher model, e.g., T5 (Raffel et al. 2020), LLaMA (Touvron et al. 2023). Experimental results prove that larger language models indeed provide more powerful causal relation reasoning ability. Our code and dataset are released in https://github.com/cubenlp/KD-ECS.

In summary, our contributions are as follows:

- We present a high-quality and fine-grained benchmark corpus FG-RECCON for fine-grained emotion-causal span pair extraction in conversation scenario with the aid of ChatGPT and manual annotation.
- We propose a novel distillation method **KD-ECS**, using coarse-grained causal reasoning objective to enhance the fine-grained emotion and causal span extraction in conversation.
- We conduct extensive experiments to demonstrate the superiority of our proposed method and explore the causal reasoning ability of LLMs with different size.

Related Work

Emotion and Cause Recognition in Conversation

The **ECSPE** task consists of two subtasks: emotion recognition (Shen et al. 2021; Ghosal et al. 2019; Zhu et al. 2021) and causal span recognition (Poria et al. 2021). These two subtasks can be combined in a "pipeline" format. (Shen et al. 2021; Hu, Wei, and Huai 2021) focusd on improving the modeling of conversation history to recognize emotion. (Poria et al. 2021) provided the RECCON dataset for causal span recognition. (Poria et al. 2021) treated the span-level cause identification as a machine reading comprehension task. (Li et al. 2022) proposed a knowledge enhanced conversation graph to recognize the causal utterances when emotion was given. Different from these work, we jointly identify the emotion and causal span pairs.

Other works jointly extract emotions and causes in the ECPE task (Xia and Ding 2019) on document-level datasets (Gao et al. 2017; Gui et al. 2014), which was initially proposed as an utterance pairing task. MGSAG (Bao et al. 2022) proposed to incorporate fine-grained and coarse-grained semantic features jointly without regard to distance limitation. UECA (Zheng et al. 2022) designed prompt templates to predict *is/isn't* in a Question Answering format. These methods all performed coarse-grained utterance recognition. In contrast to these methods, we employ knowledge distillation to learn coarse-grained causal relationships between utterances, aiming to identify finer-grained causal spans.

Prompt Learning and Knowledge Distillation

Prompt learning methods have been proved to extract knowledge from language models (Liu et al. 2021a; Schick and Schütze 2020). (Zhou et al. 2022) used prompt learning to predict connective words for implict discourse relationship recognition, effectively improving the model's ability to infer various relationships between utterances.

Knowledge distillation was proposed by (Hinton, Vinyals, and Dean 2015) for transferring knowledge from teacher models to sstudent models to improve performance. The architecture of teacher-student models has also been used as a special form of transfer learning for domain migration (Choi, Choi, and Lee 2022). Recently, Large Language Models (LLMs) have shown excellent performance in generalization across various tasks (Bubeck et al. 2023). In order to improve the performance of models in specific domains, many research works have focused on distilling the knowledge of teacher LLMs into student models. (Jiang et al. 2023) proposed a method for transferring knowledge from a complex, closed-source large language model to a smaller language model, which achieved promising results.

Problem Formulation

Let $\{u_1, u_2, \cdots, u_n\}$ be a series of utterances in a conversation, where n is the total number of utterances. The speaker id of utterance u_i is denoted by sp_i. Each utterance u_i is paired with the utterance $u_j (j < i)$ from the conversational history $his(u_i)$, which is called a candidate pair. The ECSPE task is to recognize the emotion emo_u_i of u_i and extract the specific span from the candidate pairs where emo_u_i is not *Neutral* and u_j is corresponding cause utterance of u_i. Specifically, the causal span of u_i is formalized as the start and end positions (P_{st}, P_{ed}) of the maximum causal substring in the utterance.

Corpus Construction

This section describes the process of annotating the RECCON dataset with additional fine-grained utterances and causes, and then obtaining the FG−RECCON dataset.

Annotation Process

For each dialogue in the dataset, we use a combination of ChatGPT[1] automatic labeling and manual modifications to to enhance emotion granularity and annotate causal spans.

