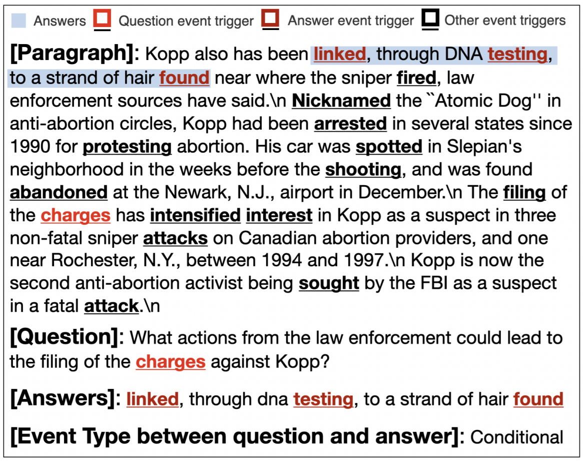
Event-Centric Question Answering via Contrastive Learning and Invertible Event Transformation

Junru Lu[1]**, Xingwei Tan**[1]**, Gabriele Pergola**[1]**, Lin Gui**[2] **and Yulan He**[1,2,3]
[1]Department of Computer Science, University of Warwick, UK
[2]Department of Informatics, King's College London, UK
[3]The Alan Turing Institute, UK
{Junru.Lu, Xingwei.Tan, Gabriele.Pergola}@warwick.ac.uk
{lin.1.gui, yulan.he}@kcl.ac.uk

Abstract

Human reading comprehension often requires reasoning of event semantic relations in narratives, represented by Event-centric Question-Answering (QA). To address event-centric QA, we propose a novel QA model with contrastive learning and invertible event transformation, call TranCLR. Our proposed model utilizes an invertible transformation matrix to project semantic vectors of events into a common event embedding space, trained with contrastive learning, and thus naturally inject event semantic knowledge into mainstream QA pipelines. The transformation matrix is fine-tuned with the annotated event relation types between events that occurred in questions and those in answers, using event-aware question vectors. Experimental results on the **E**vent **S**eman**t**ic **R**elation **R**easoning (ESTER) dataset show significant improvements in both generative and extractive settings compared to the existing strong baselines, achieving over 8.4% gain in the token-level F1 score and 3.0% gain in Exact Match (EM) score under the multi-answer setting. Qualitative analysis reveals the high quality of the generated answers by TranCLR, demonstrating the feasibility of injecting event knowledge into QA model learning. Our code and models can be found at https://github.com/LuJunru/TranCLR.


▨ Answers □ Question event trigger □ Answer event trigger □ Other event triggers

[Paragraph]: Kopp also has been **linked**, through DNA **testing**, to a strand of hair **found** near where the sniper **fired**, law enforcement sources have said.\n **Nicknamed** the ``Atomic Dog'' in anti-abortion circles, Kopp had been **arrested** in several states since 1990 for **protesting** abortion. His car was **spotted** in Slepian's neighborhood in the weeks before the **shooting**, and was found **abandoned** at the Newark, N.J., airport in December.\n The **filing** of the **charges** has **intensified** **interest** in Kopp as a suspect in three non-fatal sniper **attacks** on Canadian abortion providers, and one near Rochester, N.Y., between 1994 and 1997.\n Kopp is now the second anti-abortion activist being **sought** by the FBI as a suspect in a fatal **attack**.\n

[Question]: What actions from the law enforcement could lead to the filing of the **charges** against Kopp?

[Answers]: **linked**, through dna **testing**, to a strand of hair **found**

[Event Type between question and answer]: Conditional

Figure 1: An event-centric QA example from the ESTER dataset (Han et al., 2021). All event triggers are highlighted in bold and underlined in the paragraph. The question event trigger and answer event triggers are further highlighted in red colors with different shades. In-text answer is smeared with blue.

1 Introduction

Since 2019, many larger-scale pre-trained language models (PLMs) (Devlin et al., 2018; Raffel et al., 2019; Lu et al., 2020; Pergola et al., 2021b) have been introduced to address the Question-Answering (QA) tasks, reaching performance on par with humans on entity-centric QA datasets such as SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), and NewsQA (Trischler et al., 2016), in which answers are often entities extracted from text. A raising challenge is to research and develop new PLM-based frameworks tackling more difficult QA settings in real-world scenarios. One direction is to go beyond entity-centric QA and explore QA tasks focusing on high cognitive level information such as **events**. A recently introduced **E**vent **S**eman**t**ic **R**elation **R**easoning (ESTER) dataset (Han et al., 2021) facilitates the development of Machine Reading Comprehension (MRC) models for event-centric QA. The dataset contains event-centric question-answers annotated with event semantic relation type labels. Figure 1[1] shows an example instance from the dataset. The main challenge is to effectively explore event semantic knowledge to answer event-centric questions. In the example illustrated in Figure 1, an MRC or QA model needs to first understand that the question asks for a potential answer event which holds a *conditional* relation with the main event '*charges*' mentioned in the question. It then needs to identify events in the paragraph which have the

[1] Better viewing in color.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2377–2389
December 7-11, 2022 ©2022 Association for Computational Linguistics

conditional relation with the question event trigger '*charges*', in this case, '*linked*', '*testing*' and '*found*'. Finally, it needs to generate the answer involving the identified events in natural language. It is easy for humans to understand narratives by constructing a situational logic chain capturing how events evolve and relate to each other in text. Yet, existing QA models only learn shallow semantic cues based on word token statistics gathered from large-scale text corpora (Niven and Kao, 2019), but are not able to grasp high-level concepts such as events. Preliminary experimental results using the pre-trained T5 language model on the ESTER dataset show that there remains a large gap over 15% between machine and human performance (Han et al., 2021).

Intuitively, it is possible to inject event information through a multi-task learning framework where event-related tasks such as event relation type detection and event embedding learning could be potentially useful to guide the QA model to generate better answers. For example, event relation type detection aims to detect the desired event semantic relation given a question, while event embedding learning aims to push events holding the desired semantic relation closer in the new event embedding space. However, in a QA model, the answer generator is usually built on a PLM in which the original event representations learned in the PLM should be preserved. That is, we want to map event representations onto a new event embedding space in order to inherently capture their semantic relations specified by an input question, but at the same time, we need to keep the original event representations learned by the PLM in order to generate coherent answers. To deal with this dilemma, we propose an invertible transformation operator, which makes it possible to learn new event embeddings without changing the mutual information of any given event pairs, making it effective in injecting event information for event-centric QA.

