Zero-Shot Cross-Lingual Document-Level Event Causality Identification with Heterogeneous Graph Contrastive Transfer Learning

Zhitao He[1,2], **Pengfei Cao**[1,2], **Zhuoran Jin**[1,2], **Yubo Chen**[1,2]
Kang Liu[1,2,3], **Zhiqiang Zhang**[4], **Mengshu Sun**[4], **Jun Zhao**[1,2*]
[1] The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Shanghai Artificial Intelligence Laboratory, China
[4] Ant Group, Hangzhou, China
{zhitao.he, pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

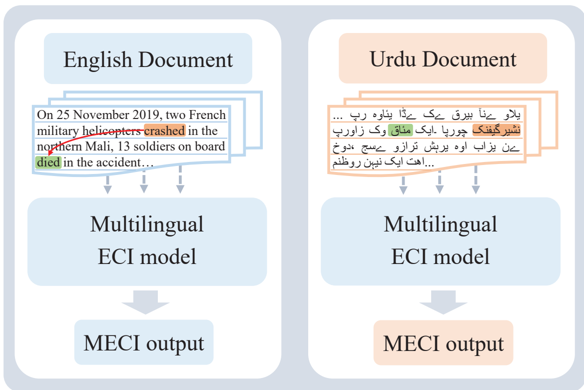
Event Causality Identification (ECI) refers to the detection of causal relations between events in texts. However, most existing studies focus on sentence-level ECI with high-resource languages, leaving more challenging document-level ECI (DECI) with low-resource languages under-explored. In this paper, we propose a Heterogeneous **G**raph **I**nteraction Model with **M**ulti-granularity **C**ontrastive Transfer Learning (GIMC) for zero-shot cross-lingual document-level ECI. Specifically, we introduce a heterogeneous graph interaction network to model the long-distance dependencies between events that are scattered over a document. Then, to improve cross-lingual transferability of causal knowledge learned from the source language, we propose a multi-granularity contrastive transfer learning module to align the causal representations across languages. Extensive experiments show our framework outperforms the previous state-of-the-art model by 9.4% and 8.2% of average F1 score on monolingual and multilingual scenarios respectively. Notably, in the multilingual scenario, our zero-shot framework even exceeds GPT-3.5 with few-shot learning by 24.3% in overall performance.

Keywords: zero-shot cross-lingual, document-level, event causality identification

1. Introduction

Event Causality Identification (ECI) is an important task in natural language processing (NLP), which can facilitate various applications, including explainable question answering (Yang et al., 2018b), intelligent search (Rudnik et al., 2019) and complex reasoning (Dalvi et al., 2021). Most previous methods (Kadowaki et al., 2019; Liu et al., 2020a; Zuo et al., 2021; Cao et al., 2021) focus on sentence-level with English corpora. Nevertheless, a substantial number of causal relations are expressed by multiple sentences. For instance, there are approximately 68.7% of causal relationships in English corpora (Caselli and Vossen, 2017) are attributed to inter-sentence event pairs. Hence, identifying causality of events at the document-level is necessary, which gains increasing attention recently (Tran Phu and Nguyen, 2021; Fan et al., 2022; Boxi Cao et al., 2024; Chen et al., 2022).

However, training an ECI model typically relies on a large amount of data, especially for document-level, which makes it hard to adapt to low-resource languages. While pre-trained language models have exhibited remarkable capabilities across various tasks (Brown et al., 2020;



(a) Training **(b) Zero-shot Testing**

Figure 1: Zero-shot cross-lingual document-level event causality identification. Data is non-parallel.

Chowdhery et al., 2022; Wang et al., 2023; Tan et al., 2023; Liang Zhang et al., 2023), they still struggle in multilingual setting, as evidenced in recent studies (Chang et al., 2023; Huang et al., 2023; Zhang et al., 2023). Even the powerful ChatGPT exhibits quite limited performance in multilingual relation prediction task, with an average accuracy of only 37%. Particularly for low-resource languages,

*Corresponding author

LREC-COLING 2024, pages 17833–17850

the model performs even worse, such as achieving only a 6.3% accuracy in Urdu (Lai et al., 2023). Moreover, low-resource languages face a critical shortage of training data, making it challenging to enhance the document-level ECI performance of language models. Therefore, in this paper we focus on zero-shot cross-lingual document-level ECI, aiming to efficiently transfer the causality knowledge from the source languages to any other languages (e.g. low-resource/less-studied languages) under limited language resources. As shown in Figure 1, the model is trained with the annotated data in source language and directly applied to target language. While significant efforts have been made for monolingual setting (Tran Phu and Nguyen, 2021; Chen et al., 2022), two critical challenges arise when we apply monolingual ECI to zero-shot cross-lingual document-level setting:

(1) **Language-agnostic causal knowledge alignment**. Each language has its own characteristics. The multilingual ECI models, trained in source language, inevitably tend to learn language-specific knowledge rather than pure language-agnostic knowledge (i.e., causal knowledge). Thus, the trained ECI model may only perform well in source language. For example, different languages have distinct distributions of distance between causal events. According to statistics, the majority distance between causal events in English corpora is 40 words, whereas in Turkish, it is 33 words, and in Danish, it is 23 words. Furthermore, based on our error analysis of vanilla pre-trained langugae model, when we employ model trained on English to test on Danish, more than half (53%) of the false positive causal events are separated by more than 23 words, which means the ECI model trained in English corpora tend to predict a causal relation between two events that are far away, leading negatively impact on the prediction in Danish.

(2) **Causal events scattering**. As shown in Figure 2(a), the intra-sentence causality between "crashed" and "died" can be easily predicted. By contrast, event "crashed" and "killing" are located in sentence 1 and 5 respectively, the long distance and the interference of irrelevant events such as "Beirut barracks bombings" and "pursuing militants" make it difficult to directly model the long-distance dependencies between mentioned events. Fortunately, as shown in Figure 2(b), we can efficiently obtain the informative phrases (i.e., the underlined "military helicopter" and "soldier") related to scattered causal events "crashed" and "killing" by processing dependency structure with heuristic rules. We find that these informative phrases are similar and can serve as intermediate bridges for connecting the two events, which indicates the importance of informative phrases in modeling long-distance dependencies between events.

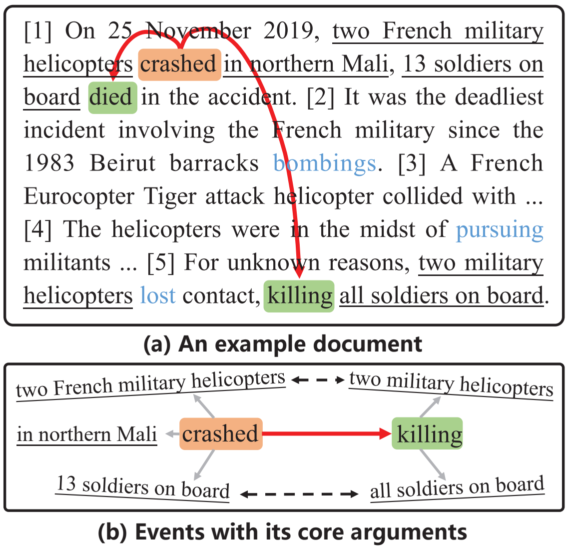


Figure 2: We show event-relevant sentences. Events are colored and we underline the informative phrases related to colored events. Red lines and gray lines indicate the causal relations and arguments of events, respectively.

To tackle the aforementioned challenges, we propose a Heterogeneous **G**raph **I**nteraction Model with **M**ulti-granularity **C**ontrastive Transfer Learning (GIMC). Specifically, we introduce a ***Multi-granularity Contrastive Transfer Learning Module*** to align causal representations across languages for transferring language-agnostic causal knowledge in statement and aspect levels. Meanwhile, we design a ***Heterogeneous Graph Interaction network*** with informative phrase nodes, sentence nodes, statement nodes, event pair nodes to model the long-distance dependencies between events. Extensive experiments show that our zero-shot framework outperforms previous models and even exceeds GPT-3.5 with zero/few-shot prompt.

We summarize the contributions as follows:

- We propose a novel heterogeneous graph interaction model with multi-granularity contrastive transfer learning (GIMC) to simultaneously address document-level and zero-shot cross-lingual event causality identification.
- We introduce a multi-granularity contrastive learning module to facilitate the cross-lingual transfer of language-agnostic causal knowledge, and construct a heterogeneous graph interaction network with four kinds of semantically rich nodes to model long-distance dependencies between events.
- Extensive experiments on the widely used multilingual ECI dataset show the effectiveness of our proposed model. F1 scores are improved

by an average of 9.4% and 8.2% in monolingual and multilingual scenarios.

