Emotion Cause Pair Extraction Using a Sequence to Sequence Model

CSE 842 Final Report

Jay Ho hophuc@msu.edu
David Khankin khankind@msu.edu

November 28th, 2023

1 Introduction

In 2019, Rui Xia and Zixiang Ding from Nanjing University of Science and Technology in China proposed a new task to extract the potential pairs of emotions and their corresponding causes from a conversation [2]. The main challenge of this task is in handling the joint process of emotion and cause extraction given just a conversation. These two tasks independently have been approached and studied, with models that are dedicated to simply one of these tasks or the other. The consideration of studying both is important because, as stated by Fanfan et al, "the way of annotating emotions first and then extracting causes ignores the fact that emotions and causes are mutually indicative" [5].

2 Problem Description

Emotion-Cause Pair Extraction is a task in natural language processing that involves identifying and extracting pairs of emotions and corresponding causes from textual data such as a conversation. This includes recognizing different emotion categories according to Ekman's six basic emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise [9]. From this, cause refers to the objective event or subjective argument that triggers the corresponding emotion. The existence of this problem stems from the complexity inherent in human emotions, especially as emotions can manifest in a variety of ways, and without context specific information, a given phrase can carry different weights and meaning to it. Extracting emotion-cause pairs requires handling this ambiguity, alongside implementing sophisticated natural language processing techniques.

Human-computer conversational interaction is rapidly gaining popularity as programs such as ChatGPT are being developed and made publicly available. This task's significance lies in understanding human behavior, as in doing so, we can create emotion-aware systems which can respond with more empathy to user questions and make interactions between human and AI feel far more human-like, leading to more natural and comfortable conversation.

3 Implementation

From the previous models proposed for this ECPE task, one that stands out is the Unified Target-Oritented Sequence-to-Sequence model (UTOS) proposed by Cheng et al [4]. The main aspect of this model is that it performs the entire pair labeling process for emotion and cause at the same time, as opposed to past two-step approaches, as mentioned in Section 5 Related Work, that focus on labeling emotions first and then finding their corresponding causes.

Our training data has emotion cause pairs that are already identified along with conversations split into utterances. Therefore, the model learns to predict emotion and cause pairs based on features presented in the training data itself. Our initial data is in a text file that contains all conversations, utterances, and emotion-cause pairs, and for the purpose of our model we convert this into JSON format. Our goal is to use the existing UTOS model as a backbone and modify its layers and details in order to achieve a higher accuracy on our data.

3.1 Model

UTOS MODEL STRUCTURE As shown in Figure 1, the UTOS model consists of a BERT en-

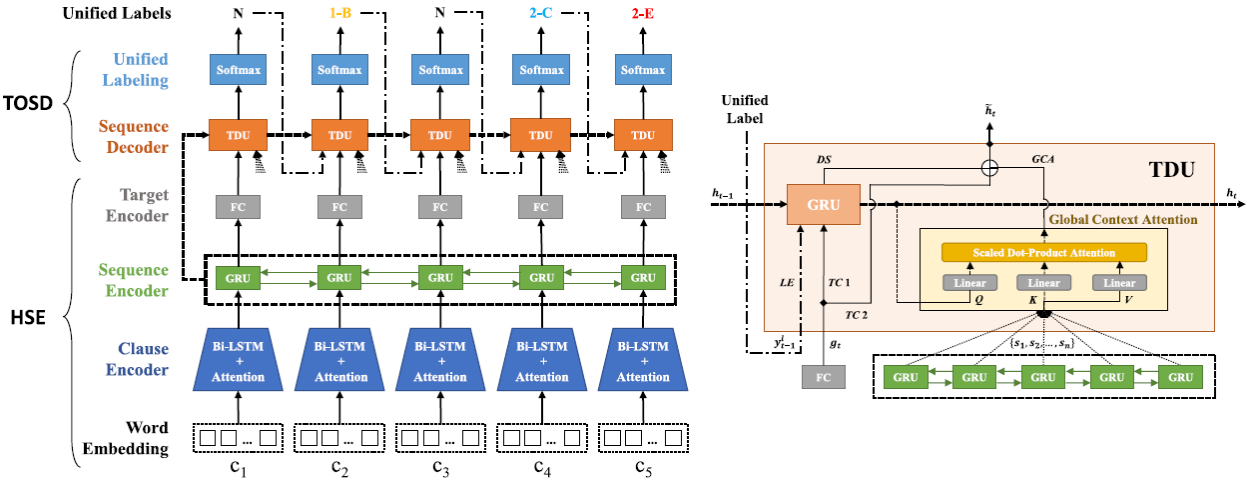

Figure 1: Left is architecture of UTOS model. Right is details of Target Decoder Unit [4].