More concretely, to leverage the event semantic knowledge into QA models, we propose a novel multi-task learning framework, named TranCLR (Fig. 2), combining a general-purpose QA model, with an event invertible transformation operator to encode event relations across questions and paragraphs. It builds on the UnifiedQA (Khashabi et al., 2020) model for answer generation, and employs an invertible event transformation operator to project the hidden representations from the UnifiedQA encoder onto a new *event embedding space*. The transformed representations are then used for (i) contrasting learning and for (ii) event relation type classification. The contrastive learning mechanism is adopted to realign the event vectors, strengthening the relations between the events mentioned in questions and those candidate answer events in paragraphs and improving the generalization to out-of-distribution event relations. On the other hand, the event relation type classification is used to further fine-tune the transformation matrix through contextualized question representations. The combination of the transformation operator, along with contrastive learning and event relation type classification, leads the model to focus on the textual and relation features characterizing the event occurrences in text, and results in an overall boost in performance on event-centric QA tasks.

Our contributions can be summarized as follows: **(1)** We introduce a novel multi-task learning framework for event-centric QA, TranCLR, in which we design an invertible event transformation operator and a contrastive learning mechanism, further combined with event relation type classification, to perform better reasoning on event semantic relations; **(2)** We conduct an experimental assessment on the ESTER dataset showing that TranCLR boosts the performance of QA models compared to strong existing PLM-based QA baselines, achieving over 8.4% and 3.0% gain in the token-level F1 and EM score respectively under the multi-answer setting; **(3)** Visualization of event-aware token semantic vectors verifies the effectiveness of event knowledge injection. We further show the advantages of our framework tailored for event-centric learning on both zero- and few-shot learning, and adaptation ability on out-of-domain event-centric questions.

2 Related work

This work is related to two lines of research: event-centric QA, and contrastive learning.

Event-centric QA The growing interest into event understanding has recently led to the development of new resources for event-centric QA and event relation extraction. Souza Costa et al. (2020) proposed *EventQA*, an event-centric QA dataset to access semantic information stores in knowledge graphs. The questions are created via a random walking on the EventKG (Gottschalk and Demidova, 2019), then manually translated into natural language. Ning et al. (2020a) modified and converted an event temporal relation extraction dataset – MATRES (Ning et al., 2018) into a reading comprehension format focused on event temporal ordering questions, named TORQUE. Instead of solely focusing on simple arguments or temporal relations, the ESTER dataset (Han et al., 2021) was developed to highlight how events are semantically related in terms of five most common semantic relations: *Causal, Conditional, Counterfactual, Sub-event,* and *Coreference* relations. Aforementioned work built dataset baselines with popular entity-based PLMs, and thus leave significant performance gaps compared with human evaluation. Asai and Hajishirzi (2020), Dua et al. (2021) and Shang et al. (2021) leverage features of closely related questions to capture temporal difference to deal with certain types of event-centric questions. Compared to the existing works, we target to various types of event-centric questions. Therefore, we introduce an invertible event transformation to (i) model the event semantic relations through an auxiliary classification task, and to (ii) realign the event latent representations via contrastive learning in the space of the transformed events.

Contrastive Learning Approaches to contrastive learning for text focus on the generation of positive and negative training pairs from pretrained language models. For example, Clark et al. (2020) proposed a new pretraining framework named ELECTRA, which defines a new generative training task, i.e., Replacement Token Detection (RTD), with the aim of determining whether a token was originally replaced by the language model. Based on ELECTRA, Meng et al. (2021) designed two new pretraining tasks: the Correct Language Modeling (CLM), aiming at restoring a corrupted sentence; and the a contrastive learning-based task, in which the positive pairs are made of recovered sentences and corresponding previously corrupted sentences. Similarly, Qin et al. (2020) designed another contrastive learning framework ERICA for document-level text understanding, via specific entity discrimination pre-training task and relation discrimination pre-training task. Chen et al. (2022) proposed a two-stage framework, integrating answer-aware span-based contrastive learning for cross-lingual machine reading comprehension. Wu et al. (2020) and Fang et al. (2020) designed a framework similar to SimCLR (Chen et al., 2020) to generate sentence representations by applying several data augmentation strategies to create contrastive pairs, such as word deleting and swapping, back-translation and synonym replacement. Yet, Gao et al. (2021) reported that simply using dropout masks twice within a PLM can led to rather reliable positive pairs. Our work adopts standard contrastive learning framework. The positive and negative pairs of events are composed directly from the different text sections: questions, paragraphs, and answers within the paragraph.

3 Methodology

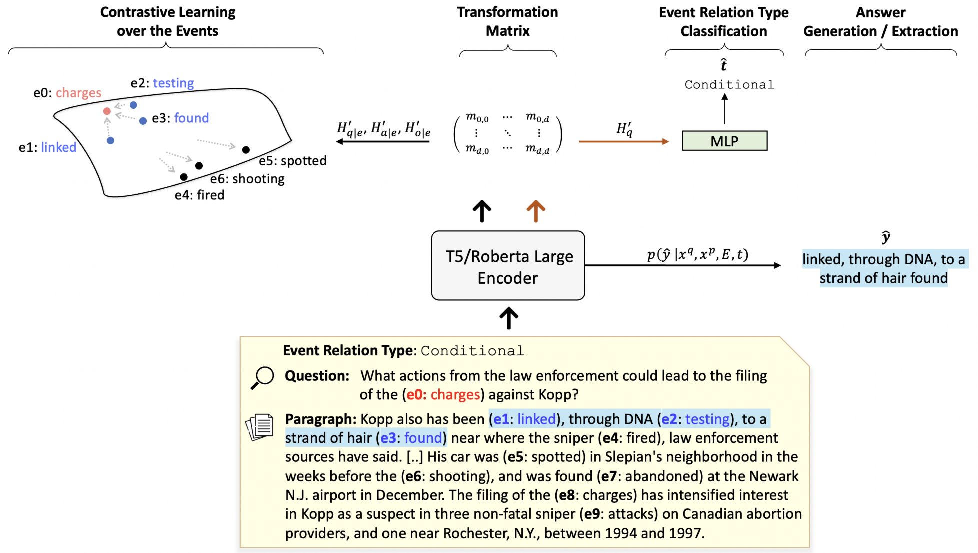
In this section, we first define the task of event-centric QA and then present our proposed TranCLR model. We build our model mainly based on the ESTER dataset (Han et al., 2021).