Fine-grained Emotion. Following (Zhou, Lan, and Wu 2018), we select 16 labels with high frequency of occurrence in the dataset. First, we construct the following natural language template and automatically annotated emotions through chatGPT:" *Identify the emotion in each turn of the following dialogue. The emotion labels can only be selected from the following: sadness, ... , praise, others. The dialogue is as follows:...* ". In the validation set, ChatGPT achieves a fine-grained emotion classification accuracy of 69.3%.

To improve annotation accuracy, three experts conduct a manual review. After training on the criteria, they sample and discusse 100 dialogues to align their understanding. They annotate the entire dataset independently, achieving an inter-annotator agreement with a kappa coefficient of 76.9%. Discrepancies are resolved through consensus voting.

Causal Spans. We add extra spans for utterances that were previously labeled *neutral* but now labeled fine-grained emotion. We follow the annotation from RECCON (Poria et al. 2021) to formulate the annotation specifications for causal spans as shown in Figure 1(I). To aggregate the causal spans, we take the union of the candidate spans from different annotators as the final causal span only when the size of their intersection is at least 50% of the size of the shortest candidate span. Otherwise, a third annotator was brought in to determine the final span from the existing spans. We add $1,217$ new utterance–causal span (UCS) pairs of emotion and causal spans and corrected 21 wrong pairings on the original dataset. For a detailed overview of our causal span annotation process and the proposed dataset, please refer to our open-source repository.

Data Statistics and Analysis

We present an overview of our FG−RECCON dataset. The fine-grained emotion distribution after annotation is shown in the Table 1. After annotation, 57.56% of utterances which are originally labelled *happiness* are further subdivided into positive emotions such as *like* and *gratitude*, and 23.21% of utterances which are originally labelled *neutral* are further subdivided into fine-grained emotions. Table 2 shows the statistical information of the datasets. Average utterance contains 3.6 phrase segments, with causal span length accounting for 31% of causal utterance.

FG-RECCON			
Emotions	count	Emotions	count
Sadness	$342(3.0\%)$	*Gratitude*	$587(5.3\%)$
Worry	$336(2.9\%)$	*Acceptance*	$318(2.9\%)$
Anger	$371(3.3\%)$	*Like*	$768(6.9\%)$
Disapproval	$389(3.4\%)$	*Curiosity*	$575(5.2\%)$
Disgust	$137(1.2\%)$	*Neutral*	$4,026(36.3\%)$
Happiness	$1,985(17.5\%)$	*Surprise*	$491(4.4\%)$
Sympathy	$238(2.1\%)$	*Praise*	$310(2.8\%)$
Hope	$404(3.6\%)$	*Others*	$36(0.3\%)$

Table 1: The distribution of fine-grained emotion labels.

Dataset	Elements	Train	Val	Test
RECCON	Positive UCS pairs	7,269	341	1,894
	Negative UCS pairs	20,648	838	5,330
FG−RECCON	Positive UCS pairs	8,186	392	2,132
	Negative UCS pairs	20,646	838	5,330
	Num. of Dialogues	834	47	225
	Num. of Utterances	8,206	493	2,405

Table 2: Statistics of the FG−RECCON and RECCON dataset.

Method

In this section, we provide a detailed introduction to our **KD-ECS** method. Considering the consistency of the utterance-level and span-level causal relations, we use the method of knowledge distillation (Hinton, Vinyals, and Dean 2015) to narrow the scope of fine-grained identification. We first present the teacher and the student model for emotion and causal span prediction and then describe our framework of using knowledge distillation at the coarse-grained level. The **KD-ECS** model is illustrated in Figure 2.

Teacher Model

The aim of the teacher model is to learn utterance-level causal reasoning, which benefits subsequent span recognition by reducing the search scope.

Inspired by (Zhou et al. 2022),we view the **ECPE** task as identifying discourse causal relationships between utterances by predicting causal connectives. Therefore, we employ a prompt learning approach for the teacher model to bridge the gap between connective prediction in the pre-training and fine-tuning stage and make better use of the knowledge of **PLMs**.