coder [1] to capture the utterance representation, a sequence encoder and target-oriented sequence decoder, and a softmax layer to produce the predictive probability distribution on a set of unified labels which is the unified labeling process. The target-oriented sequence decoder employs a target decoder unit that consists of a global context attention layer to get the final utterance representation for the unified labeling process.

MODEL MODIFICATIONS As the base UTOS model used BERT-chinese as its utterance encoder, we swapped this to BERT-base-cased from Hugging Face, as you can see in Figure 2, which allows us to handle English data. We also ran this model using BERT-base-uncased, but as our data is cased, we expectantly achieved a lower result.

In order to achieve a baseline result, we stripped down the UTOS model to minimize its interaction with the data. At this level, we removed the original attention layer in the target decoder unit of the sequence decoder and leaky ReLU layer. After we removed those layers, only the GRU layers remain in the sequence encoder and decoder.

Building upon the original model, we improved the UTOS model by adding the BiLSTM layer after the BERT encoder to capture more context of the utterance, and we replaced the original attention layer with Pytorch's multihead attention layer, which has 5 heads as opposed to 1 as the original did. This is represented in Figure 2, which showcases the changes we made from the original UTOS model.

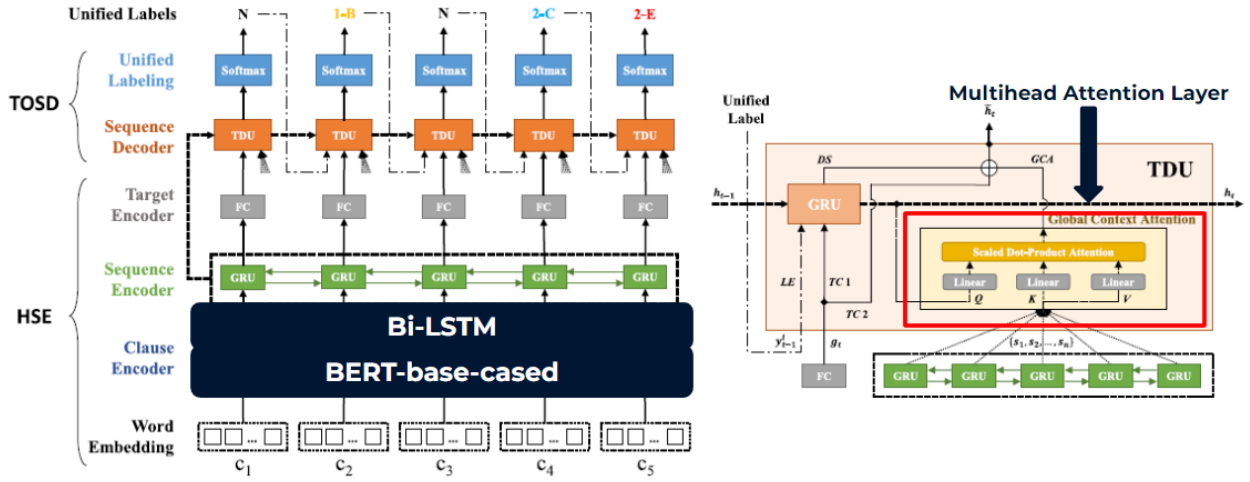

Figure 2: Showcase of Modifications Implemented to UTOS Model.

These changes allow the model to capture more patterns and information from different representation subspaces simultaneously [8]. We tried various other number of heads, but found that 5 resulted in the highest F1 score for our data.

MODEL INPUT The model input is a group of conversations, and as stated, each conversation is split into utterances where each utterance is a list of words. This is to separate conversations, and furthermore split the conversation into multiple pieces of the data for the model to train on.