3.1 Task Formulation

Event-centric QA can be formulated as question answering centred on the understanding of event semantic relations. The task can be mathematically defined as: given a text passage $\boldsymbol{x}^p$ and an answerable event-centric question $\boldsymbol{x}^q$, a model is asked to provide one or more answers $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_A\}$, where A denotes the total number of answers to the given question. In the ESTER dataset, event triggers in text passages, questions and answers are annotated, $\boldsymbol{E} = \{\boldsymbol{e}^p_1, \cdots, \boldsymbol{e}^p_{C_p}, \boldsymbol{e}^q, \boldsymbol{e}^a_1, \cdots, \boldsymbol{e}^a_{C_a}\}$, where C_p and C_a denote the total number of events in the text passage and the answer, respectively. Each question only contains a single event $\boldsymbol{e}^q$. Since answers are parts of the paragraph in ESTER, paragraph event triggers also include answer event triggers. In addition, the relation type of the question event and answer events, $\boldsymbol{t} \in \mathcal{T}$, is also annotated. In the ESTER dataset, there are 5 event semantic relations types: *Causal*, *Conditional*, *Counterfactual*, *Subevent*, and *Co-reference*.

3.2 TranCLR

We propose a novel framework for event-centric question answering, called `TranCLR`, which is a multitask model via contrastive learning and invertible event transformation. The overall framework is shown in Figure 2. Following settings in the ESTER work (Han et al., 2021), we adopt T5-large (Raffel et al., 2019) as an encoder-decoder backbone for the generative setting (i.e., answer generation), and RoBERTa-large (Liu et al., 2019) as an encoder for the extractive setting (i.e., answer extraction). The T5-large model will be fine-tuned in a universal generative style (Khashabi


Contrastive Learning over the Events
Transformation Matrix
Event Relation Type Classification
Answer Generation / Extraction
e0: charges
e2: testing
e3: found
e1: linked
e5: spotted
e6: shooting
e4: fired
$H'_{q|e}, H'_{a|e}, H'_{o|e}$
$\begin{pmatrix} m_{0,0} & \cdots & m_{0,d} \\ \vdots & \ddots & \vdots \\ m_{d,0} & \cdots & m_{d,d} \end{pmatrix}$
H'_q
MLP
$\hat{t}$
Conditional
T5/Roberta Large Encoder
$p(\hat{y}\,|x^q, x^p, E, t)$
$\hat{y}$
linked, through DNA, to a strand of hair found

Event Relation Type: `Conditional`

Question: What actions from the law enforcement could lead to the filing of the (**e0:** charges) against Kopp?

Paragraph: Kopp also has been (**e1**: linked), through DNA (**e2**: testing), to a strand of hair (**e3**: found) near where the sniper (**e4**: fired), law enforcement sources have said. [..] His car was (**e5**: spotted) in Slepian's neighborhood in the weeks before the (**e6**: shooting), and was found (**e7**: abandoned) at the Newark N.J. airport in December. The filing of the (**e8**: charges) has intensified interest in Kopp as a suspect in three non-fatal sniper (**e9**: attacks) on Canadian abortion providers, and one near Rochester, N.Y., between 1994 and 1997.

Figure 2: The overall `TranCLR` architecture. The input to encoder is the concatenation of the event relation type, the question, and the paragraph. The resulting hidden vectors are used to provide answers. Simultaneously, the hidden representations are projected through a transformation matrix and used for both contrastive learning and event relation type classification. The contrastive learning mechanism realigns the event vectors to strengthen the relations between the event occurred in question and candidate answer events in paragraphs; while event relation type classification predicts the event relation type given a transformed question representation.

et al., 2020), therefore named as UnifiedQA-T5-large. During training, the input sequence consists of question-answer event relation type label $\boldsymbol{t}$, question $\boldsymbol{x}^q$ and passage $\boldsymbol{x}^p$ with ":", "\n", "</s>" and "<s>" special tokens. We use $\boldsymbol{x} = \{\boldsymbol{t}{:}\boldsymbol{x}^q\backslash n\boldsymbol{x}^p\}$ and $\{$<s>$\boldsymbol{t}{:}\boldsymbol{x}^q$</s></s>$\boldsymbol{x}^p\}$ to denote the whole input sequence for the generative and the extractive settings, respectively. Let N_x be the length of $\boldsymbol{x}$, and d be the dimension of hidden state vectors, $H \in \mathbb{R}^{N_x \times d}$ is the contextual hidden states of the encoder. The target label for the generative setting is the concatenation of all answers $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_A\}$ with each separated by a ";" special token, while labels for the extractive setting are $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_{\boldsymbol{x}^p}\}$, following the "B-I-O" or "I-O" tagging format.

After getting the hidden states H of an input sequence $\boldsymbol{x}$ via the UnifiedQA-T5-large or the RoBERTa-large encoder, we simultaneously train the model with contrastive learning for two tasks, the main QA task, and the auxiliary task for event relation type classification. Therefore, our model is designed to maximize the probability $p(\hat{\boldsymbol{y}}|\boldsymbol{x}^q, \boldsymbol{x}^p, \boldsymbol{E}, \boldsymbol{t})$ of the generated answers or the predicted labels given a question $\boldsymbol{x}^q$, the supporting paragraph $\boldsymbol{x}^p$, all event triggers in the materials $\boldsymbol{E}$, and question-answer event relation type label $\boldsymbol{t}$. The event relation type $\boldsymbol{t}$ can be considered as a prefix or prompt to the input in prompt-based learning. It is worth noting that the annotated event triggers are **only** used in training, but not in inference.

The key to event-centric QA is to perform reasoning on the semantic relation of the event found in the question and those candidate events in the paired text passage. For example, for the question in Figure 2, "*What actions from the law enforcemnet could lead to the filling of the charges against Kopp?*", we would expect the QA model to generate the answer which contains the event(s) that exhibit the *Conditional* relation with the event *charges* mentioned in the question. This is somewhat similar to node prediction in knowledge graph embedding learning, that is, given the head event $\boldsymbol{e}^q$ in the question and the relation type t, we aim to locate the tail event $\boldsymbol{e}^p$ in the text passage to generate the desired answer. Inspired by knowledge embedding learning methods such as TransE (2013), we propose to transform event embeddings using a transformation matrix and introduce an auxiliary task for event relation type classification. Using the transformation matrix has two advantages. First, the token-level hidden states are preserved in the original embedding space which are important for semantic-based QA. Second, the transformed event embeddings allow the identification of common features for more general event relation type classification in the new event embedding space. In what follows, we describe our proposed invertible event transformation operator, contrastive learning, and event relation type classification in more detail.