Prompt Template. Given two utterances u_i, u_j in the dialogue and the emotion labels emo_h_i, emo_h_j, we use

[1]https://platform.openai.com/docs/models/gpt-3-5

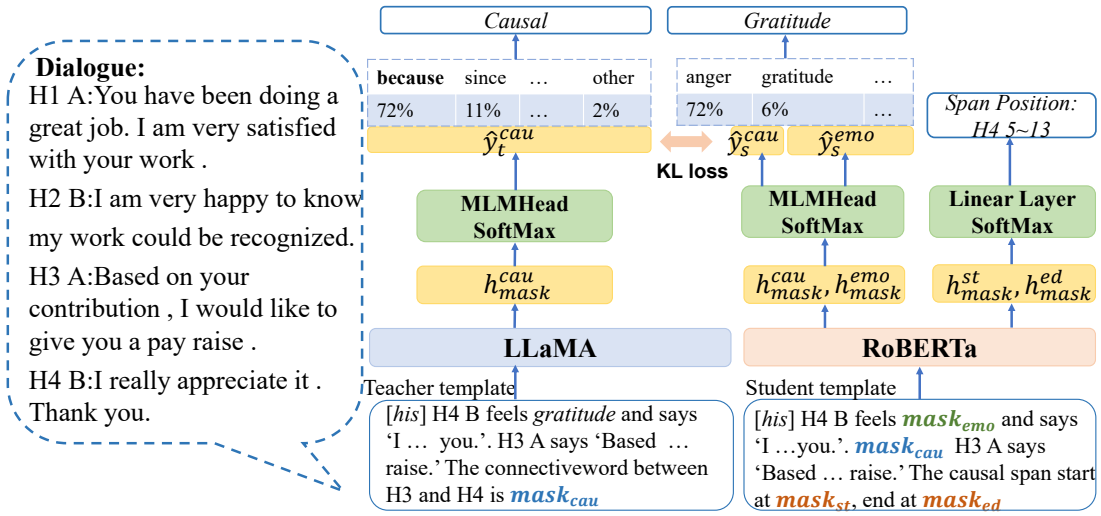


Figure 2: The architecture of our KD-ECS model, which contains two modules: (1) A teacher model for utterance-level causality recognition and (2) a student model for joint recognition of emotion and causal spans.

the following prompt template to identify whether it is an emotion-cause pair:

$$t_{prompt} = \text{"}[his]\ d_i\ sp_i \text{ feels } emo\_u_i \text{ and says } u_i.\ d_j\ sp_j \text{ says } u_j.\text{The connective word between } d_i \text{ and } d_j \text{ is } \langle mask_{cau} \rangle\ .\text{"} \quad (1)$$

$[his]$ represents the dialogue history, which is joint by historical utterances with the special token $[sep]$. d_i is the dialogue turn id. $mask_{cau}$ represents the masked causal token that needs to be predicted. Then, we input the prompt t_{prompt} to a language model, e.g., LLaMA (Touvron et al. 2023) to obtain the representation $h^{cau}_{mask\_t}$ for $mask_{cause}$.

Connective Word Prediction. The objective of the teacher model is to predict whether the connective words between two utterances in the dialogue. Thus, we use the MLMHead layer to map $h^c_{mask\_t}$ to the tokens in the vocabulary:

$$e^{cau}_{mask\_t} = MLMHead(h^{cau}_{mask\_t}), \quad (2)$$

Where $e^{cau}_{mask\_t} \in \mathbb{R}^{|V_c|}$, and $|V_c|$ represents the vocabulary size. During training, a softmax layer is applied to normalize the probability logits:

$$\hat{y}^{cau}_{ti} = Softmax(e^{cau}_{ti}). \quad (3)$$

We train the teacher model using cross-entropy to calculate the loss between the teacher model's predicted results and the selected golden connective words.

$$\mathcal{L}^{cau}_t = -\frac{1}{\mathcal{M}} \sum_{i \in \mathcal{M}} y_i log \hat{y}^{cau}_{ti}, \quad (4)$$

where $\mathcal{M}$ denotes the set of masked tokens and y_i represents the golden label. We use LLAMA as the teacher model and utilize LORA (Hu et al. 2021) to fine-tune it.

Answer Mapping. In prompt learning, we establish a mapping between the retrieved answer and its associated output (Liu et al. 2021a). We manually select answer words to construct emotion and causal discrete answer space. For fine-grained emotion prediction in the student model, we directly use each type of emotion word as answer words. For the prediction of connective words, we choose the most frequent and less ambiguous connective words, which are masked as a single token to unify the inputs and outputs of the model and to facilitate the mapping between words and causal labels. Specific word choices are shown in Table 3.