2. Methodology

Figure 4 shows the architecture of GIMC which consists of a heterogeneous graph interaction network (left) and a multi-granularity contrastive transfer learning module on graph (right). We first encode the document using multilingual pre-trained language model, then construct the heterogeneous graph interaction network with four types of nodes. Finally, we leverage statement-level and aspect-level casual pattern contrastive learning to facilitate the cross-lingual transfer of causal knowledge.

2.1. Informative Phrase Extraction

Given a sentence $s = \{w_j\}_{j=1}^{|s|} \in \mathcal{D}$, we use the multilingual NLP toolkit Trankit (Nguyen et al., 2021), which has an overall performance of about 93% across different languages in sentence parsing, to obtain the dependency tree as shown in Figure 3(a). Previous studies only exploit the nodes in dependency tree as language-independent information to enhance ECI systems (Gao et al., 2019; Tran Phu and Nguyen, 2021) and overlook the rich semantics of dependency relations.

Thus, based on the semantics of the dependency relations (De Marneffe and Manning, 2008; De Marneffe et al., 2014; Schuster and Manning, 2016), we further process the dependency structure with heuristic rules to obtain a simplified dependency tree with informative phrases, as shown in Figure 3(b). We first analyze all dependency relations and their subtypes[1], retaining 19 semantically rich and indicative dependency relations, e.g., *nsubj* (nominal subject), *obj* (object), *obl* (oblique nominal). In this way, we extract the informative phrases and the corresponding dependency relations that indicates the relevant arguments of the events of interest (e.g., the subject (nsubj) of the event "$\mathrm{crashed}$" is "$\mathrm{two\ French\ military\ helicopters}$"). The complete list of dependency relations is in Appendix A.

2.2. Heterogeneous Graph Interaction Network

We build a heterogeneous graph interaction network $\mathcal{G}$ which contains informative phrase nodes, sentence nodes, statement nodes and event pair nodes. After inserting special tags "<t>" and "</t>" at the start and end of all events to mark the event positions, we transform each word $w_i \in \mathcal{D}$ into the embedding $\boldsymbol{x}_i$ using pre-trained language model.

[1] https://universaldependencies.org/u/dep/index.html

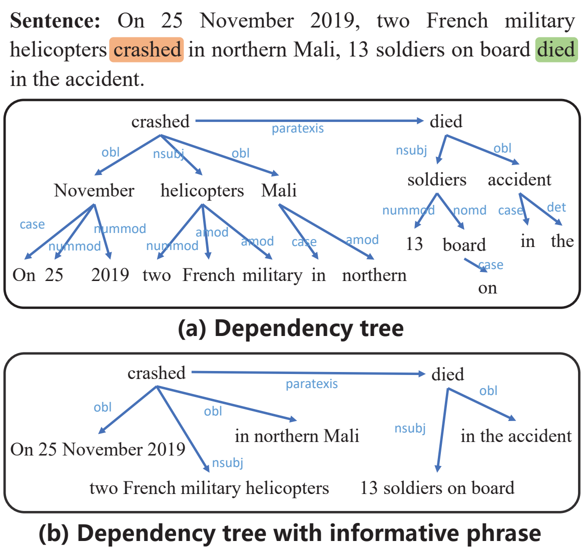


Figure 3: Example of informative phrase extraction

For each informative phrase node p which needs to incorporate its role information r_p, we initialize its embedding by $\boldsymbol{h}_p^{(0)} = \mathrm{Mean}(\{\boldsymbol{x}_j\}_{j\in p}) + \mathrm{Emb}(r_p)$. For each sentence node s, we initialize its embedding $\boldsymbol{h}_s^{(0)} = \mathrm{Mean}(\{\boldsymbol{x}_j\}_{j\in s})$. We define statement as a sentence containing two events to adapt to cross-sentence cases, such as the sentence 1 where the event "$\mathrm{crashed}$" and the event "died" are located, or a concatenation of two sentences to include event pairs that across sentences, such as event "$\mathrm{crashed}$" in sentence 1 and event "$\mathrm{killing}$" in sentence 5. We initialize statement node embedding $\boldsymbol{h}_{st}^{(0)} = \mathrm{Mean}(\{\boldsymbol{x}_j\}_{j\in st})$. For event pair (i, j) with their representation $(\boldsymbol{e}_i, \boldsymbol{e}_j)$, following Chen et al. (2022), we initialize $\boldsymbol{v}_{i,j}^{(0)} = \boldsymbol{W}_v[\boldsymbol{e}_i||\boldsymbol{e}_j]$, $\boldsymbol{W}_v$ are trainable parameters and $||$ denotes concatenation.

To capture the interactions among these nodes, we introduce six types of edges:

Phrase-Phrase Edge (P-P) The P-P edges are derived from the dependency structure. Informative phrase nodes are connected to each other by the dependency relations.

Sentence-Phrase Edge (S-P) The phrase node is connected to its sentence node. The S-P edges enable to model the local contextual information of a phrase in its corresponding sentence.

Phrase-Events Edge (P-E) The informative phrase containing an event mention can be seen as a more complete expression of the event. We model this information via the P-E edge.

Sentence-Events Edge (S-E) If there are sentences containing any events of the current event pair, the event pair node is connected to the nodes of those sentences. We model the local context information of events with S-E edges.

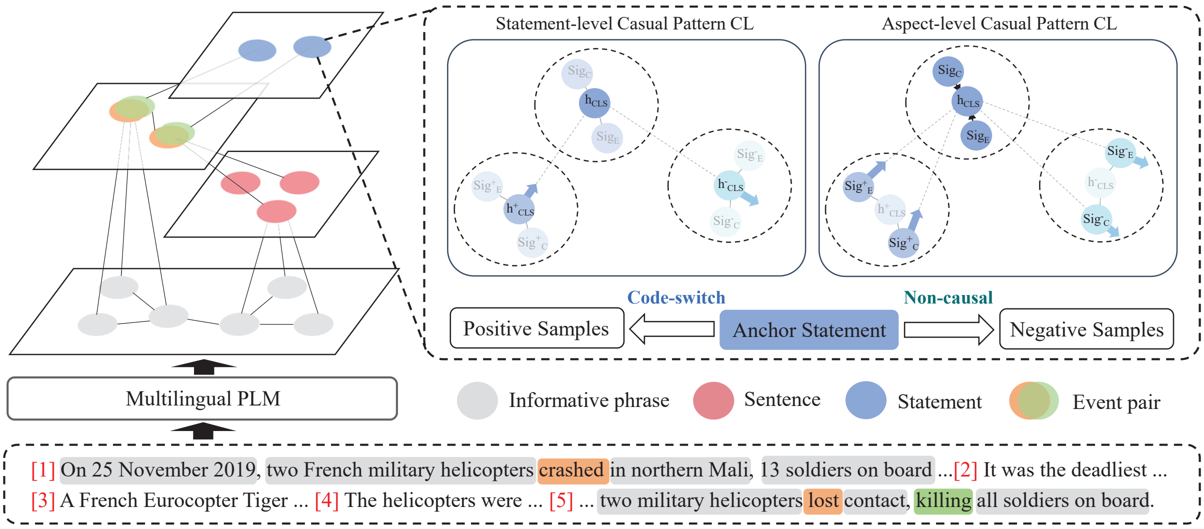


Figure 4: Overview of our proposed heterogeneous graph interaction network with multi-granularity contrastive transfer learning (GIMC) for zero-shot cross-lingual document-level ECI.

Statement-Events Edge (St-E) We connect the event node and its statement node. The St-E edge is expected to model the process of transfer the causality in statement to event pair to avoid the model from overfitting the causality in event pair.

Events-Events Edge (E-E) There is an E-E edge between two event pair nodes only if the two corresponding event pairs share at least one event. E-E edge can promote the effective transmission of the causal information in event pairs.

After construct heterogeneous graph, we apply Graph Attention Networks v2 (GATv2) (Brody et al., 2022) to model the global interactions. For each edge (j, i), the scoring function $f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ indicates the importance of the features of the neighbor $j \in \mathcal{N}_i$ to the node i:

$$f(\boldsymbol{h}_i, \boldsymbol{h}_j) = \boldsymbol{a}^\top \mathrm{LeakyReLU}(\boldsymbol{W} \cdot [\boldsymbol{h}_i || \boldsymbol{h}_j]) \tag{1}$$

where $\boldsymbol{W}$ is denoted as $[\boldsymbol{W}_l || \boldsymbol{W}_r], \boldsymbol{W}_l \in \mathbb{R}^{d' \times d}, \boldsymbol{W}_r \in \mathbb{R}^{d' \times d}, \boldsymbol{a} \in \mathbb{R}^{d'}$ are trainable parameters. $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are the representations of node i and j respectively. Then the attention function is defined:

$$\alpha_{ij} = \frac{\exp(f(\boldsymbol{h}_i, \boldsymbol{h}_j))}{\sum_{j' \in \mathcal{N}_i} \exp(f(\boldsymbol{h}_i, \boldsymbol{h}'_j))} \tag{2}$$

Then, we compute a weighted average of the transformed features of the neighbor nodes as the new representation of node i, using the normalized attention scores:

$$\boldsymbol{h}'_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}_r \boldsymbol{h}_j \tag{3}$$

We employ K separate attention heads and concatenate their outputs as the output of node i:

$$\boldsymbol{h}'_i = \boldsymbol{W}_o \Big(\Big\Vert_{k=1}^{K} \boldsymbol{h}'^k_i \Big) \tag{4}$$

where $\boldsymbol{W}_o$ are trainable parameters.