3.2.1 Invertible Event Transformation

We propose an invertible transformation which aims to map event representations onto a new event embedding space in which the desired event semantic relations are inherently encoded. Let $H_{q|e} \in \mathbb{R}^{C_q \times d}$, $H_{a|e} \in \mathbb{R}^{C_a \times d}$ and $H_{o|e} \in \mathbb{R}^{C_o \times d}$ be part of the hidden state vectors H representing the embeddings of the question event, the answer events and other events in the text passage, in which C_q refers to the number of event triggers in the question, and the sum of C_a and C_o refers to the total number of event triggers in the text passage. Additionally, let $M \in \mathbb{R}^{d \times d}$ be the transformation matrix, $H_{q|e}, H_{a|e}$ and $H_{o|e}$ are mapped onto a new event embedding space by: $H^{'}_{q|e} = MH_{q|e} + b_M$, $H^{'}_{a|e} = MH_{a|e} + b_M$, and $H^{'}_{o|e} = MH_{o|e} + b_M$, where $b_M \in \mathbb{R}^d$ is the bias term. The singularity of the random matrix can be guaranteed by (Tao and Vu, 2008) with high probability (confirmed by our experimental results as well). Therefore, we do not need any regularisation terms to guarantee the rank of the transformation matrix in the training process. Since the linear transformation is invertible, we have the following properties (the proof can be found in Appendix A),

Property 1. For any event representation e obtained from a PLM, and its transformed new embedding e', we have $S(e') = S(e) + \log(|M|)$, where S is the entropy of a given event.

Property 1 guarantees that the projected representation of a given event has a smoother distribution which makes it easier to find a separatrix in a hyperspace in the auxiliary task of event relation type classification, since $|M|$ is usually large than 1. The distribution of outliers, i.e., low frequency words, will be smoothed by this invertible transformation as well.

Property 2. For any event representation pair e_1 and e_2 obtained from a PLM, and their transformed representations e'_1 and e'_2, we have $I(e'_1, e'_2) = I(e_1, e_2)$, where I is the mutual information of the given event pair.

Property 2 guarantees that for any event pairs, the projection will not change the mutual information, which represents event relations encoded in the original PLM. Since the projection is a bijection and invertible, the separatrix from the learned auxiliary task will be converted to the hidden states to guide the answer generation directly.

3.2.2 Contrastive Learning

After mapping event representations onto a new event embedding space by the aforementioned invertible event transformation, we can then form positive event pairs (h_i, h_j) by selecting the transformed question event h_i from $H^{'}_{q|e}$ and the transformed answer event h_j from $H^{'}_{a|e}$. We can also form negative event pairs (h_i, h_k) and (h_k, h_j) by randomly sampling h_k from the transformed event vectors of other events $H^{'}_{o|e}$. Let L_{cl} denote the loss of contrastive learning:

$$L_{cl} = \frac{1}{Z} \sum_{i=1}^{|H^{'}_{q|e}|+|H^{'}_{a|e}|} \sum_{j=1}^{|H^{'}_{q|e}|+|H^{'}_{a|e}|} [l_{cl:(i,j)} + l_{cl:(j,i)}] \quad (1)$$

where $l_{cl:(i,j)}$ denotes the loss for positive pair on event vectors h_i and h_j, $l_{cl:(i,j)} = -\log[\exp(\cos(h_i, h_j)/\tau)/s_i]$, $\cos(\cdot)$ denotes the cosine similarity function, s_i denotes the sum of cosine similarity of the positive event pair (h_i, h_j) and that of negative event pairs (h_i, h_k), τ is the temperature hyperparameter to adjust the penalty of negative pairs, and $Z = 2(|H^{'}_{q|e}| + |H^{'}_{a|e}|)$. L_{cl} sums over the contrastive loss of all possible event pairs in the training set.

`TranCLR` takes question event vector and answer event vectors as the source of positive pairs, while takes other event vectors as the source of negative events. The purpose of contrastive learning is to better employ the event information as hint for the QA task. Therefore, a good transformation matrix is essential. We introduce an auxiliary event relation type classification task in order to train a better transformation matrix.

3.2.3 Event Relation Type Classification

As shown in the analysis of the n-gram word and token statistics conducted on the ESTER dataset (Han et al., 2021), the questions already encode sufficient information to detect the type of event relations referred. Therefore, the idea is to apply the same transformation matrix, used on the event vectors, also on the hidden vectors encoding the question, and then use the results for event relation type classification. We first predict the event relation type, $\hat{t}$ by feeding the tranformed question vector to a feed-forward layer and a softmax layer. We then define the cross entropy loss of event relation type classification, denoted as L_{tc}:

$$L_{tc} = -\sum_{n=1}^{N} t_n log\hat{t_n} \quad (2)$$

3.2.4 Final Objective Function

For answer generation, the model operates on the hidden state vectors H in the original embedding space encoded by UnifiedQA-T5-large (or RoBERTa-large) to generate (or extract) the answer(s), $\hat{\boldsymbol{y}}$. Let L_{qa} denote the loss of the main question answering task:

$$L_{qa} = -\frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}_i \log \hat{\boldsymbol{y}}_{\boldsymbol{i}} \quad (3)$$

where $T = N_a + A - 1$ is the total token length of A ground truth answers separated by $A - 1$ ";" special tokens under the generative setting, while $T = \boldsymbol{x}^p$ is the total token length of the supporting paragraph $\boldsymbol{x}^p$ under the extractive setting. For the latter, we further extract all tokens marked as "BI" or "I" predictions as answers. The final loss is defined as:

$$L = L_{qa} + \lambda_{tc}L_{tc} + \lambda_{cl}L_{cl} \quad (4)$$

where λ_{tc}, λ_{cl} are hyperparameters to control the contribution of individual loss terms.

4 Experiments

In this section, we will first introduce the experimental setup including the dataset used and the hyperparameter setting, followed by the discussion of experimental results and ablation studies.

4.1 Experimental Setup

Dataset We use the event-centric QA dataset, ESTER (Han et al., 2021), for our experiments. The dataset contains 6k human-annotated event-centric questions with an average length of 10 tokens over 1.9k paragraphs with a maximum of 340 tokens. All event triggers have been marked over the questions, paragraphs and answers. Besides, the dataset provides the event type label for each question from the five common event relation types: *Causal*, *Conditional*, *Counterfactual*, *Sub-event*, and *Co-reference*, and collects over 10k event relation pairs. Each of the aforementioned event relation types contribute to 43.1%, 21.3%, 7.1%, 15.6% and 12.9% of questions, respectively. Most of the questions have 1-2 in-paragraph answers, while the *Sub-event* type questions have more than 3 answers on average. ESTER has been officially split into the training, development and test sets, with 4,547, 301 and 1,170 instances, respectively.[2] Table A1 in Appendix B reports the statistics of 5 event types in ESTER.