Task	Relation	Target words
Connectives prediction	Causal	*because,as, since,thus.*
	Non-causal	*other*
Emotion recognition		*sadness,worry,anger,hope, disgust,happiness,sympathy, like,gratitude,acceptance, disapproval,curiosity,neutral, surprise,praise,others,*

Table 3: Settings of answer words.

Student Model

The goal of the student model is to perform emotion recognition and causal span recognition. And meanwhile, learns the inter-utterance causal relation knowledge of the teacher model through knowledge distillation.

Prompt Template. Different from the prompt template of the teacher model, we add the prediction of emotion and causal span. We construct the following prompt template to recognize emotion and causal spans in the dialogue:

$$s_{prompt} = \text{"}[his]\ d_i\ sp_i \text{ feels } \langle mask_{emo} \rangle \text{ and says } u_i\ \langle mask_{cau} \rangle\ d_j\ sp_j \text{ says } u_j. \text{ The causal span start at } \langle mask_{st} \rangle \text{ , end at } \langle mask_{ed} \rangle\ .\text{"} \quad (5)$$

Where $mask_{emo}$ represents the masked emotion token that needs to be predicted and $mask_{st}, mask_{ed}$ represent the masked position of the causal span. The student model learns knowledge of the teacher model's causal reasoning from $mask_{cau}$ by predicting the connective words.

We feed s_{prompt} to RoBERTa (Liu et al. 2019) to obtain the representation of $mask_{cau}$ token $h^{cau}_{mask\_s}$ for further knowledge distillation. Meanwhile, we obtain the representation of $mask_{emo}$ token h^{emo}_{mask}, and the representations $h^{st}_{mask}, h^{ed}_{mask}$ for $mask_{st}$ and $mask_{ed}$ to make emotion and causal span prediction.

Emotion and causal span Prediction. Same to the teacher model, we use an MLMHead to map h^{emo}_{mask} and $h^{cau}_{mask\_s}$ to the token in vocabulary and normalize the probability logits using softmax:

$$\hat{y}^{emo}_{si} = Softmax(MLMHead(h^{emo}_{mask})), \quad (6)$$

$$\hat{y}^{cau}_{si} = Softmax(MLMHead(h^{cau}_{mask})). \quad (7)$$

For the prediction of spans, we use a linear layer and softmax layer to predict the logits for the entire sequence length, and select the highest logits as the start and end positions.

$$\hat{P}^{st}_{si} = Softmax(h^{st}_{si}W_{st} + b_{st}), \quad (8)$$

$$\hat{P}^{ed}_{si} = Softmax(h^{ed}_{si}W_{ed} + b_{ed}), \quad (9)$$

where $W_{st}, W_{ed} \in \mathbb{R}^{dim*l}$, dim represents the dimension of the hidden layer, and l represents the length of the entire historical utterance. Afterwards we use cross-entropy to calculate the loss for both emotion and causal span prediction:

$$\mathcal{L}^s_{emo} = -\frac{1}{\mathcal{M}} \sum_{i \in \mathcal{M}} y^{emo}_i log \hat{y}^{emo}_{si}, \quad (10)$$

$$\mathcal{L}^s_{sp} = -\frac{1}{\mathcal{M}} \sum_{i \in \mathcal{M}} (y^{st}_i log \hat{P}^{st}_{si} + y^{ed}_i log \hat{P}^{ed}_{si}). \quad (11)$$

Knowledge Distillation

As shown in Figure 2, our **KD-ECS** model consists of two components: the teacher model, which aims to learn utterance-level causal reasoning from a coarse-grained dataset, and the student model, which acquires the causal relation knowledge between utterances from the teacher by approximating its predicted vector output for the connectives.