2.3. Multi-granularity Contrastive Learning Module

Inspired by sentence-level ECI studies (Liu et al., 2020a; Zuo et al., 2021), which leverage the context-specific causal patterns of statements, for example, we can leverage the causal pattern "The [EVENT] **generates** [EVENT] ..." to identify the causation between **traffic congestion** and **environmental pollution** in a new statement "The traffic congestion generates environmental pollution and economic loss". We generalize the causal pattern to the multilingual space, using contrastive learning to explicitly align causal representations across language. As shown in Figure 4, our contrastive transfer learning module consists of two part: 1) *statement-level* causal pattern contrastive learning to align statement causal representation across languages, 2) *aspect-level* causal pattern contrastive learning to align the representation between fine-grained causal patterns and statement. We select appropriate positive and negative samples for anchor statement:

Positive Samples Given each anchor causal statement, we follow Qin et al. (2020) to use MUSE bilingual dictionaries to generate multilingual code-switched statements as the positive samples. Unlike Qin et al. (2020) which randomly replace a word at a time, we operate on phrases in statements. Specifically, given a phrase, we first randomly select a bilingual dictionary (e.g., en-da). Then, we switch each word in the phrase one by one using that bilingual dictionary, which is intuitively expected to maintain more complete semantic information. Taking the informative phrase in Figure 3 as example, "$\mathrm{two\ French\ military\ helicopters}$" $\xrightarrow{da}$ "$\mathrm{to\ franske\ militær\ helikoptere}$".

Negative Samples To make the negative examples more discriminative, we select non-causal statements within the same document as negative examples that do not have any text overlap with anchor statement. Furthermore, as we expect the model to focus on transfering across languages, we generate multilingual negative examples by code-switching to augment the list of negative samples.

1) *Statement-level* causal pattern contrastive learning As Zuo et al. (2021) learn context-specific causal patterns from causal statements, we propose a statement-level causal pattern contrastive learning loss to explicitly align causal representations of anchor statement with the generated positive sample. Formally, this is formulated as:

$$\mathcal{L}_{\mathrm{S}} = -\sum_{j=1}^{n} log \frac{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{CLS}+})}{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{CLS}+}) + \sum_{k=0}^{K-1} s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^k_{\mathrm{CLS}-})}$$

where $s(\cdot)$ denotes dot product, n is the number of corresponding positive samples, K is the number of negative samples. $\boldsymbol{h}_{\mathrm{CLS}}$ is the embedding of CLS token, which serves as the initial state of statement and contains context-specific causal patterns.

2) *Aspect-level* causal pattern contrastive learning Many previous studies (Cao et al., 2021; Chen et al., 2022) focus on exploring the causality of event pairs. Recently, Liu et al. (2020a) exploits event-agnostic, context-specific patterns which achieve promising performance. Thus, we categorize causal patterns into two aspects: event pair and event-agnostic context. Using the mentioned statement "$\mathrm{The\ traffic\ congestion\ generates\ environmental\ pollution\ and\ economic\ loss}$" as an example, one of its causal patterns is the event pair aspect that "$\mathbf{traffic\ congestion}$" and "$\mathbf{environmental\ pollution}$", another causal pattern is the event-agnostic context aspect "$\mathrm{The\ [EVENT]}$ $\mathbf{generates}$ $\mathrm{[EVENT]}$ ..." Specifically, we split each statement into event pair and event-masked context, and get representations of these two types of aspects after encoding by multilingual pre-trained language model. We further introduce the *aspect-level* causal pattern contrastive learning loss:

$$\mathcal{L}_{\mathrm{AspE}} = -\sum_{j=1}^{n} log \frac{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{AspE}+})}{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{AspE}+}) + \sum_{k=0}^{K-1} s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^k_{\mathrm{AspE}-})}$$

$$\mathcal{L}_{\mathrm{AspC}} = -\sum_{j=1}^{n} log \frac{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{AspC}+})}{s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^j_{\mathrm{AspC}+}) + \sum_{k=0}^{K-1} s(\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}^k_{\mathrm{AspC}-})}$$

We consider contrastive learning loss from both event pair aspect ($\mathcal{L}_{\mathrm{AspE}}$) and event-agnostic context aspect ($\mathcal{L}_{\mathrm{AspC}}$).

2.4. Training

We concatenate event pair node representation $\boldsymbol{v}_{e_{i,j}}$, statement node representation $\boldsymbol{h}_{\mathrm{CLS}}$. For training, we adopt cross entropy as loss function:

$$\boldsymbol{p}_{e_i,e_j} = \mathrm{softmax}(\boldsymbol{W}_p[\boldsymbol{v}_{e_{i,j}} || \boldsymbol{h}_{\mathrm{CLS}}]) \tag{5}$$

$$\mathcal{L}_{\mathrm{C}} = -\sum_{\boldsymbol{e}_i, e_j \in E_s} \boldsymbol{y}_{e_i,e_j} \log(\boldsymbol{p}_{e_i,e_j}) \tag{6}$$

where $\boldsymbol{p}_{e_i,e_j}$ is the predicted probability of causality between events e_i and e_j. E_s is the set of events, $\boldsymbol{y}_{e_i,e_j}$ is a one-hot vector representing the gold label between e_i and e_j. We sum the losses as follows:

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{C}} + \mathcal{L}_{\mathrm{S}} + \mathcal{L}_{\mathrm{AspE}} + \mathcal{L}_{\mathrm{AspC}} \tag{7}$$

In our implementations, we employ the base versions of the language-specific pre-trained language models (PLMs) and the multilingual PLMs. The learning rate is initialized as 1e-3 with a linear decay. We use the AdamW algorithm (Loshchilov and Hutter, 2017) to optimize model parameters. The batch size is set to 1, the number of GATv2 layers is 3. The number of training epochs is 60. Each experiment is conducted on NVIDIA GeForce RTX 3090 GPUs.

3. Experiments

3.1. Dataset

As the cross-lingual DECI is under-explored, the dataset we build upon is currently the only large-scale multi-lingual ECI dataset Lai et al. (2022) that employs a consistent annotation standard. It comprises as many as 3591 documents of five typologically diverse languages, i.e., English, Danish, Spanish, Turkish, and Urdu, the details are shown in Table 1. This dataset is not only larger but also more challenging, as a majority of events are 10 to 50 words away from each other in documents and there are clear divergences between the distance distributions of causal events over languages (we list the majority distance of causal events below).

Language	Document	Relation	Event	Distance
Danish	519	1377	6909	22
English	438	2050	8732	40
Spanish	746	1312	11839	39
Turkish	1357	5337	14179	33
Urdu	531	979	4975	23

Table 1: Statistics for the MECI dataset.

	Model	English			Danish			Spanish			Turkish			Urdu			AVG
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
*	PLM	35.6	44.9	39.7	23.2	23.0	23.1	42.7	44.6	43.6	40.4	56.0	46.9	20.2	33.5	25.2	35.7
*	ERGO	54.7	65.7	55.7	36.4	21.3	26.9	62.3	44.0	51.6	61.3	50.2	55.2	37.4	35.3	36.3	45.1
*	RichGCN	48.1	**69.5**	56.8	27.1	35.0	30.6	59.8	48.2	53.4	54.7	62.0	58.1	31.1	**47.9**	37.7	47.3
	GIMC (base)	**64.9**	57.0	**60.7**	**53.9**	**45.3**	**49.2**	**78.5**	**55.2**	**64.8**	**74.1**	**62.3**	**67.7**	**48.3**	42.7	**45.3**	**57.5**
XLMR	PLM	48.7	59.9	53.7	35.9	36.2	36.0	50.6	49.1	49.9	44.0	59.4	50.5	40.4	43.2	41.8	46.4
XLMR	Know	39.3	42.6	40.9	31.4	11.4	16.7	39.9	28.4	33.2	36.5	46.7	41.0	41.1	22.2	28.9	32.1
XLMR	ERGO	55.0	57.5	56.2	39.3	28.1	32.8	44.5	42.4	43.9	54.7	51.5	53.1	49.6	35.8	41.6	45.5
XLMR	RichGCN	50.6	**68.0**	58.1	31.9	50.0	38.9	50.7	55.0	52.8	50.5	**64.6**	56.7	37.7	**56.0**	45.1	50.3
	GIMC (base)	**61.5**	58.4	**59.9**	**52.1**	**50.5**	**51.3**	**81.1**	**56.4**	**66.5**	**68.7**	58.2	**63.0**	**64.3**	42.6	**51.2**	**58.4**
mBERT	PLM	38.4	46.0	41.9	25.2	26.6	25.9	43.9	41.5	42.7	36.2	48.7	41.6	31.9	34.3	33.0	37.0
mBERT	Know	35.8	56.7	43.9	25.8	36.0	30.1	39.7	38.3	39.0	39.7	46.9	43.0	36.7	35.3	36.0	38.4
mBERT	ERGO	58.2	49.0	53.2	34.4	24.6	28.7	56.3	39.7	46.6	52.7	45.5	48.8	43.4	41.6	42.5	43.9
mBERT	RichGCN	48.4	**67.1**	56.2	29.7	38.0	33.4	51.2	52.0	51.6	50.0	59.9	54.5	40.1	**50.0**	44.5	48.0
	GIMC (base)	**63.4**	54.8	**58.8**	**60.2**	**45.2**	**51.6**	**77.5**	**55.7**	**64.8**	**70.1**	**60.1**	**64.7**	**62.1**	42.4	**50.4**	**58.1**
	GPT-3.5 (zero-shot)	24.6	79.4	37.6	10.0	66.5	17.4	7.3	74.2	13.3	27.0	69.3	38.8	15.7	63.8	25.2	26.4