Evaluation Metrics We use the same metrics introduced in ESTER (Han et al., 2021): F_1^T, $HIT@1$ and EM defined for the multi-answer setting. F_1^T calculates unigram-level token overlap between generated answers and the ground truth answers, $HIT@1$ measures whether the leftmost answer contains a correct event trigger, and Exact Match EM checks if any predict answer matches exactly the corresponding ground truth answer.

Baseline The baselines we use are the seq2seq pipeline built on the UnifiedQA-T5-large and the RoBERTa-large models introduced in ESTER (Han et al., 2021).[3] Hyperparameter setting for our models can be found in Appendix B.

4.2 Results

4.2.1 Overall Comparison

Model	F_1^T	$HIT@1$	EM
Generative setting			
UnifiedQA-large (Han et al., 2021)	66.8	**87.2**	24.4
UnifiedQA-large (our run)	65.8	86.7	24.6
UnifiedQA-large TranCLR	**74.2**	86.4	**25.6**
UnifiedQA-large TranCLR (-prefix)	69.6	81.4	21.6
UnifiedQA-large TranCLR (-TC)	**74.6**	**87.4**	24.6
UnifiedQA-large TranCLR (-CL)	72.8	84.7	**25.6**
UnifiedQA-large TranCLR (-TransM)	66.8	77.7	20.3
Extractive setting			
RoBERTa-large (Han et al., 2021)	68.8	66.7	16.7
RoBERTa-large (our run)	67.0	69.4	17.9
RoBERTa-large (IO)	73.7	77.4	15.3
RoBERTa-large (IO) TranCLR	**74.7**	**80.4**	**18.3**

Table 1: Main results of experiments on ESTER dataset. (Han et al., 2021) takes "B-I-O" labels for extractive QA, while we found "I-O" labels work better. UnifiedQA-large TranCLR (-*) refer to ablation studies for generative QA, where -prefix, -TC, -CL and -TransM refer to the removal of the event type label prefix, question-answer event type classification, contrastive learning, and the transformation matrix, respectively.

[2] As only the training set and the development set have been released, we fine-tune our model on the training set and evaluate on the development set.

[3] As the event-centric QA task has only been recently introduced, no other approach has been proposed for ESTER.

	UnifiedQA-large (our run)			UL TranCLR			UL TranCLR (-TC&CL)			UL TranCLR (-prefix)		
Type	F_1^T	$HIT@1$	EM	F_1^T	$HIT@1$	EM	F_1^T	$HIT@1$	EM	F_1^T	$HIT@1$	EM
Causal (39.2%)	72.8	**90.7**	**31.4**	**80.5**	89.8	30.5	71.4	88.1	**31.4**	75.0	82.2	25.4
Conditional (19.3%)	58.7	84.5	19.0	66.6	86.2	**24.1**	63.5	**89.7**	22.4	**69.2**	86.2	19.0
Counterfactual (9.3%)	65.7	**78.6**	35.7	**75.9**	**78.6**	32.1	65.1	75.0	**39.3**	65.2	67.9	28.6
Sub-event (19.6%)	59.0	89.8	**13.6**	**70.6**	89.8	**13.6**	66.2	**93.2**	**13.6**	65.0	89.8	11.9
Co-reference (12.6%)	65.2	**79.0**	21.1	**70.6**	76.3	**26.3**	65.8	76.3	23.7	63.6	68.4	23.7
All (100%)	65.8	**86.7**	24.6	**74.2**	86.4	25.6	67.6	**86.7**	**25.9**	69.6	81.4	21.6

Table 2: Results from various models on 5 different event relation types on the development set. UL refer to the abbreviation of UnifiedQA-large. UL TranCLR outperforms UnifiedQA-large baseline significantly in F_1^T across all event relation types. The ablated versions of UL TranCLR show mixed results in $HIT@1$ and EM.

We show the overall evaluation results in Table 1. Our model achieves impressive results compared with the previous baseline, gaining about 8% improvement in F_1^T under the generative setting, 3.0% improvement in EM and $HIT@1$ scores under the extractive setting.[4] We have the following observations: (1) event-based contrastive learning brings a significant gain, enabling a better reasoning of semantic relations between event triggers in questions and candidate answers in text, since the question event and the answer event, although bearing very different semantic meanings, are pushed closer in their projected new event embedding space. This is evident from the drastically improved F_1^T score in the main QA task; (2) both event relation type classification and contrastive learning are indispensable since the combination of them achieves more balanced results across all metrics in answer evaluation, showing that the auxiliary event relation type classification task leads to a better learned transformation matrix; (3) prompt-based learning using the event relation type as prefix of input[5] is effective as the additional information can better guide the model to answer questions which are much more difficult than traditional factoid questions; and finally (4) the use of the transformation matrix makes it possible to simultaneously learn representations in both the original embedding space and the new event-centric space, leading to better QA results.

4.2.2 Zero-shot and Few-shot Learning

In this section, we assess the ability of TranCLR for zero-shot and few-shot learning, i.e., without the training data or with very few training instances. Figure 3 shows the F_1^T and EM values of TranCLR and the baseline UnifiedQA-large with the varying size of the training set. It can be observed that in zero-shot learning, both models give similar EM results and TranCLR slightly outperforms UnifiedQA-large in F_1^T. With only 500 training instances, TranCLR is able to generate more accurate answers, beating UnifiedQA-large by 3% in EM, demonstrating the benefit of making effective use of event information for better reasoning of event semantic relations. The performance gap however gradually diminishes with the increasing size of the training set. Nevertheless, TranCLR is able to generate answers containing more overlapped information with the ground truth with more training instances, evidenced by the increased performance gains compared to UnifiedQA-large, reaching nearly 8% in F_1^T when using the full training set. This shows that our proposed contrastive learning combined with the auxiliary task of event semantic type classification can better capture event semantic knowledge which guides the decoder to generate answers closer to the ground truth.

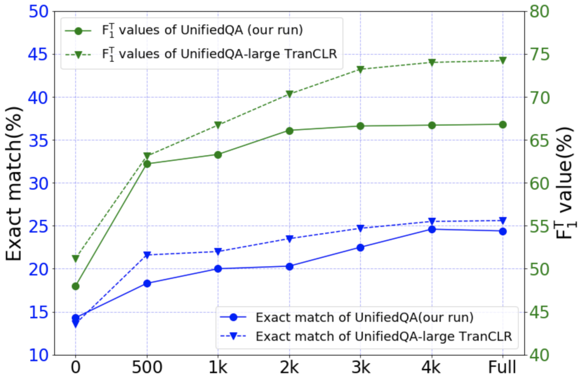


Figure 3: Zero-shot and Few-shot learning performance of UnifiedQA-large TranCLR and UnifiedQA-large.

