Event Temporal Relation Extraction with Bayesian Translational Model

Xingwei Tan[1], **Gabriele Pergola**[1], **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
{Xingwei.Tan, Gabriele.Pergola.1}@warwick.ac.uk
yulan.he@kcl.ac.uk

Abstract

Existing models to extract temporal relations between events lack a principled method to incorporate external knowledge. In this study, we introduce *Bayesian-Trans*, a Bayesian learning-based method that models the temporal relation representations as latent variables and infers their values via Bayesian inference and *translational functions*. Compared to conventional neural approaches, instead of performing point estimation to find the best set parameters, the proposed model infers the parameters' posterior distribution directly, enhancing the model's capability to encode and express uncertainty about the predictions. Experimental results on the three widely used datasets show that Bayesian-Trans outperforms existing approaches for event temporal relation extraction. We additionally present detailed analyses on uncertainty quantification, comparison of priors, and ablation studies, illustrating the benefits of the proposed approach.[1]

1 Introduction

Understanding events and how they evolve in time has been shown beneficial for natural language understanding (NLU) and for a growing number of related tasks (Cheng et al., 2013; Wang et al., 2018; Ning et al., 2020; Geva et al., 2021; Sun et al., 2022). Howeover, events often form complex structures with each other through various temporal relations, which is challenging to track even for humans (Wang et al., 2020a).

One of the main difficulties is the wide variety of linguistic expressions of temporal relations across different contexts. Although many of them share some linguistic similarities, most of the topics in which they occur are characterized by some shared but unspoken knowledge that determines how temporal information is expressed. For example, when it comes to health, prevention is widely

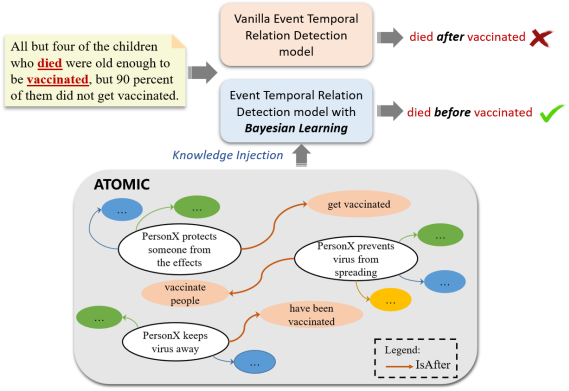

Figure 1: Comparison between with or without external knowledge incorporation on event relation extraction.

practised, with many treatments (e.g., vaccinations) being effective only if administered *before* the onset of a disorder. On the contrary, in the automotive industry, it is common that most people repair their car *after* a problem occurs. However, despite its simplicity, such commonsense knowledge is rarely stated explicitly in text and varies greatly across different domains. For example, in Figure 1, a detection model lacking the commonsense knowledge that *vaccination* can protect people from infection, tends to get confused by the complex linguistic structures in the excerpt and returns the wrong prediction entailing that '*died*' happens `after` '*vaccinated*'. Instead, with the consideration of prior temporal knowledge involving the *vaccination* event from an external knowledge source ATOMIC (Hwang et al., 2021), a model gives the correct prediction that '*died*' occurs `before` '*vaccinated*'.

Methods proposed in recent studies for event relation extraction are mostly end-to-end neural architectures making rather limited use of such commonsense knowledge (Han et al., 2019a,b). Only a few works have explored the incorporation of external knowledge to mitigate the scarcity of event annotations (Ning et al., 2019; Wang et al., 2020b).

[1]Experimental source code is available at https://github.com/Xingwei-Warwick/Bayesian-Trans

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1125–1138
May 2-6, 2023

Nevertheless, these approaches typically update the event representations with knowledge features derived from external sources, lacking a principled way of updating models' beliefs in seeing more data in the domains of interests.

In this work, we posit that the Bayesian learning framework combined with translational models can provide a principled methodology to incorporate knowledge and mitigate the lack of annotated data for event temporal relations. Translational models, such as TransE (Bordes et al., 2013), are energy-based models based on the intuition that the relations between entities can be naturally represented by geometric translations in the embedding space. More concretely, a relation between a *head entity* and a *tail entity* holds if there exists a *translational* operation bringing the *head* close to the *tail* vector.

Specifically, we introduce a novel Bayesian Translational model (Bayesian-Trans) for event temporal relation extraction. Compared to conventional neural translational models, which only yield a point estimation of the network parameters, the Bayesian architecture can be seen as an ensemble of an infinite number of neural predictors, drawing samples from the posterior distribution of the translational parameters, refining its belief over the initial prior. As a result, event temporal relations are determined by the stochastic translational parameters drawn from posterior distributions. Additionally, such posteriors are conditioned upon the prior learned on external knowledge graphs, providing the commonsense knowledge required to interpret more accurately the temporal information across different contexts. As shown in the results obtained from the experimental evaluation on three commonly used datasets for event temporal relation extraction, the combination of translational models and Bayesian learning is particularly beneficial when tailored to the detection of event relations. Moreover, a favorable by-product of our Bayesian-Trans model is the inherent capability to express degrees of uncertainty, avoiding the overconfident predictions on out-of-distribution context. Our contributions are summarized in the following:

- We formulate a novel Bayesian translational model for the extraction of event temporal relations, in which event temporal relations are modeled through the stochastic translational parameters, considered as latent variables in Bayesian inference.
- We devise and explore 3 different priors under Bayesian framework to study how to effectively incorporate knowledge about events.
- We conduct thorough experimental evaluations on three benchmarking event temporal datasets and show that Bayesian-Trans achieves state-of-the-art performance on all of them. We also provide comprehensive analyses of multiple aspects of the proposed model.