Table 2: Monolingual performance on MECI dataset. We report the results using language-specific PLMs ("*"), XLMR and mBERT as the backbone respectively. AVG is the average F1 score for five languages.

	Model	English → English			English → Danish			English → Spanish			English → Turkish			English → Urdu			AVG	Δ
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F		
XLMR	PLM	48.7	59.9	53.7	20.1	**59.2**	30.1	16.0	**66.4**	25.8	36.1	**60.5**	45.2	25.7	**62.0**	36.3	38.2	19.4
XLMR	Know	39.3	42.6	40.9	13.3	42.1	20.3	10.4	47.3	17.1	25.8	57.6	35.7	19.3	54.5	28.5	28.5	15.5
XLMR	ERGO	55.0	57.5	56.2	36.4	34.6	35.5	34.0	52.8	41.4	45.3	40.9	43.0	38.2	42.9	40.0	43.2	16.2
XLMR	RichGCN	50.6	**68.0**	58.1	28.5	43.7	34.5	22.7	62.4	33.3	46.4	55.0	50.3	38.6	55.2	45.5	44.3	17.2
	GIMC (Ours)	**64.2**	54.3	**58.8**	**44.5**	44.1	**44.3**	**69.0**	40.4	**51.0**	**56.7**	46.7	**51.2**	**61.5**	38.5	**47.3**	**50.5**	**10.4**
mBERT	PLM	38.4	46.0	41.9	12.4	35.4	18.4	11.4	63.3	19.3	21.5	47.6	29.6	17.0	44.2	24.6	26.7	18.9
mBERT	Know	35.8	56.7	43.9	7.8	**62.0**	13.8	7.2	**69.4**	13.0	20.4	**55.5**	29.9	14.2	**61.5**	23.0	24.7	24.0
mBERT	ERGO	58.2	49.0	53.2	32.0	28.7	30.3	35.4	41.3	38.1	46.3	33.0	38.6	35.3	36.4	35.8	39.2	17.5
mBERT	RichGCN	48.4	**67.1**	56.2	23.7	45.3	31.1	20.6	58.6	30.5	44.5	52.0	48.0	35.0	56.8	43.3	41.8	18.0
	GIMC (Ours)	**66.6**	54.0	**59.6**	**56.3**	41.9	**48.0**	**58.2**	50.2	**53.9**	**61.6**	42.5	**50.3**	**55.7**	43.4	**48.8**	**52.1**	**9.4**
	GPT-3.5 (few-shot)	27.9	83.2	41.8	12.7	75.0	21.8	8.8	84.9	15.9	27.3	80.2	40.8	18.2	75.6	29.4	29.9	14.8

Table 3: Zero-shot cross-lingual document-level ECI performance with English as the source language. We report the average (AVG) and fluctuation (Δ) of F1 score under the different multilingual PLMs.

3.2. Experimental Settings

We utilize the base versions of all PLMs and evaluate our model in the following settings:

Monolingual learning setting The training and test data are in same language. We use language specific PLMs, i.e., BotXO2[2] for Danish, BERT (Devlin et al., 2019) for English, BETO (Canete et al., 2020) for Spanish, BERTurk (Schweter, 2020) for Turkish, and UrduHack[3] for Urdu.

Multilingual learning setting The ECI models are trained in source language and directly tested in target language. We utilize the multilingual PLMs, i.e., mBERT (Devlin et al., 2019) or XLMR (Conneau et al., 2020) as the backbone.

3.3. Baselines

We compare our method with the vanilla PLM, LLM and three strong monolingual baselines (we replace their backbones with multilingual PLM):

(1) **PLM** By leveraging the embeddings $\boldsymbol{h}_i, \boldsymbol{h}_j$ of two events, the overall representation vector is formed $\boldsymbol{h}_{i,j} = [\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{h}_i - \boldsymbol{h}_j, \boldsymbol{h}_i * \boldsymbol{h}_j]$ for ECI.

(2) **Know** Liu et al. (2020a) retrieves external knowledge from knowledge base and event-agnostic, context-specific patterns to enrich the representations of events for causality identification.

(3) **RichGCN** Tran Phu and Nguyen (2021) implement several interaction graphs, by learning a linear combination of the adjacency matrices of these graphs to get a final graph and uses graph convolutional network (GCN) to capture capture relevant context information.

(4) **ERGO** Chen et al. (2022) define a pair of events as a node and build a complete event relational graph to capture the causation transitivity among event pairs via a graph transformer.

(5) **GPT-3.5** We use gpt-3.5-turbo with zero-shot / few-shot learning on this complex multilingual task. The prompt is shown in Figure 5 and language-specific prompts can be found in Appendix B.

[2]https://huggingface.co/Maltehb/danish-bert-botxo

[3]https://github.com/urduhack/urduhack

Model	Danish → English			Danish → Danish			Danish → Spanish			Danish → Turkish			Danish → Urdu			AVG	Δ
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F		
PLM	43.6	25.9	32.5	32.1	25.6	28.5	28.7	35.7	31.8	29.6	22.5	25.6	28.6	20.3	23.7	28.4	3.8
ERGO	67.1	30.0	41.4	34.4	24.6	28.7	57.3	33.1	41.9	52.5	31.0	38.9	39.2	28.5	33.0	36.7	10.1
RichGCN	52.8	40.0	45.3	29.7	29.0	33.4	34.4	**49.3**	40.5	44.4	**42.3**	43.3	32.4	**54.7**	40.6	40.7	9.0
GIMC (Ours)	**63.7**	**42.7**	**51.1**	**61.1**	**40.4**	**48.6**	**65.8**	44.9	**53.4**	**63.3**	37.1	**46.8**	**57.6**	41.6	**48.3**	**49.6**	**2.4**
GPT-3.5 (few-shot)	23.2	80.8	36.1	10.4	79.5	18.4	7.9	85.6	14.4	25.2	79.7	38.3	17.2	82.7	28.5	27.1	12.3
Model	**Spanish → English**			**Spanish → Danish**			**Spanish → Spanish**			**Spanish → Turkish**			**Spanish → Urdu**			**AVG**	**Δ**
PLM	57.1	23.2	33.0	27.8	16.3	20.6	51.3	45.5	48.2	52.9	18.5	27.4	35.0	21.7	26.8	31.2	21.3
ERGO	71.0	21.7	33.3	42.5	20.1	27.2	56.3	39.7	46.6	55.3	21.4	30.8	44.9	24.6	31.7	33.9	15.8
RichGCN	56.4	31.1	40.2	26.9	27.7	27.3	51.2	**52.0**	51.6	52.5	29.9	38.1	42.1	33.7	37.5	38.9	**15.8**
GIMC (Ours)	**75.6**	**36.7**	**49.4**	**65.5**	**35.8**	**46.3**	**81.4**	51.3	**62.9**	**70.0**	**33.4**	**45.2**	**66.1**	**36.6**	**47.1**	**50.2**	15.9
GPT-3.5 (few-shot)	23.7	85.7	37.2	9.9	77.5	17.5	6.9	84.0	12.8	23.9	80.0	36.9	15.5	80.6	26.0	26.1	16.6
Model	**Turkish → English**			**Turkish → Danish**			**Turkish → Spanish**			**Turkish → Turkish**			**Turkish → Urdu**			**AVG**	**Δ**
PLM	34.3	43.3	38.3	24.5	33.5	28.3	23.4	45.4	30.9	39.8	50.9	44.7	27.2	43.5	33.5	35.1	11.9
ERGO	49.6	32.6	39.3	27.6	28.1	27.8	29.1	40.8	33.9	52.7	45.5	48.8	38.0	45.8	41.5	38.2	13.1
RichGCN	44.5	**61.3**	51.6	20.3	**48.4**	28.6	22.6	**61.3**	33.1	50.0	**59.9**	54.5	36.4	**59.2**	45.1	42.5	14.9
GIMC (Ours)	**58.8**	47.7	**52.7**	**64.8**	39.6	**49.2**	**61.8**	55.8	**58.6**	**76.2**	56.2	**64.7**	**62.3**	54.2	**58.0**	**56.6**	**10.1**
GPT-3.5 (few-shot)	17.1	82.1	28.3	5.8	70.2	10.8	4.0	76.4	7.7	17.7	78.4	28.9	32.2	100	48.7	24.9	14.9
Model	**Urdu → English**			**Urdu → Danish**			**Urdu → Spanish**			**Urdu → Turkish**			**Urdu → Urdu**			**AVG**	**Δ**
PLM	38.8	25.5	30.8	18.7	21.1	19.8	30.4	23.7	26.6	29.5	21.4	24.8	45.1	33.6	38.5	28.1	13.0
ERGO	52.7	25.0	33.9	22.2	22.8	22.5	47.5	27.5	34.8	51.3	33.0	40.1	43.4	41.6	42.5	34.7	9.7
RichGCN	**54.5**	34.8	42.5	26.2	**29.3**	27.7	34.7	41.0	37.6	56.4	**32.6**	**41.3**	40.1	**50.0**	44.5	38.7	**7.2**
GIMC (Ours)	52.5	**38.3**	**44.3**	**50.5**	28.7	**36.6**	**51.3**	**46.3**	**48.7**	**61.1**	24.7	35.2	**60.5**	43.8	**50.8**	**43.1**	9.6
GPT-3.5 (few-shot)	20.7	89.0	33.7	8.3	84.6	15.1	5.6	87.7	10.6	21.8	87.1	35.0	8.3	84.6	15.1	21.9	10.7