4.2.3 Results per Event Relation Type

In Table 2, we provide detailed comparison of results under various event relation types. We can

[4] Evaluation code: https://github.com/PlusLabNLP/ESTER/tree/master/code

[5] During inference, event relation type is detected automatically from a given question.

Paragraph: Kopp also has been **linked, through DNA testing, to a strand of hair found** near where the sniper fired, law enforcement sources have said.\n Nicknamed the “Atomic Dog” in anti-abortion circles, Kopp had been arrested in several states since 1990 for protesting abortion. His car was spotted in Slepian’s neighborhood in the weeks before the shooting, and was found abandoned at the Newark, N.J., airport in December.\n The filing of the charges has intensified interest in Kopp as a suspect in three non-fatal sniper attacks on Canadian abortion providers, and one near Rochester, N.Y. , between 1994 and 1997.\n Kopp is now the second anti-abortion activist being sought by the FBI as a suspect in a fatal attack.\n
Question: What actions from the law enforcement could lead to the filing of the charges against Kopp? **Ground Truth Answer**: linked, through DNA testing, to a strand of hair found **Question-answer event relation type**: *Conditional*
UnifiedQA-T5-large (our run): 1: arrested in several states since 1990 for protesting abortion; 2: his car was spotted in slepian’s neighborhood; 3: was found abandoned at the newark, n.j., airport in december **UnifiedQA-T5-large TranCLR**: 1: arrested in several states since 1990; 2: link, through dna testing; 3: strand of hair found near where the sniper fired

Table 3: Example answers generated by different models. `TranCLR` injected with event knowledge accurately grab the news narrative, and generate answers that cover 100% content of ground truth. In contrast, UnifiedQA-T5 without event-related learning is confused with question-related context information in the paragraph.

observe that the results on the ‘*Causal*’ type, being the largest category, are much better compared to other event relation types. Our proposed `TranCLR` achieves the best F_1^T scores across all event relation types compared to the baseline UnifiedQA-large, with the increment in the range of 5.4-11.6%. The largest performance improvement of 11.6% is observed on the most difficult ‘*Sub-event*’ type in which questions have more than 3 answers on average. By analyzing the results, we found that it is sometimes quite difficult to distinguish between ‘*Conditional*’ and ‘*Counterfactual*’ types. As such, adding the event relation type as prefix may confuse the model. In terms of the $HIT@1$ results, `TranCLR` with prefix only (i.e., `-TC&CL`) improves upon the baseline by over 5% and nearly 4% for the ‘*Conditional*’ and the ‘*Sub-event*’ types respectively. We also observe that using the event relation type as prefix in prompt-based learning is very effective in boosting the EM scores, especially for the ‘*Counterfactual*’ type in which nearly 4% improvement is obtained compared to UnifiedQA-large. For the ‘*Sub-event*’ type where multiple answers are expected, there is no improvement in EM in our models compared to the baseline.

4.2.4 Visualisation of Event Embeddings

In Figure 4[6], we visualize the learned event embeddings in the semantic space during the training. It can be observed that with the increasing number of training epochs, event triggers are grouped into few clusters from an evenly distributed initial state. Most irrelevant event nodes are pushed aside as they are used as negative samples in contrastive learning, while question events and answer events are pulled together. The visualization reveals reasonable process of event knowledge distillation.

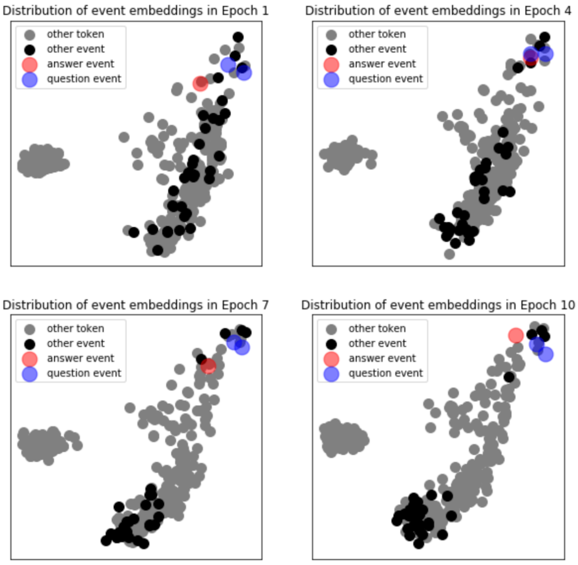


Figure 4: Distributions of events in the semantic space. Question events, answer events, other events and the remaining non-event tokens are shown in blue, red, black and gray, respectively.

4.3 Qualitative Analysis

We further perform qualitative analysis in Table 3 over the example illustrated in Figure 1 and 2[7]. All models generate more than one answers. More concretely, `TranCLR` manages to generate text covering 100% of the ground truth answer, while UnifiedQA-large is unable to generate coherent answers as the model failed to detect the semantic relations between the question event trigger "*charges*" and the answer event triggers "*link*", "*testing*" and "*found*".

4.4 Generalization Evaluation

To further evaluate the generalization capability of our event knowledge distillation paradigm, we apply the `TranCLR` model trained on ESTER for

[6] Better viewing in color.

[7] More example outputs are presented in Appendix C.

Model	F_1	EM
RoBERTa-large	10.0	0.0
RoBERTa-large (Han et al., 2021)	20.0	4.1
RoBERTa-large TranCLR (ESTER)	**28.7**	**15.6**

Table 4: Zero-shot inference results on the TORQUE development set. RoBERTa-large (the first result row) is not trained on any event QA data, while the other models are trained on ESTER only.

zero-shot inference on unseen QA data focusing on event temporal relations, using the TORQUE dataset (Ning et al., 2020b). TORQUE focuses on questions about event temporal relations such as *"what happened before/after [some event]?"*. For each question, the dataset provides a two-sentence supporting passage and passage event annotations. The answers are simply event mentions in the form of words/phrases, rather than longer text spans as in the ESTER dataset. We perform zero-shot inference on TORQUE without fine-tuning the models on its training set. It can be observed from Table 4 that compared with RoBERTa-large without trained on any event QA data (first result row), fine-tuning RoBERTa-large on ESTER (second result row) improves F_1 by 10%. Nevertheless, our proposed TranCLR exhibits a significantly better event understanding capability, achieving 8.7% and 11.5% further gains in F_1 and EM scores, respectively. It is worth mentioning that the ESTER dataset does not contain any questions about event temporal relations. The results show the strong generalization capabilities of TranCLR and further verify the effectiveness of our proposed framework for event semantic reasoning.