In the training stage, we use the method of offline knowledge distillation(Gou et al. 2021), The teacher model learn utterance-level causal reasoning from a coarse-grained dataset, which benefits subsequent span recognition by reducing the search scope. The student model requires to serve testing inference to make fine-grained causal reasoning without causality between coarse-grained uterances, therefore, after fine-tuning and obtaining a well-performing teacher model, we force the student model to produce vectorized outputs similar to the results of the teacher model. We calculate $\mathcal{L}_{KD}$ by using Kullback-Leibler divergence (Hershey and Olsen 2007) to measure the gap in connectives prediction between the student and teacher model.

$$\mathcal{L}_{KD} = \sum_{i}^{K} \hat{L}^{cau}_{ti} log(\hat{L}^{cau}_{ti} / \hat{L}^{cau}_{si}), \quad (12)$$

Where $\hat{L} = Softmax(e_{mask}/\tau)$ and e_{mask} are the pre-softmax logits output by the model in Formula 2 and Formular 7, and τ is the temperature rate parameter used to alleviate the issue of class imbalance in knowledge distillation.

In the training stage, The student model performs joint learning of emotion and causality, and approaches the logits of the teacher model. The loss function of the student model is as follows:

$$\mathcal{L}_s = \alpha(\mathcal{L}^s_{emo} + \mathcal{L}^s_{span}) + (1-\alpha)\tau^2 \mathcal{L}_{KD}, \quad (13)$$

where α is the coefficient that balances these two terms, $\mathcal{L}^s_{emo}$ and $\mathcal{L}^s_{span}$ are the loss in Formula 10 and Formula 11.

Experiments

Dataset and Evaluation Metrics

We perform fine-grained emotion and causal span recognition on the `FG-RECCON` dataset, and coarse-grained emotion and causal utterance recognition on the `RECCON` dataset. Detailed information about the datasets is presented in Section Corpus Construction.

We use accuracy (*Acc*) to measure the performance of emotion recognition. Following (Poria et al. 2021), we report the F_1 scores of both negative and positive causal span pairs and the *macro*_F_1 score of them. We report exact match (*EM*) to examine the percentage of causal spans exactly extracted by the model out of from the gold standard data.

Baselines

To validate the effectiveness of our approach, we compare it with recent models for emotion cause recognition.

- The first set is the methods originally proposed as baselines for causal span extraction on the RECCON dataset (Poria et al. 2021). **Roberta-base** (Liu et al. 2019) and **Span-bert Fine-tuned on SQuAD** (Joshi et al. 2020) formulated the extraction of causal spans in dialogues as a machine reading comprehension (MRC) task.
- The second set is constituted by methods achieved good performance in ECPE tasks. We select **ECPE-2D** (Ding, Xia, and Yu 2020), **MGSAG** (Bao et al. 2022) and **KEC** (Li et al. 2022). We make a simple modification to these models' final predictions, changing the utterance-level classification task to a span extraction task. **UECA** (Zheng et al. 2022) can't be applied in the **ECSPE** task, we show its performance on the ECPE task on the `RECCON` dataset.
- The final set of baselines comprises large language models, specifically **LLaMA** (Touvron et al. 2023) and **ChatGPT**, selected to generate natural language expressions for emotional reasoning. For LLaMA, We take the dataset in the form of instruction data to fine-turn it using LoRA (Hu et al. 2021). For ChatGPT, we set corresponding natural language prompts and give an example for few-shot learning.

Implementation Details

Our model uses *RobertaForMaskedLM* (Liu et al. 2019) as the backbone, which is aqcuired from *huggingface Trans-*

Model	FG-RECCON(ECSPE)					RECCON(ECPE)			
	Emo *Acc*	*EM*	$Pos.\ F_1$	$Neg.\ F_1$	*macro_*F_1	Emo *Acc*	$Pos.\ F_1$	$Neg.\ F_1$	*macro_*F_1
Roberta	40.30	26.08	53.18	84.75	68.96	55.98	49.77	85.28	67.53
SpanBert	42.87	28.05	54.21	85.26	69.74	-	-	-	-
ECPE-2D*	44.26	28.11	55.18	86.4	70.79	52.76	52.39	95.86	73.62
MGSAG*	46.74	29.37	55.69	87.93	71.81	61.11	56.56	91.86	74.21
UECA	-	-	-	-	-	57.62	57.39	90.45	73.92
KEC*	47.02	30.19	56.94	88.29	72.62	59.33	59.47	92.74	76.11
LLAMA-7B	47.87	31.25	56.82	88.70	72.76	59.61	59.02	91.60	75.31
LLAMA-13B	48.10	32.09	57.14	88.75	72.94	60.48	59.83	92.24	76.03
ChatGPT	**67.20**	33.78	58.57	89.42	73.99	**76.39**	60.44	92.32	76.38
KD-ECS	48.26	**33.85**	**58.62**	**89.70**	**74.16**	61.04	**60.51**	**92.87**	**76.69**