Table 4: The Performance in low-resource settings. We separately report the cross-lingual performance with low-source language Danish, Spanish, Turkish, and Urdu as source language.

Task Description:
Answer the causal relationship between event mentions from the given passage. The relation has to be of the type: "CauseEffect" means that the former event is the cause of the latter event,"EffectCause" means that the former event is the effect of the latter event,"NoRel" means that the two events have no causal relationship.
Here are an example:
Passage: One of the <T0> surviving <T0> passengers , an eight - year - old in 1978 , was <T1> awarded <T1> US $ 900,000 in <T2> damages <T2> from the airline by a Portland jury in 1984…
Filling the relations:
{"T0-T1": "<One of the above mentioned relations>", "T1-T0": "<One of the above mentioned relations>", "T2-T1": "<One of the above mentioned relations>", … }
Filled relations:
{"T0-T1": "CauseEffect","T1-T0": "EffectCause", "T2-T1": "CauseEffect", … }
Note: <T>event<T> marks the position of event mention in passage, with T as the code name of the event mention. Fills these following relations and returns the filled json file.
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 5: Few-shot prompt for English. We set language specific prompts for different source languages. The passage and relations are truncated.

3.4. Main Results

(1) The Performance in Monolingual Setting The results is illustrated in Table 2, in this monolingual setting, we observe that our model GIMC (base version), which removes contrastive learning module, achieves the best performance across all languages, and outperforms RichGCN across all settings with an average improvement of 9.4% in F1 score. It is worth noting that we test the effectiveness of our method with different PLMs as the backbone. The significant improvement of our method over the vanilla PLM (18.8%) suggests that the success of our model can be attributed to the effectiveness of our method rather than the backbone. These baselines without well-designed cross-lingual module, rely almost on multilingual backbones, perform poorly when transferring to other languages. However our model can cope well with all language settings, even for Danish (an average increase of 16.4% compared with RichGCN).

(2) The Performance in High-Resource Setting As shown in Table 3, the entire GIMC performs surprisingly well on all languages. Comparing with the results of the baseline RichGCN with mBERT, GIMC pushes the average F1 score from 41.8% to 52.1% and effectively reduces performance fluctuations (i.e., Δ) during transfer to other languages. Note that the performance of RichGCN with mBERT drops significantly for three target languages Danish (by 25.1%), Spanish (by 25.7%), and Urdu (by 12.9%), we observed the same trend in our experiments with GPT-3.5, demonstrating that despite extensive training on high-resource language English, LLM fails to achieve plug-and-play functionality when faced with complex cross-lingual tasks, which is consistent with Lai et al. (2023).

(3) The Performance in Low-Resource Setting Since English as high-resource language may benefit more from PLMs than other languages, we further conduct experiments with low-resource lan-

Model	English	Danish	Spanish	Turkish	Urdu	AVG	Δ
GIMC	**59.6**	48.0	**53.9**	**50.3**	**48.8**	**52.1**	**9.4**
GIMC(GCN)	58.6	47.5	50.4	47.1	48.3	50.4	10.2
- P-P	57.3	43.2	49.2	46.6	45.4	48.3	11.2
- S-P	58.3	46.4	53.0	49.2	46.3	50.6	9.6
- P-E	58.8	47.5	52.2	48.3	47.2	50.8	10.0
- S-E	57.9	45.5	51.5	46.1	47.6	49.7	10.2
- St-E	58.9	**48.4**	51.2	47.5	46.1	50.4	10.6
- E-E	57.5	43.3	47.6	45.7	45.5	47.9	12.0

Table 5: The ablation experiments of GIMC's heterogeneous graph interaction network. We directly report the F1 scores (%) in the zero-shot cross-lingual setting using English as source language.

Model	English	Danish	Spanish	Turkish	Urdu	AVG	Δ
GIMC	59.6	**48.0**	**53.9**	50.3	**48.8**	**52.1**	**9.4**
GIMC$_{word}$	**61.2**	47.2	51.6	**50.6**	48.3	51.8	11.8
GIMC(base)	58.8	44.2	49.7	46.8	47.6	49.4	11.7

Table 6: Performance of GIMC on ablation study for cross-lingual contrastive transfer learning module.

guage as source language. We utilize mBERT as our backbone due to its superior performance in the above experiments. As shown in Table 4, GIMC improves average performance and fluctuation in all language settings and achieves double first in monolingual performance (by 64.7%) and cross-lingual transfer performance (by 56.6%) on Turkish. Contrarily, the GPT-3.5 performs unsatisfactorily (with an average performance 24.9% lower than GIMC). Moreover, The drop of RichGCN for Urdu → Turkish is smaller compared to ours. It suggests that the baseline might be sensitive to certain specific languages, whereas our model demonstrates consistently strong performance across all settings.

3.5. Effects of Heterogeneous Graph Interaction Network

To investigate the effectiveness of our graph interaction network, we conduct the ablation studies as

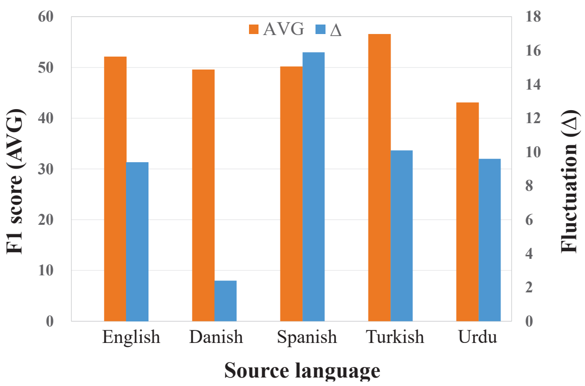


Figure 6: Average F1 score (AVG) and fluctuation (Δ) of GIMC with different source language.

shown in Table 5, the F1 score would decreases 1.3% ~3.8% without the top five types of edges. Besides, removing the E-E edge causes an significant drop by 4.2%, which is consistent with "preserving transitivity of causation" (Paul et al., 2013; Chen et al., 2022), each edge represents a possible causal pattern transitivity between the event pair. Despite a slight decrease in performance when using other information aggregation methods such as GCN, our method is still competitive. Ablation experiments demonstrates that our graph interaction network helps improve the performance.

3.6. Effects of Cross-lingual Contrastive Transfer Learning Module

As depicted in the Table 6, the overall performance of the phrase-level code switching is almost the same as that of the word-level code switching, but its performance in cross-lingual transfer is more stable (Δ decreased by 2.4%). The bottom row of the table shows the performance of the model without contrastive learning module, F1 decreased by 2.7% and delta increased by 2.3% in GIMC(base), suggesting that contrastive learning module of our model does play an important role in the cross-lingual transfer of causal knowledge.