MODEL OUTPUT The model output is a sequence of unified labels, where each label is generated by a softmax layer, which takes the final representation of a target utterance from the sequence encoder and produces the predictive probability distribution on the set of unified labels. Based on the UTOS paper, authors designed a special set of unified labels to be 1-E, 1-C, 1-B, 2-E, 2-C, 2-B, $\cdots$, k-E, k-C, k-B in N where N indicates neither emotion or cause is identified in the utterance, and each unified label contains the content part and the pairing part [4]. The content part can be labeled as E (emotion), C (cause), or B (both) to indicate the type of content, and the pairing part can be labeled as 1,2, ..., k to indicate how to pair utterances [4].

Referring to Figure 3 as an example, we can see 5 utterances within a conversation, and two different pairs identified. The first pair is identified in utterance c_2, and is labeled as 1-B. This identifies it as the first label found, and the B represents that both the emotion and cause come from utterance c_2. The second pair found is labeled as c_5, c_4. The unified labeling scheme represents the corresponding utterances as 2-C and 2-E, where 2-E is labeled on utterance c_5, showing this is where the emotion is coming from, and 2-C is labeled on utterance c_4, showing that this is where the cause originates from.

3.2 Data

In this experiment, we used an English dataset from Singh et al [7]. The original UTOS model was trained on the Chinese dataset from Xia and Ding [2]. The Chinese dataset is smaller than the English dataset [7] that we decided to use as shown in Table 1. We chose this dataset as it has been used for the ECPE task prior to this instance, and followed the same general format as the original UTOS Chinese dataset.

EXPERIMENTAL SETTING The dataset is split into two parts: 90% for training and 10% for testing. The original authors' BertAdam optimizer is swapped to the AdamW optimizer from Hugging Face Transformers and is trained with various learning rates such as $2e-5, 3e-5, 3e-6$. It is found that the learning rate $3e-6$ provided the best result. We trained the modified model with 50 epochs and used early stopping technique. The learning rate of BERT is $2e-5$, and the learning rate for other layers in the model, such as the fully connected layer and Bi-LSTM layer, are set to $1e-4$. Other than these parameters, we used the original setting of the UTOS model.

4 Evaluation

To evaluate model results, precision, recall, and F1 score are used as evaluation metrics. Precision is the total number of correct pairs divided by the total number of predicted pairs, recall is the total number of correct pairs divided by the total number of annotated pairs, and F1 is calculated by the algorithm shown in the following formulas.

$$P = \frac{\sum correctpairs}{\sum predictedpairs}$$

$$R = \frac{\sum correctpairs}{\sum annotatedpairs}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Predicted pairs is the total number of predicted emotion-cause pairs, annotated pairs is the number of labeled emotion-cause pairs in the dataset, and correct pairs is the number of pairs that are predicted accurately.

4.1 Results

As seen in Table 2, we showcase the models mentioned in Model Modifications from 3.1: UTOS-eng (original model run on english data), UTOS-sim (simplified model), and UTOS-com (improved version of the model). From this table one can see that our complex model (UTOS-com) had a trade-off. The precision of each task decreased, but we had a significant enough increase in recall that we are able to see an overall more accurate result. This means that when UTOS-eng and UTOS-sim predict emotion cause pair, it is more likely to be accurate, but they tend to miss out on more pairs overall, whereas