Table 4: Experimental results for **ECSPE** task on FG−RECCON and **ECPE** task on RECCON datasets. The best results of each part are underlined. The models marked with an '*' represent that we have made simple modifications to the models to adapt to the current task.

Model	*EM*	$Pos.\ F_1$	*macro_*F_1
KD-ECS	**33.85**	**58.62**	**74.16**
w/o KD	30.49	56.27	72.74
w/o connectives	29.13	55.34	71.59
w/o $mask_{cau}$	28.68	54.76	71.04

Table 5: Ablation study on FG−RECCON dataset.

formers[2]. We adopt the AdamW optimizer (Loshchilov and Hutter 2017), set learning rate as $1e^{-5}$ and batch size as 16. For the hyperparameter in knowledge distillation, we set $\alpha = 0.6$ and $\tau = 2$ by grid search. The entire project is based on the PyTorch Lightning framework[3], and all other settings are default parameters. We conduct experiments using 5 random seeds and select the model with the best performance on the validation set. We then evaluate this model on the test set to report its results.

Experimental Results

Table 4 displays our main results. For the **ECSPE** task, our method achieves state-of-the-art (SOTA) performance. Specifically, compared with the best-performing model KEC in the second set, our model improves the accuracy of emotion and span extraction by 1.24% and 3.66%, and has a 1.68% improvement on $Pos.\ F_1$. This indicates that our knowledge distillation method enables us to learn more implicit causal knowledge from large models. Meanwhile, our model demonstrates more significant improvements compared to the baseline proposed by (Poria et al. 2021). Our model has a 4.41% improvement on $Pos.\ F_1$ compared to SpanBERT. This suggests that compared to MRC mechanisms, our prompt-based approach improves the reasoning ability of the bert-based model. Our model's accuracy in fine-grained emotion recognition is not able to match that of large language models because the large model itself has a good emotion recognition ability (Bubeck et al. 2023), and FG−RECCON is marked by ChatGPT. But, we achieve similar performance in identifying causal spans with ChatGPT, which indicates that we obtain the ability to perform causal reasoning through knowledge distillation from the relatively LLaMA model. Additionally, upon inspecting ChatGPT's results, we find that when given one turn prompt, it often forcefully interpret the relationship between each utterance in the dialogue history and the emotion utterance. This leads to many incorrect causal spans, resulting in ChatGPT's suboptimal performance. Perhaps this issue could be further improved through chain of thought or multi-step prompts.

Besides, through knowledge distillation, our **KD-ECS** model learns utterance-level cause recognition from the teacher model, enabling it to achieve good performance on the coarse-grained **ECPE** task as well. From Tabel 4, our model improves 1.04% on $Pos.\ F_1$ compared to KEC. The accuracy of emotion is still not superior to that of ChatGPT, but in cause identification, our method also achieves comparable performance with the large model. This proves that knowledge distillation directly improves the effect of utterance-level cause recognition in students' model.

Ablation Study

To analyze the performance of different modules, we conduct experiments on the following modifications: (1)**w/o KD**: We remove the teacher model and perform joint learning of utterance and span-level causal relation recognition. (2)**w/o connectives**: We further remove the prediction of causal connective words and use [cls] for utterance-level causal relation recognition. (3)**w/o** $mask_{cau}$: We only perform joint recognition of emotion and causal spans. In all experiments, we evaluate the performance of the models on the FG−RECCON dataset.