3.7. Cross-lingual Transfer Effect of Different Languages

We plot the performance of zero-shot cross-lingual transfer using different languages as source language in Figure 6. GIMC achieves best cross-lingual performance on Turkish, due to more data and possibly greater linguistic transferability. The large gap between the performance of model on Spanish in monolingual setting and cross-lingual setting leads to its high fluctuation. The drop on Danish much smaller than on other languages, e.g. da-en, which is partly attribute to the effectiveness of our cross-language module, and partly may because of the features of the language such as language families (both Danish and English belong to the Germanic language family). Due to its distinct morphology and syntax compared to the other four languages, Urdu exhibits below-average performance, warranting further attention in future research.

4. Related Work

Event Causality Identification Most event causality identification (ECI) methods focus on sentence-level, and the corresponding datasets are mainly English corpora. Liu et al. (2020a) proposes a mention masking generalization mechanism to learn event-agnostic, context-specific patterns. Zuo et al. (2021) designs a self-supervised framework to learn context-specific causal patterns from causal statements. However, these sentence-level models fail to predict the causality expressed by multiple sentences. So Gao et al. (2019) leverages Integer Linear Programming (ILP) to model document-level structures for ECI. Tran Phu and Nguyen (2021) constructs several document-level interaction graphs and uses GCN to capture relevant context information. Chen et al. (2022) proposes a concise graph network with event pairs as nodes to model the transitivity of causation.

Cross-lingual Transfer Learning Recently, cross-lingual contextualized embeddings have achieved promising results. e.g., mBERT (Devlin et al., 2019), XLMR (Conneau et al., 2020). Further studies also consider aligning representations between source and target languages during fine-tuning. Liu et al. (2020b) proposes task-related parallel word pairs to generate code-switching sentences for learning the interlingual semantics across languages. Qin et al. (2020) leverage a data augmentation framework to generate multi-lingual codeswitching data to fine-tune mBERT. Qin et al. (2022) proposes a global-local contrastive learning framework for cross-lingual spoken language understanding.

5. Conclusion

In this paper, we propose Heterogeneous Graph Interaction Model with Multi-granularity Contrastive Transfer Learning (GIMC) to simultaneously address document-level and zero-shot cross-lingual ECI. We design a multi-granularity contrastive transfer learning module to align causal representations across language. We construct a heterogeneous graph interaction network to model long-distance dependencies between events. Experiments on Multilingual dataset show the effectiveness of our method in both monolingual and multilingual scenarios. Despite extensive training on high-resource language English, GPT-3.5 with few-shot prompts fails to achieve plug-and-play functionality in cross-lingual task, it appears more practical to develop smaller task-specific models for complex multilingual problems.

6. Limitations

This paper are towards the task of zero-shot cross-lingual document-level event causality identification. This work has certain limitations. Firstly, due to lack of dataset, we only conduct experiments on MECI, which is the largest multilingual event causality identification dataset by far, however, the dataset only contain five languages across diverse typologies. It would be beneficial to test the effectiveness of our method on other low-resource language in the future. Secondly, limited by computation resources, our experiments only employ the language-specific PLMs as previous work and the base versions of multilingual PLMs, mBERT and XLMR. Different pre-trained language model as encoder could produce different results. Thirdly, constrained by the high computational cost of API calls, we only conducted analysis on GPT-3.5 with few-shot prompts, thus unable to provide comparisons with other popular LLMs such as GPT-4 (OpenAI, 2023). Furthermore, while we crafted task-specific prompts for GPT-3.5, it could be intriguing to explore additional prompts for testing purposes. We expect future research to encompass a broader range of languages and models.

7. Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2022ZD0160503), the National Natural Science Foundation of China (No. 62176257). This work is also supported by the Youth Innovation Promotion Association CAS.

8. Bibliographical References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

Hongyu Lin Boxi Cao, Xianpei Han, and Le Sun. 2024. The life cycle of knowledge in big language models: A survey. *Machine Intelligence Research*, pages 1–22.

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks? In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021.

Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872, Online. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. CHEER: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Anthony Colas, Ali Sadeghian, Yue Wang, and Daisy Zhe Wang. 2021. Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. *arXiv preprint arXiv:2111.00276*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

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.

Chuang Fan, Daoxing Liu, Libo Qin, Yue Zhang, and Ruifeng Xu. 2022. Towards event-level causal relation identification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1828–1833.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, Toronto, Canada. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and

Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Ludan Ruan Liang Zhang, Anwen Hu, and Qin Jin. 2023. Multimodal pretraining from monolingual to multilingual. *Machine Intelligence Research*, 20:220–232.

Jian Liu, Yubo Chen, and Jun Zhao. 2020a. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Laurie Ann Paul, Ned Hall, and Edward Jonathan Hall. 2013. *Causation: A user's guide*. Oxford University Press.

Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng, and Jian Liao. 2023. Enhancing event causality identification with event causal label and event pair interaction graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10314–10322, Toronto, Canada. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. Gl-clef: A global-local contrastive learning framework for cross-lingual spoken language understanding. *arXiv preprint arXiv:2204.08325*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. 2019. Searching news articles using an event knowledge graph leveraged by wikidata. In *Companion proceedings of the 2019 world wide web conference*, pages 1232–1239.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378.

Stefan Schweter. 2020. Berturk - bert models for turkish.

R Speer, J Chin, and C ConceptNet Havasi. 2016. 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (December 2016)*, pages 4444–4451.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018b. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Changsen Yuan, He-Yan Huang, Yixin Cao, and Yonggang Wen. 2023. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16222–16234.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.

Kun Zhao, Donghong Ji, Fazhi He, Yijiang Liu, and Yafeng Ren. 2021. Document-level event causality identification via graph inference mechanism. *Information Sciences*, 561:115–129.

Bo Zhou, Chenhao Wang, Yubo Chen, Kang Liu, Jun Zhao, Jiexin Xu, Xiaojian Jiang, and Qiuxia Li. 2022. Augmentation, retrieval, generation: Event sequence prediction with a three-stage sequence-to-sequence approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1865–1874.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

9. Language Resource References

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. MECI: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A. Dependency Relations

The table 7 lists the 19 dependency relations we used in this paper. The upper part of the table follows the main organizing principles of the UD taxonomy such that rows correspond to functional categories in relation to the head while columns correspond to structural categories of the dependent. The lower part of the table lists relations that are not dependency relations in the narrow sense.

<table>
<tr><th></th><th>Nominals</th><th>Clauses</th><th>Modifier words</th></tr>
<tr><td>Core arguments</td><td>nsubj
nsubj:pass
obj
iobj</td><td>csubj</td><td></td></tr>
<tr><td>Non-core dependents</td><td>obl
obl:loc
obl:tmod
obl:npmod
dislocated</td><td>advcl</td><td>advmod</td></tr>
<tr><td>Nominal dependents</td><td>appos</td><td>acl
acl:relcl</td><td></td></tr>
<tr><th>Coordination</th><th colspan="2">Loose</th><th>Other</th></tr>
<tr><td>conj</td><td colspan="2">list
parataxis</td><td>root</td></tr>
</table>

Table 7: Dependencies with rich semantics that we reserved in this paper

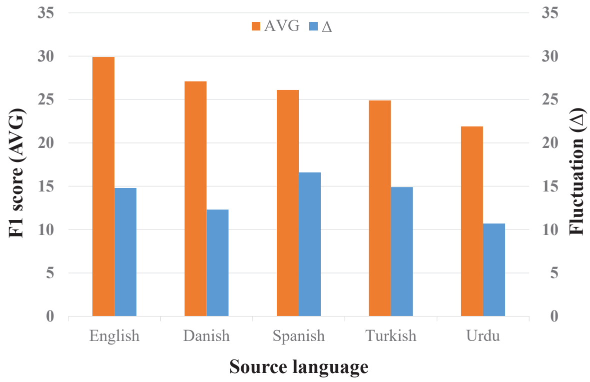


Figure 7: Average F1 score(AVG) and fluctuation(Δ) of GPT-3.5 with different language specific prompt.

B. Experiment Details of GPT-3.5

B.1. Prompts

In the monolingual setting, we employ a zero-shot prompt as illustrated in Figure 8. We provide a task description and specify the required output format. In the multilingual setting, we have designed five different language-specific few-shot prompts. In each prompt, we provide a document and 13 event pairs, where the events mentioned in the document are marked with special symbols (i.e., <T>). Among the 13 event pairs, 8 are causal event pairs, and 5 are unrelated event pairs.

B.2. Analysis

Models were queried through the OpenAI API between December 2022 and May 2023. We illustrate the multilingual performance of GPT-3.5 in Figure 7. It can be observed that GPT-3.5 performs the best in English and the worst in Urdu. This indicates that GPT-3.5 has insufficient understanding of languages other than English, especially low-resource language. Moreover, even in the case of high-resource language English, the overall performance of GPT-3.5 is below 30%. In addition to the possibility of imperfect prompt design, this outcome is highly likely to be attributed to the complexity of the task and the differences between languages.