5 Conclusions

In this paper, we have proposed a novel framework, called TranCLR, to tackle the event-centric QA task on the ESTER dataset (Han et al., 2021). The core idea of TranCLR is to effectively explore the event knowledge in both questions and context through event-centric contrastive learning and the auxiliary task of event type classification. Our experimental results show superior performance of TranCLR on event-centric QA compared to the strong baseline, gaining 8.4% and 3% absolute improvements in F_1^T and EM scores respectively. Further zero-short inference and qualitative analysis verify the promising event semantic understanding and reasoning capability of our model.

Limitations

Although we have verified the promising event semantic understanding and reasoning capability of TranCLR trained on ESTER for both in-domain event semantic relations and the out-of-domain event temporal relation, it is worth further exploring whether the model indeed captures event semantic relations and does not just generate answers by the matching of spurious patterns. Adversarial attacks could be explored in the future to assess the possible backdoor of the model in order to evaluate its robustness (Tan et al., 2021; Pergola et al., 2021a; Bartolo et al., 2021).

Our current work is built on the ESTER dataset where each question is paired with a single paragraph. In reality, event-centric QA may require the gathering of evidence scattered over multiple paragraphs and reasoning over more sophisticated event chains or graphs. Such complex event semantic relations is beyond what our proposed event-centric contrastive learning could capture. To develop new methodologies for dealing with more challenging event-centric QA, efforts need to be devoted to develop a dataset under a more realistic setting.

Acknowledgement

This work was funded in part by the the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/1, EP/V048597/1, EP/X019063/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1).

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. *arXiv preprint arXiv:2004.10157*.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. *arXiv preprint arXiv:2104.08678*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Nuo Chen, Linjun Shou, Ming Gong, and Jian Pei. 2022. From good to best: Two-stage training for cross-lingual machine reading comprehension. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10501–10508.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Pradeep Dasigi, Sameer Singh, and Matt Gardner. 2021. Learning with instance bundles for reading comprehension. *arXiv preprint arXiv:2104.08735*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Simon Gottschalk and Elena Demidova. 2019. Eventkg–the hub of event knowledge on the web–and biographical timeline generation. *Semantic Web*, 10(6):1039–1070.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. Ester: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. 2020. CHIME: Cross-passage hierarchical memory network for generative review question answering. pages 2547–2560.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020a. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020b. Torque: A reading comprehension dataset of temporal ordering questions. *arXiv preprint arXiv:2005.00242*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Gabriele Pergola, Lin Gui, and Yulan He. 2021a. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2870–2883, Online. Association for Computational Linguistics.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021b. Boosting low-resource biomedical QA via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985, Online. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2020. Erica: improving entity and relation understanding for pre-trained language models via contrastive learning. *arXiv preprint arXiv:2012.15022*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Chao Shang, Peng Qi, Guangtao Wang, Jing Huang, Youzheng Wu, and Bowen Zhou. 2021. Open temporal relation extraction for question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 3157–3164, New York, NY, USA. Association for Computing Machinery.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terence Tao and Van Vu. 2008. On the singularity probability of random bernoulli matrices.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Appendix

A Proof of Properties of the Invertible Transformation

Definition: For any event representation e obtained from a Pre-trained Language Model (PLM), let e' be the transformed representation, $e' = M \cdot e + b$, where M is the transformation matrix and b is a bias, the projection of linear transformation is invertible.

Property 1 For any event representation e obtained from a PLM, and its transformed representation e', we have $S(e') = S(e) + \log(|M|)$, where S is the entropy of the given event.

Proof. Assume that M is an identity matrix, we then have $e' = e + b$. Hence,

$$\begin{aligned} S(e') &= -\int p_{e'}(e+b)\log p_{e'}((e+b)d(e+b) \\ &= -\int p_e(e)\log p_e(e)d(e) \\ &= S(e), \end{aligned}$$

where the p_e and $p_{e'}$ represent the probability space before and after transformation. Therefore, the bias term will not change the entropy after projection. Then, we only need to consider a general transformation matrix M,

$$\begin{aligned} S(e') &= -\mathbb{E}[\log p_{e'}(M \cdot e)] \\ &= -\mathbb{E}[\log(M^{-1}p_e(M^{-1} \cdot e'))] \\ &= -\mathbb{E}[\log(M^{-1}p_e(e)] \\ &= S(e) + \log(|M|) \end{aligned}$$

□

Property 2 For any event representation pair e_1 and e_2 obtained from a PLM, and their transformed representations e'_1 and e'_2, we have $I(e'_1, e'_2) = I(e_1, e_2)$, where the I is the mutual information of the given event pair.

Proof.

$$\begin{aligned} I(e'_1, e'_2) &= S(e'_1) - S(e'_1|e'_2) \\ &= S(e'_1) - S(e'_1|e'_2) + \log(|M|) - \log(|M|) \end{aligned}$$

According to the proof of **Property 1**, we have $S(e) = S(e') - \log(|M|)$. Hence,

$$\begin{aligned} I(e'_1, e'_2) &= S(e'_1) - \log(|M|) - (S(e'_1|e'_2) - \log(|M|)) \\ &= S(e_1) - S(e_1|e_2) \\ &= I(e_1, e_2) \end{aligned}$$

□

B Experimental Setting

Hyperparameters Our hyperparameter setting follows what has been reported in Han et al. (2021).

For generative setting, the hidden size of Unified-T5-large is 1,024 and the corresponding vocabulary size is 32,128. The random seed is 5. The batch size is set to 2 and the accumulation steps 3 on 2 quadro_rtx_6000 GPUs. The optimizer of all models is BertAdam[8] with $\beta 1 = 0.9, \beta 2 = 0.999$, and ϵ=1e-6. Except for parameters of weights in layer normalization and bias in all layers, all other trainable parameters are decayed with a rate of 0.95 during training. The learning rate is increased linearly from 0 to 5e-5 in the first 10% total training steps and then linearly decreased to 0.

Similarly, for extractive setting, the hidden size of Roberta-large is 1,024 and the vocabulary size is 50,265. The random seed is 23. The batch size is 8 with accumulation steps 2 on same 2 quadro_rtx_6000 GPUs. The optimizer and decaying strategy remain same as generative models. The learning rate is changed to 1e-5. In addition, following Han et al. (2021), we adopt label weight 4 for "B" and "I" label to reduce label unbalance.