As shown in Table 5, the model with all modules exhibited better performance. Specifically, when the teacher model was removed, the *EM* performance of the model decreased by 3.36% and F_1 decreased by 1.42%. This indicates that the

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

[3]https://github.com/Lightning-AI/lightning

causal reasoning knowledge learned by the student model better assist span extraction compared to direct joint learning. Further removing the prediction of causal connectives result in 4.72% decrease in the *EM* performance, confirming that causal connectives are crucial linguistic cues for identifying discourse relations (Zhou et al. 2022). Removing the utterance-level causal relation recognition part resulte in 5.17% decrease in *EM* compared with the init model, indicating that there is a progressive relationship between utterance and span-level causal relation recognition.

Additionally, compared to direct joint learning, our knowledge distillation-based approach allows us to train the teacher model on existing coarse-grained data. This enables more stable and significant improvements even with a small amount of fine-grained training data. As shown in Figure 3, when only 10% of the fine-grained training data is used, our **KD-ECS** method outperforms the **ECS** method of joint learning by a 20% increase in $Pos.\ F1$. As the percentage of training data increases, the **KD-ECS** method outperforms joint learning methods, indicating that knowledge distillation can enhance model performance by leveraging existing knowledge, even in situations with limited samples. When trained on the complete FG−RECCON dataset, the **KD-ECS** method achieves a 2.35% improvement.

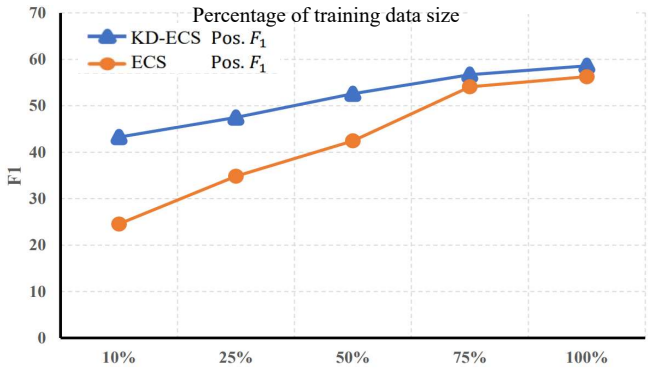


Figure 3: Comparison of methods based on knowledge distillation and joint learning under different scale training sets.

Discussion

Influence of Fine-Grained Emotion

In this section, we analyze the impact of fine-grained emotion on identifying the reasons behind emotions. We identify the causal span and compare the results with the given fine-grained and coarse-grained emotions respectively on the FG−RECCON dataset. The experimental results are shown in Table 6. We run two baselines proposed by (Poria et al. 2021) and our **KD-ECS** model for the experiments. It can be observed that when fine-grained emotion is provided, the model shows improvement in identifying reason spans. Specifically, our model shows an increase of 1.29% in span identification accuracy and 0.82% in $F1_{pos}$. The experiments demonstrate that distinguishing fine-grained emotion can provide more information when searching for reasons.

Model	FG-EM		CG-EM	
	EM	$Pos.F_1$	*EM*	$Pos.F_1$
Roberta	32.91	59.10	32.63	58.17
SpanBert	35.48	61.05	34.64	60.00
KD-ECS	**36.77**	**62.14**	**35.48**	**61.32**

Table 6: Experiment results for identifying the causal span with the given emotions on the FG−RECCON dataset, where "FG-EM" and "CG-EM" represent results on fine-grained and coarse-grained emotion labels respectively.

Influence of Teacher Model Scales

In this section, we investigate the effect of different sizes of teacher models on the knowledge distillation performance. We select five teacher models, namely RoBERTa-base, RoBERTa-large (Liu et al. 2019), t5-large (Raffel et al. 2020), LLaMA-7b, and LLaMA-13b (Touvron et al. 2023). As shown in Figure 4, with the increasing of the teacher model's size, the $Pos.\ F_1$ score of the student model is also significantly improved, but it will eventually become saturated. The best performance is achieved with LLaMA-7b. (Qiu et al. 2022) attributed this to the difference in capabilities between the teacher and student models. Specifically, small student models have difficulty understanding the higher-order knowledge extracted by a excessively large model. This issue could potentially be addressed in the future using dynamic knowledge distillation (Qiu et al. 2022) to overcome the performance bottlenecks.