Task Description:
Answer the causal relationship between event mentions from the given passage. The relation has to be of the type: "CauseEffect" means that the former event is the cause of the latter event,"EffectCause" means that the former event is the effect of the latter event,"NoRel" means that the two events have no causal relationship.

Note: <T>event<T> marks the position of event mention in passage, with T as the code name of the event mention. Your output must only be the relation of the two given event mentions. Fills these following relations and returns the filled json file. Do not add any redundant information except filled json file.

Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 8: Zero-shot prompt for GPT-3.5

Task Description:
Answer the causal relationship between event mentions from the given passage. The relation has to be of the type: "CauseEffect" means that the former event is the cause of the latter event,"EffectCause" means that the former event is the effect of the latter event,"NoRel" means that the two events have no causal relationship.

Here are an example:
Passage: One of the <T0> surviving <T0> passengers , an eight - year - old in 1978 , was <T1> awarded <T1> US $ 900,000 in <T2> damages <T2> from the airline by a Portland jury in 1984 . She was <T3> injured <T3> and both of her parents were <T4> killed <T4> . <T5> Published <T5> in February 2018 , _ Crash Course _ by Julie Whipple <T7> focuses <T7> on the <T8> events <T8> of the night of the <T9> crash <T9> , the <T10> investigation <T10> , and aftermath of the <T11> crash <T11> . - The events of <T13> Flight <T13> 173 were <T14> featured <T14> in " Focused on Failure " , a season - 12 (2012) episode of the Canadian TV series _ Mayday _ (called _ Air Emergency _ in the US and _ Air Crash Investigation _ in the UK and elsewhere around the world) .

Filling the relations:
{"T0-T1": "<One of the above mentioned relations>", "T1-T0": "<One of the above mentioned relations>", "T2-T1": "<One of the above mentioned relations>","T1-T2": "<One of the above mentioned relations>", "T13-T14": "<One of the above mentioned relations>","T14-T13": "<One of the above mentioned relations>","T5-T7": "<One of the above mentioned relations>","T7-T5": "<One of the above mentioned relations>","T0-T10": "<One of the above mentioned relations>","T1-T7": "<One of the above mentioned relations>","T11-T1": "<One of the above mentioned relations>","T11-T9": "<One of the above mentioned relations>","T13-T18": "<One of the above mentioned relations>"}

Filled relations:
{"T0-T1": "CauseEffect","T1-T0": "EffectCause", "T2-T1": "CauseEffect","T1-T2": "EffectCause","T13-T14":"CauseEffect","T14-T13":"EffectCause","T5-T7": "CauseEffect","T7-T5": "EffectCause","T0-T10": "NoRel","T1-T7": "NoRel","T11-T1": "NoRel","T11-T9": "NoRel","T13-T18": "NoRel"}

Note: <T>event<T> marks the position of event mention in passage, with T as the code name of the event mention. Your output must only be the relation of the two given event mentions. Fills these following relations and returns the filled json file. Do not add any redundant information except filled json file.
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 9: English few-shot prompt

Task Description:
Besvar årsagssammenhængen mellem begivenhedsomtaler fra den givne passage. Relationen skal være af typen: "CauseEffect" betyder, at den førstnævnte hændelse er årsagen til den sidstnævnte hændelse,"EffectCause" betyder, at den førstnævnte hændelse er virkningen af den sidstnævnte hændelse,"NoRel" betyder, at de to hændelser har ingen årsagssammenhæng.

Here are an example:
Passage: Det nordamerikanske luftrum holdt <T0> lukket <T0> i flere dage efter <T1> angrebene <T1> , hvilket førte til en <T2> nedskæring <T2> på næsten 20 % i lufttrafikkens kapacitet . Som følge af tvillingetårnenes <T12> kollaps <T12> blev mere end 2.500 forurenende stoffer , herunder kendte kræftfremkaldende stoffer , <T11> spredt <T11> ud over Manhattan . Dette har <T3> ført <T3> til invaliderende <T4> sygdomme <T4> blandt redningsarbejderne . For eksempel <T5> døde <T5> NYPD-officeren Frank Macrì af lungekræft , der havde <T13> bredt sig i <T13> hele hans krop , den 3. september 2007 ; hans familie har <T6> taget <T6> hans <T7> kræftsygdom <T7> som et resultat af den lange arbejdstid , som han har <T8> haft <T8> på stedet . <T9> Effekter <T9> på helbredet <T10> udvidede <T10> sig ligeledes til nogle beboere , studerende og kontorarbejdere i Lower Manhattan samt i nærheden af Chinatown .

Filling the relations:
{"T1-T2": "<En af de ovennævnte relationer>", "T2-T1": "<En af de ovennævnte relationer>", "T1-T0": "<En af de ovennævnte relationer> ","T0-T1": "<En af de ovennævnte relationer>", "T12-T11": "<En af de ovennævnte relationer>","T11-T12": "<En af de ovennævnte relationer >","T11-T3": "<En af de ovennævnte relationer>","T3-T11": "<En af de ovennævnte relationer>","T0-T13": "<En af de ovenfor nævnte relations>","T0-T3": "<En af de ovennævnte relationer>","T0-T9": "<En af de ovennævnte relationer>","T1-T13": "<En af ovenstående nævnte relationer>","T1-T3": "<En af de ovennævnte relationer>"}

Filled relations:
{'T1-T2': 'CauseEffect', 'T2-T1': 'EffectCause', 'T1-T0': 'CauseEffect', 'T0-T1': 'EffectCause', 'T12-T11': 'CauseEffect', 'T11-T12': 'EffectCause', 'T11-T3': 'CauseEffect', 'T3-T11': 'EffectCause','T0-T13': 'NoRel', 'T0-T3': 'NoRel', 'T0-T9': 'NoRel', 'T1-T13': 'NoRel', 'T1-T3': 'NoRel'}

Note: <T>hændelse<T> markerer positionen for hændelsesomtale i passage, med T som kodenavn på hændelsesomtalen. Dit output må kun være relationen mellem de to givne begivenhedsomtaler. Udfylder disse følgende relationer og returnerer den udfyldte json-fil. Tilføj ikke nogen overflødig information undtagen udfyldt json-fil.
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 10: Danish few-shot prompt

Task Description:
Responda la relación causal entre las menciones de eventos del pasaje dado. La relación tiene que ser del tipo: "CauseEffect" significa que el primer evento es la causa del último evento, "EffectCause" significa que el primer evento es el efecto del último evento, "NoRel" significa que los dos eventos tienen sin relación de causalidad.

Here are an example:
Passage: El VUELO 821 DE AEROFLOT , <T0> operado <T0> por la aerolínea Aeroflot-Nord , se <T1> estrelló <T1> el durante su <T2> aproximación <T2> al aeropuerto Bolshoe Savino de la ciudad de Perm , Rusia . Este <T11> desastre <T11> aéreo provocó el <T4> cambio <T4> de marca de Aeroflot-Nord en Nordavia , efectivo el 1 de diciembre de 2009 . Todos los 83 pasajeros y 5 miembros de la tripulación <T5> fallecieron <T5> . La aeronave , un Boeing 737 con <T6> registro <T6> VP-BKO , <T7> despegó <T7> del aeropuerto de Moscú-Sheremetyevo con destino a la ciudad de Perm , al pie de los montes Urales . La causa principal del <T8> accidente <T8> fue que ambos pilotos habían <T9> perdido <T9> la orientación espacial debido a su <T10> inexperiencia <T10> con el tipo de indicador de actitud occidental en el avión .

Filling the relations:
{'T11-T4': '<Una de las relaciones mencionadas anteriormente>', 'T4-T11': '< Una de las relaciones mencionadas anteriormente>', 'T1-T5': '< Una de las relaciones mencionadas anteriormente>', 'T5-T1': '< Una de las relaciones mencionadas anteriormente>', 'T9-T8': '< Una de las relaciones mencionadas anteriormente>', 'T8-T9': '< Una de las relaciones mencionadas anteriormente>', 'T10-T9': '< Una de las relaciones mencionadas anteriormente>', 'T9-T10': '< Una de las relaciones mencionadas anteriormente>', 'T0-T2': '< Una de las relaciones mencionadas anteriormente>', 'T0-T5': '< Una de las relaciones mencionadas anteriormente>', 'T0-T7': '< Una de las relaciones mencionadas anteriormente>', 'T0-T9': '< Una de las relaciones mencionadas anteriormente>', 'T1-T2': '< Una de las relaciones mencionadas anteriormente>'}

Filled relations:
{'T11-T4': 'CauseEffect', 'T4-T11': 'EffectCause', 'T1-T5': 'CauseEffect', 'T5-T1': 'EffectCause', 'T9-T8': 'CauseEffect', 'T8-T9': 'EffectCause', 'T10-T9': 'CauseEffect', 'T9-T10': 'EffectCause', 'T0-T2': 'NoRel', 'T0-T5': 'NoRel', 'T0-T7': 'NoRel', 'T0-T9': 'NoRel', 'T1-T2': 'NoRel'}

Note: <T>event<T> marca la posición de la mención del evento en el pasaje, con T como el nombre en clave de la mención del evento. Su salida solo debe ser la relación de las dos menciones de eventos dadas. Rellena las siguientes relaciones y devuelve el archivo json relleno. No agregue ninguna información redundante, excepto el archivo json lleno.
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 11: Spanish few-shot prompt

Task Description:
Verilen pasajdaki olaylar arasındaki nedensel ilişkiyi yanıtlayın. İlişki şu türden olmalıdır: "CauseEffect", önceki olayın sonraki olayın nedeni olduğu anlamına gelir, "EffectCause", önceki olayın sonraki olayın etkisi olduğu anlamına gelir, "NoRel", iki olayın olduğu anlamına gelir nedensel ilişki yok.