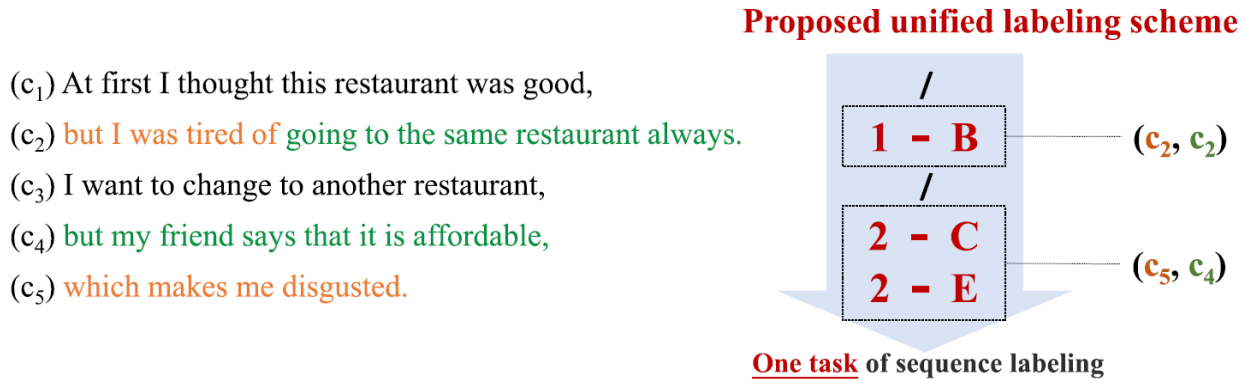

Figure 3: Example of the unified labeling scheme performed by the model [4].

UTOS-com was made to take more risks and managed to find more pairs overall, but had less independent accuracy in doing so.

MULTIHEAD ATTENTION To evaluate the effect of the attention mechanism in the UTOS model, we trained the model with a various number of heads in the multihead attention layer. In Table 3, UTOS-Att4, UTOS-Att5, and UTOS-Att6 are the UTOS base models with 4, 5, and 6 heads in the attention layer without the Bi-LSTM in the sequence encoder. As seen in Table 3, UTOS-Att6 has the highest recall and F1-score on emotion extraction, cause extraction, and emotion-cause pair extraction. While UTOS-Att5 has the highest precision in all categories which is another instance of precision-recall trade-off. Therefore, we decided to move further by adding the Bi-LSTM layer in the sequence encoder. Adding heads to the attention layer, and introducing the Bi-LSTM layer, both have the effect of lowering the precision but increasing the recall, yet do so in alternate ways. We found that with the Bi-LSTM layer addition, the UTOS model with the 5-heads attention layer (UTOS-com) yield the best result as shown in 2.

Bi-LSTM BERT is a transformer-based model that captures contextual information from left to right context words. However, BERT processes words in parallel and its attention mechanism is not sequential. Therefore, we believe that adding a Bi-LSTM layer can further capture the sequential patterns that BERT may miss in the utterance representation produced by BERT, and increase our recall score. UTOS-com contains the Bi-LSTM layer, and as can be seen in Table 2 and Table 3, this seems to hold true. In Table3, which has the Bi-LSTM layer removed from UTOS-com, we simply have lower recall scores overall, and as we can see in UTOS-com from Table 2, which includes both the attention layer and the Bi-LSTM layer, we have overall higher recall scores, and a higher F1 score.

4.2 Problems

The base UTOS model has one issue with its design that results in a lower recall. The issue is within the implemented unified labeling system, where each utterance is restricted to a single label. This means, if an utterance is tagged with a label such as 1-B, signifying the emotion and cause from the first pair come from this utterance, it cannot simultaneously hold labels for 2-E or 2-C, representing the emotion or cause for the second pair, respectively. This limitation becomes an issue when considering scenarios where the cause or emotion for pair 2 might be expressed in a different utterance, dependent on the initial one. However, it's important to note that

Table 1: Statistics of Datasets

Statistic	UTOS Chinese	English
Number of conversations	1945	2843
Number of pairs	N/A	3215
Number of conversations with 1 pair	1746	2537
Number of conversations with 2 pairs	177	256
Number of conversations with more than 2 pairs	22	50

Table 2: UTOS Model Performances

Model	Emotion Extraction			Cause Extraction			Emotion-Cause Pair Extraction		
	P	R	F1	P	R	F1	P	R	F1
UTOS-eng	0.9072	0.3088	0.4607	**0.9691**	0.3113	0.4712	**0.9072**	0.2914	0.4411
UTOS-sim	**0.9082**	0.3123	0.4648	0.9388	0.3046	0.4600	0.8878	0.2881	0.4350
UTOS-com	0.7635	**0.3965**	**0.5219**	0.8243	**0.4040**	**0.5422**	0.7500	**0.3675**	**0.4933**

the number of such instances where this is present is remarkably low, as a majority of conversations involve just one pair, and even in cases with multiple pairs, the likelihood of encountering this specific issue is not guaranteed. While we do not have exact numbers on how often this issue occurs within our English dataset, the original UTOS models authors found that this issue occurred in only one conversation from their chinese dataset, and implies that the unified labeling scheme is practical [4].