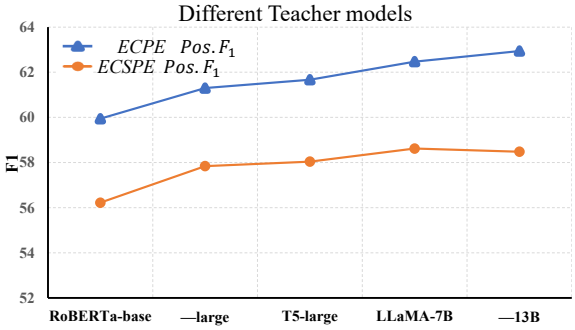


Figure 4: $Pos.\ F_1$ of KD-ECS with various teachers on the FG−RECCON dataset.

Conclusion

In this paper, we propose a novel distillation method for fine-grained emotion and causal span recognition. Our proposed method leverages the utterance-level causal reasoning ability of a teacher model to guide the performance of a student model.Additionally, to facilitate further research, we present the FG−RECCON dataset for this task using a combination of ChatGPT annotation and manual annotation methods based on the RECCON dataset. Detailed experimental results demonstrate the state-of-the-art performance of our model on both FG−RECCON and RECCON datasets.

Acknowledgments

We would like to thank the anonymous reviewers for helpful questions and comments. This work was supported by Ant Group Research Fund & the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901).

References

Bao, Y.; Ma, Q.; Wei, L.; Zhou, W.; and Hu, S. 2022. Multi-Granularity Semantic Aware Graph Model for Reducing Position Bias in Emotion-Cause Pair Extraction. *arXiv preprint arXiv:2205.02132*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Choi, D.; Choi, H.; and Lee, H. 2022. Domain Knowledge Transferring for Pre-trained Language Model via Calibrated Activation Boundary Distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1658–1669.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Ding, Z.; Xia, R.; and Yu, J. 2020. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3161–3170.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200.

Gao, J.; Liu, Y.; Deng, H.; Wang, W.; Cao, Y.; Du, J.; and Xu, R. 2021. Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 807–819. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Gao, Q.; Hu, J.; Xu, R.; Gui, L.; He, Y.; Wong, K.-F.; and Lu, Q. 2017. Overview of NTCIR-13 ECA Task. In *NTCIR*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129: 1789–1819.

Gui, L.; Yuan, L.; Xu, R.; Liu, B.; Lu, Q.; and Zhou, Y. 2014. Emotion cause detection with linguistic construction in chinese weibo text. In *Natural Language Processing and Chinese Computing: Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings 3*, 457–464. Springer.

Hershey, J. R.; and Olsen, P. A. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, IV–317. IEEE.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, D.; Wei, L.; and Huai, X. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Jiang, Y.; Chan, C.; Chen, M.; and Wang, W. 2023. Lion: Adversarial Distillation of Closed-Source Large Language Model. *arXiv preprint arXiv:2305.12870*.

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8: 64–77.

Kim, H.; and Kim, B. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

Li, J.; Meng, F.; Lin, Z.; Liu, R.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022. Neutral Utterances are Also Causes: Enhancing Conversational Causal Emotion Entailment with Social Commonsense Knowledge. *arXiv preprint arXiv:2205.00759*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pitler, E.; and Nenkova, A. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 13–16. Suntec, Singapore: Association for Computational Linguistics.

Poria, S.; Majumder, N.; Hazarika, D.; Ghosal, D.; Bhardwaj, R.; Jian, S. Y. B.; Hong, P.; Ghosh, R.; Roy, A.; Chhaya, N.; et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5): 1317–1332.

Qiu, Z.; Ma, X.; Yang, K.; Liu, C.; Hou, J.; Yi, S.; and Ouyang, W. 2022. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Schick, T.; and Schütze, H. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Xia, R.; and Ding, Z. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*.

Zheng, X.; Liu, Z.; Zhang, Z.; Wang, Z.; and Wang, J. 2022. UECA-Prompt: Universal Prompt for Emotion Cause Analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, 7031–7041.

Zhou, H.; Lan, M.; Wu, Y.; Chen, Y.; and Ma, M. 2022. Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition. *arXiv preprint arXiv:2210.07032*.

Zhou, Z.; Lan, M.; and Wu, Y. 2018. A Neural Generation-based Conversation Model Using Fine-grained Emotion-guide Attention. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, 1–8. IEEE.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.