For other hyperparamters, we set empirically τ to 1.0 in contrastive learning, and λ_{tc} = 0.1, λ_{cl} = 0.1 in Eq. (4). It takes around 3 hours to fine-tune our models for 10 epochs. Parameter amounts are 356M and 738M for extractive and generative settings respectively.

Statistics of the ESTER Dataset Table A1 shows the data statistics per event relation type. The *Causal* event relation type is the largest category, while *Counterfactual* being the smallest one.

Type	Train	Dev	Test
Causal	2047	118	431
Conditional	928	58	289
Counterfactural	294	28	106
Sub-event	678	59	204
Co-reference	600	38	140
All	4547	301	1170

Table A1: The statistics of 5 event types in the ESTER dataset. We only use the training and development set. The test set is not published.

C More Generated Answers

[8] https://github.com/google-research/bert/blob/master/optimization.py

Paragraph: Another leftist South American nation, Bolivia, has also expressed a desire to join OPEC despite its modest oil production of 40,000 barrels per day.\n The short-term outlook in terms of OPEC's influence on oil prices is likely to depend on the discipline of the cartel and the actual reduction of supplies to the market.\n The cartel decided in October to reduce its output by 1.2 million bpd from the beginning of November, but analysts believe the real reduction has been only 500,000-800,000 bpd because of cheating by some members. \n The cut of 500,000 bpd in February would reduce the output from OPEC members, excluding Iraq and Angola, to 25.8 million bpd in principle.\n Some analysts expressed concern that any reduction in supplies could send prices higher in the months ahead, the peak time for **oil demand because of the northern hemisphere winter**.\n

Question: Why could a reduction in supplies send prices higher in the months ahead?

Answer 1: oil demand because of the northern hemisphere winter

Question-answer event relation type: Conditional

UnifiedQA-T5-large(our run): 1: the peak time for oil demand because of the northern hemisphere winter

UnifiedQA-T5-large TranCLR: 1: the peak time for oil demand because of the northern hemisphere winter

---

Paragraph: Lowe's decision to bring former England rugby world cup-winning coach Clive Woodward into the backroom staff was one of the reasons for his unhappiness.\n "As a bloke I got on with him but I have to say **the decision to bring him in was bizarre**. **The relationship between me and Clive was never going to work** because there were too many people undermining the structure Rupert Lowe wanted at the club," he added.\n In an open letter by Mandaric published in the Sunday Mirror the Pompey chairman said: "If I'm honest I **never wanted Harry to leave in the first place**.\n "Of all the candidates Harry is the one which that stands out. The supporters have to trust me."\n

Question: Why was Mandaric unhappy with Lowe bringing Clive onto the staff?

Answer 1: never wanted Harry to leave in the first place

Answer 2: The relationship between me and Clive was never going to work

Answer 3: the decision to bring him in was bizarre

Question-answer event relation type: Causal

UnifiedQA-T5-large(our run): 1: too many people undermining the structure rupert lowe wanted at the club

UnifiedQA-T5-large TranCLR: 1: the decision to bring him in was bizarre

---

Paragraph: The visiting U.S. Assistant Secretary of State Richard Boucher on Tuesday said the United States did not have any involvement in the attack on a religious school in Pakistan's tribal region.\n "The **Pakistani government has said they carried out** initiative to deal with serious threats from fighters who were in that location," Boucher told reporters at U.S. embassy.\n "The **Pakistani government says it has carried out this action**. And it was necessary because militants, terrorists created a training center," he said. \n He supported Pakistan's policy to engage tribal elders to establish peace in the tribal region. \n

Question: What might have made Richard Boucher say United States did not have any involvement in the attack?

Answer 1: Pakistani government has said they carried out

Answer 2: Pakistani government says it has carried out this action

Question-answer event relation type: Conditional

UnifiedQA-T5-large(our run): 1: the pakistani government has said they carried out initiative to deal with serious threats

UnifiedQA-T5-large TranCLR: 1: the pakistani government has said they carried out initiative

---

Paragraph: It was Chamara Silva who primarily kept the scoreboard ticking over with a 68-run stand for the fifth wicket with Mahela Jayawardene, and an unbeaten 57-run partnership with Prasanna Jaywardene.\n At the close, Silva was unbeaten on 79, his second half century of the match after failing to score in the first Test, while Jayawardene was not out 22.\n Lasith "Slinga" Malinga and Muttiah Muralitharan had earlier bowled Sri Lanka to a 138-run first innings lead.\n The New Zealand batsmen had no answer to the hostile pace and slinging action of Malinga at one end, and could not read Muralitharan's spin at the other, as they crumbled to be all out before lunch on the second day for 130.\n Of the New Zealand batsmen only Brendon McCullum put up any solid resistance. He was dropped on the first ball of the day without scoring and went on to post 43 before he was bowled by Muralitharan to end the innings.

Question: What led to the crumbling of New Zealand on the second day?

Answer 1: hostile pace and slinging action of Malinga

Answer 2: could not read Muralitharan's spin

Question-answer event relation type: Causal

UnifiedQA-T5-large(our run): 1: hostile pace and slinging action of malinga

UnifiedQA-T5-large TranCLR: 1: hostile pace and slinging action of malinga at one end; 2: could not read muralitharan's spin

---

Paragraph: Heavy fighting resumed in central Somalia Wednesday after retreating Islamist fighters opened fire on a force of government and Ethiopian troops, officials and residents said, as the conflict in the Horn of Africa nation entered its second week.\n Hours after the UN Security Council failed to agree on the withdrawal of foreign troops, **Islamists fighters in trenches near the town of Jowhar opened fire** to stop Somali-Ethiopian troops from advancing further southwards.\n "Very heavy fighting has erupted outside Jowhar. The Islamic forces say they will keep fighting," said Mohamed Abdi Ali, a resident of the town about 90 kilometres (55 miles) north of the Islamist-controled capital of Mogadishu.\n "Ethiopians have not started using planes yet, but we do not rule that out," he added.\n

Question: What could be expected to happen after the Somali-Ethiopian troops tried to advance southwards?

Answer 1: Islamists fighters in trenches near the town of Jowhar opened fire

Question-answer event relation type: Conditional

UnifiedQA-T5-large(our run): 1: heavy fighting resumed in central somalia

UnifiedQA-T5-large TranCLR: 1: Islamists fighters in trenches

---

Table A2: Additional generated samples from the selected models. In the first case, both models generate overlong answers. In the second case, `TranCLR` manages to generate one completely correct answer while UnifiedQA-T5-large produced a wrong answer. For the next two cases, `TranCLR` controls answer range better in the third one and is able to cover both answers in the fourth one, compared with the UnifiedQA-T5-large baseline. In the last case, only `TranCLR` generates related answer.