5 Related Work

In the context of this task, our research found many alternate approaches and methods to solve this task. Below are some of the more significant papers and models that we used as a foundation for our experiments and modifications.

Xia and Ding proposed a 2-step approach to solve this emotion-cause pair extraction problem with independent and interactive multi-task learning and filtering [2]. While this approach is effective, it is noted that a flaw of this method is cascading failure, where if an emotion is labeled wrong, the corresponding cause can no longer be expected to be extracted accurately. There have since been multiple other models and methods used to approach this problem in order to improve the accuracy of emotion-cause pair extraction. Ding et. al (2020) proposed a joint framework for ECPE with sliding window multi-label learning [3]. Cheng et. al (2021) suggested a one-step approach, a comprehensive target-oriented sequence-to-sequence model which reframed ECPE as a unified sequence labeling problem and extracted emotion-cause pairs in one pass as opposed to the two step approach previously used [4]. In 2021, Wang et. al leveraged a two-step approach from Xia and Ding [2] to extract emotion-cause pairs from the text, video, and audio of a given conversation [5]. Existing methods (Xia and Ding 2019) and end-to-end framework (Cheng et. al 2021) did not explicitly learn the relationship between various task objectives; therefore, Zheng et al proposed a universal prompt-based method to solve different tasks such as emotion-cause pair extraction, emotion cause extraction, and conditional causal relationships classification in emotion cause analysis [6].

6 Conclusion

The ECPE task has seen significant strides since 2019, when Xia and Ding [2] proposed the task. Notably, the Unified Target-Oriented Sequence-to-Sequence (UTOS) model by Cheng et al [4] introduced the simultaneous labeling of emotions and their causes within conversations. Making modifications with BERT-based encoders, Bi-LSTM layers, and multihead attention mechanisms have shown potential in increasing accuracy, albeit with trade-offs between precision and recall. Evaluation metrics such as precision, recall, and F1 scores have highlighted these trade-offs among different model configurations. Challenges remain, particularly in the limitations of the unified labeling system within the base UTOS model, but diverse approaches from multi-task learning to end-to-end sequence labeling models in ECPE research have collectively expanded our understanding and methods in extracting emotion-

Table 3: Multihead Attention Performance without Bi-LSTM in Sequence Encoder

Model	Emotion Extraction			Cause Extraction			Emotion-Cause Pair Extraction		
	P	R	F1	P	R	F1	P	R	F1
UTOS-Att4	0.763	0.3614	0.4905	0.8444	0.3775	0.5217	0.7556	0.3377	0.4668
UTOS-Att5	**0.8333**	0.3509	0.4938	**0.8678**	0.3477	0.4965	**0.7769**	0.3113	0.4444
UTOS-Att6	0.7762	**0.3895**	**0.5187**	0.8531	**0.404**	**0.5483**	0.7552	**0.3576**	**0.4854**

cause pairs from conversational text, aiming to create more empathetic human-AI interactions and furthering the field's progress.

References

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

[2] *Rui Xia and Zixiang Ding. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.*

[3] *Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. End-to-End Emotion-Cause Pair Extraction based on Sliding Window Multi-Label Learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3574–3583, Online. Association for Computational Linguistics.*

[4] *Z. Cheng, Z. Jiang, Y. Yin, N. Li and Q. Gu, "A Unified Target-Oriented Sequence-to-Sequence Model for Emotion-Cause Pair Extraction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2779-2791, 2021, doi: 10.1109/TASLP.2021.3102194.*

[5] *Wang, Fanfan, Zixiang Ding, Rui Xia, Zhaoyu Li and Jianfei Yu. "Multimodal Emotion-Cause Pair Extraction in Conversations." IEEE Transactions on Affective Computing 14 (2021): 1832-1844.*

[6] *Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. UECA-Prompt: Universal Prompt for Emotion Cause Analysis. In Proceedings of the 29th International Conference on Computational Linguistics, pages 7031–7041, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.*

[7] *Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021. An end-to-end network for emotion-cause pair extraction. In 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 84–91.*

[8] *A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.*

[9] *Ekman, P. (1992). Are there basic emotions? Psychological Review, 99(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550*