Here are an example:
Passage: Muhalefetteki Ata Meken Sosyalist Partisi milletvekili Ömürbek Tekebayev kargo enkazında Rusça ve Kırgızca etiket basılı ürünler bulunduğunu ve bu ürünlerin açıkça Kırgızistan pazarına yönelik olduğunu <T0> söyledi <T0> . Ayrıca <T1> katıldığı <T1> bir parlamento toplantısında bir Türkiye şirketinden mektup aldığını ; uçakta Bişkek'e gelmeden direkt olarak İstanbul'a gitmesine yetecek yakıtın bulunduğunu ve uçağın , kargoların Bişkek'te <T2> boşaltılması <T2> için <T12> geldiğini <T12> ve hükûmetin doğruları <T13> sakladığını <T3> <T13> söyledi <T3> . Havalimanı gümrüğünün kargodan haberi olmadığı için bazı haberler kargonun kaçak olabileceğini <T4> yazdı <T4> . 27 Ocak'ta nakliye firması Global Link Logistics , Hong Kong'dan gelen bir ACT Airlines uçağının Manas Havalimanı'na <T17> varmasını <T18> <T17> beklediğini <T5> <T18> açıkladı <T5> . Tekebayev bunun ardından havalimanında <T19> kaçakçılık <T19> veya <T20> yolsuzluk <T20> yapılıyor olabileceğini <T7> söyledi <T7> .

Filling the relations:
{"T2-T12": "<Yukarıda belirtilen ilişkilerden biri>","T12-T2": "<Yukarıda belirtilen ilişkilerden biri>","T13-T3": "<Yukarıda belirtilen ilişkilerden biri>","T3-T13": "<Yukarıda belirtilen ilişkilerden biri>","T1-T5": "<Yukarıda belirtilen ilişkilerden biri>","T5-T1": "<Yukarıda belirtilen ilişkilerden biri>","T1-T7": "<Yukarıda belirtilen ilişkilerden biri>","T7-T1": "<Yukarıda belirtilen ilişkilerden biri>","T0-T12": "<Yukarıda belirtilen ilişkilerden biri>","T0-T17": "<Yukarıda belirtilen ilişkilerden biri>","T0-T18": "<Yukarıda belirtilen ilişkilerden biri>","T0-T4": "<Yukarıda belirtilen ilişkilerden biri>","T1-T13": "<Yukarıda belirtilen ilişkilerden biri>"}

Filled relations:
{"T2-T12": "CauseEffect", "T12-T2": "EffectCause", "T13-T3": "CauseEffect","T3-T13": "EffectCause","T1-T5": "CauseEffect","T5-T1": "EffectCause","T1-T7": "CauseEffect","T7-T1": "EffectCause","T0-T12": "NoRel","T0-T17": "NoRel","T0-T18": "NoRel","T0-T4": "NoRel","T1-T13": "NoRel"}

Note: <T>event<T>, geçişte olay bahsinin konumunu, olay bahsinin kod adı olarak T ile işaretler. Çıktınız yalnızca verilen iki olay bahsinin ilişkisi olmalıdır. Aşağıdaki ilişkileri doldurur ve doldurulmuş json dosyasını döndürür. Doldurulmuş json dosyası dışında herhangi bir gereksiz bilgi eklemeyin.
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 12: Turkish few-shot prompt

Task Description:
ےہ بلطم اک "CauseEffect" :ےیہاچ انوہ اک مسق سا قلعت ۔ںید باوج اک قلعت ہجو نایمرد ےک نورکذت ےک ہعقاو ےس ےلاوح ےئگ ےئید
اک "NoRel" ،ےہ رثا اک ہعقاو رکذلا رخؤم ہعقاو ہقباس ہک ےہ ہی بلطم اک "EffectCause" ،ےہ ببس اک ہعقاو رکذلا رخؤم ہعقاو ہقباس ہک
بلطم ےہ ہک ود ںاعقاو ئیوک ہجو رشتہ ںیہن.

Here are an example:
Passage: <T0> ںیم رتلن یداو یک ناتستلب تگلگ ہقالع ےک ناتسکاپ یلامش رٹپاک یلیہ 17-یم لم اک جوف ںیناتسکاپ تقو ےک ند وک ء2015 یئم 8
ہثداح <T0> اک راکش اوہ ۔ جس ںیم ود ٹلئاپ ، اںیشینوڈنا روا اںیشیلم ےک ںوریفس یک نایویب روا نئاپلف روا ںاروس ےک ناتسکاپ ںیم تانیعت ریفس
<T5> یمخز <T5> رگید روا <T4> کالہ <T4> 8 ےس ںیم نج ، ےھت <T8> راوس <T8> ںیناتسکاپ 6 روا یکلم ریغ 11 ںیم زاورپ ، ےئوہ کالہ
تاقیقحت <T6> یک سا ، اھت ایک ببس اک <T7> ہثداح <T7> ۔ ںیہ <T3> لماش <T3> یھب ریفس ےک ڈنیلوپ روا ڈنیلاہ ںیم نویمخز ۔ ںیہ
<T6> جار ںیہ ، ابتدائی طور رپ ضعب ذرائع غالبا (ہجیس ےب یب سی اردو) ےک قباطم یلیہ رٹپاک آرمی پبلک اسکول ںیم اترتے ےئوہ عمارت ےس
<T10> ٹکرایا <T10> ۔ جس ےس یہ <T1> ہحناس <T1> رونما اوہ ۔

Filling the relations:
{"T0-T4": “<ایک سے میں تعلقات بالا مذکورہ>", "T4-T0": “<ایک سے میں تعلقات بالا مذکورہ>",
"T0-T5": "<ایک سے میں تعلقات بالا مذکورہ>", "T5-T0": “<ایک سے میں تعلقات بالا مذکورہ>",
"T7-T6": “<ایک سے میں تعلقات بالا مذکورہ>", "T6-T7": “<ایک سے میں تعلقات بالا مذکورہ>",
"T10-T7": “<ایک سے میں تعلقات بالا مذکورہ>", "T7-T10": “<ایک سے میں تعلقات بالا مذکورہ>",
"T0-T1":“<ایک سے میں تعلقات بالا مذکورہ>", "T0-T3": “<ایک سے میں تعلقات بالا مذکورہ>" ,
"T1-T3": “<ایک سے میں تعلقات بالا مذکورہ>", "T1-T5": “<ایک سے میں تعلقات بالا مذکورہ>" ,
"T1-T6": “<ایک سے میں تعلقات بالا مذکورہ>" }

Filled relations:
{"T0-T4": "CauseEffect","T4-T0": "EffectCause", "T0-T5": "CauseEffect", "T5-T0": "EffectCause", "T7-T6": "CauseEffect","T6-T7": "EffectCause","T10-T7": "CauseEffect","T7-T10": "EffectCause","T0-T1": "NoRel","T0-T3": "NoRel","T1-T3": "NoRel","T1-T5": "NoRel","T1-T6": "NoRel"}

Note: <T>ٹنویا<T> ہلاوح ںیم ہعقاو ےک ہرکذت یک نشیزوپ وک نشان دز اترک ےہ، T ےک ہتاس ہعقاو ےک ہرکذت ےک ڈوک مان ےک
طور رپ ۔آپ اک آؤٹ پٹ فرص ود ےئید ےئگ ٹنویا ےک ہرکذت اک رشتہ انوہ ےیہاچ ۔ ان درج ذیل تاقلعت وک رھبا ےہ روا رھبی ہوئی json لئاف
وک واپس اترک ےہ۔ رھبی ہوئی json لئاف ےک علاوہ وک ئیوک فالتو معلومات شامل ہن کریں۔
Passage: {passage}
Filling the relation: {relation_to_fill}
Filled relations:

Figure 13: Urdu few-shot prompt