2 Related Work

This work is related to at least three lines of research: event temporal relation detection, prior knowledge incorporation, and graph embedding.

2.1 Event Temporal Relation

Similar to entity-level relation extraction (Zeng et al., 2014; Peng et al., 2017), the latest event temporal relation extraction models are based on neural networks, but in order to learn from limited labeled data and capture complex event hierarchies, a wide range of optimization or regularization approaches have been explored. Ning et al. (2019) proposed an LSTM-based network and ensured global consistency of all the event relations in the documents by integer linear programming. Wang et al. (2020b) employed RoBERTa (Liu et al., 2019) and converted a set of predefined logic rules into differentiable objective functions to regularize the consistency of the relations inferred and explore multi-task joint training. Tan et al. (2021) proposed using hyperbolic-based methods to encode temporal information in a hyperbolic space, which has been shown to capture and model asymmetric temporal relations better than their Euclidean counterparts. Hwang et al. (2022) adopted instead a probabilistic box embeddings to extract asymmetric relations. Wen and Ji (2021) proposed to add an auxiliary task for relative time prediction of events described over an event timeline. Cao et al. (2021) developed a semi-supervised approach via an uncertainty-aware self-training framework, composing a training set of samples with actual and pseudo labels depending on the estimated uncertainty scores. None of the aforementioned approaches explored Bayesian learning for incorporating prior event temporal knowledge.

2.2 Incorporation of Prior Knowledge

Knowledge plays a key role in understanding event relations because people often skip inessential details and express event relations implicitly which

is difficult to understand without relevant knowledge. For example, TEMPROB (Ning et al., 2018b) contains temporal relation probabilistic knowledge which is encoded by Siamese network and incorporated into neural models as additional features (Ning et al., 2019; Wang et al., 2020b; Tan et al., 2021). Unlike previous works, we combine the Bayesian Neural Network with distance-based models, treating the translational parameters as latent variables to be inferred. To this end, we adopt the variational inference (Kingma and Welling, 2014a; Blei et al., 2016; Gui et al., 2019; Pergola et al., 2021a; Zhu et al., 2022), and derive the prior distribution of the temporal relation information from commonsense knowledge bases (Pergola et al., 2021b; Lu et al., 2022). Christopoulou et al. (2021) explored a similar intuition of using knowledge base priors as distant supervision signals, but the approach and the task are different.

2.3 Graph Embedding Learning

Multi-relational data are commonly interpreted in terms of directed graphs with nodes and edges representing entities and their relations, respectively. Several works have recently focused on modelling these multi-relational data with relational embeddings by detecting and encoding local and global connectivity patterns between entities.

TransE (Bordes et al., 2013) has been a seminal work adopting geometric translations of entities to represent relations in the embedding space. If a relation between a head and a tail entity holds, it is encoded via the translational parameters learned at training time. However, TransE cannot model symmetry relation well by simple addition which led to several subsequent studies exploring diverse types of transformation resulting in a family of *translational models* (Wang et al., 2014; Ji et al., 2015; Lin et al., 2015). Among them, Balazevic et al. (2019) proposed to utilize the Poincaré model, mapping the entity embeddings onto a Poincaré ball, and using the Poincaré metric to compute the score function and predict their relations. Chami et al. (2020b) further expanded the idea of embedding learning over manifolds by additionally considering reflections and rotations and redefining the translation over a learned manifold.

Although translational models are shown efficient in modeling graph relation, they provide relatively limited interaction between nodes than neural network-based methods, such as Graph Neural Networks (Estrach et al., 2014; Chami et al., 2020a). Under this framework, nodes in a graph are neural units, which can iteratively propagate information through edges, and whose representations are learnt during the training process. In particular, Relational Graph Convolutional Networks (RGCN) (Schlichtkrull et al., 2018) encode relational data through link prediction and entity classification tasks, while enforcing sparsity via a parameter-sharing technique. Although modeling knowledge graphs has been one of the main focuses of the above-mentioned graph learning approaches, they lack any systematic mechanism to inject prior knowledge and update it during training.

3 Bayesian-Trans Model

In identifying temporal relations between events, we aim at predicting the relation type of two events given in text, commonly denoted as *head* event x_h and *tail* event x_t:

$$\hat{y} = \arg\max_{y \in \mathcal{R}} p(y|x_h, x_t) \quad (1)$$

where $\mathcal{R}$ denotes a set of possible relation types, while x_h and x_t the head and tail event triggers, respectively. Assuming that a set of latent variables $\mathbf{\Lambda}$ denotes the collection of all relation-specific transformation parameters $\mathbf{\Lambda}_r$. For example, in the knowledge embedding learning model such as MuRE (Balazevic et al., 2019), the head entity is first transformed through a relation-specific matrix $\mathbf{W}_r$, followed by a relation-specific translation vector t_r, then $\mathbf{\Lambda}_r = \{\mathbf{W}_r, t_r\}$. By Bayesian learning, the probability of inferring a relation type r can be written as:

$$p(y = r|x_h, x_t) = \int_{\mathbf{\Lambda}} p(y_r|x_h, x_t, \mathbf{\Lambda}) p(\mathbf{\Lambda}|\mathcal{G}) d_{\mathbf{\Lambda}} \quad (2)$$

Here, $p(\mathbf{\Lambda}|\mathcal{G})$ denotes the prior distribution of $\mathbf{\Lambda}$ derived from an existing knowledge graph encoded as $\mathcal{G}$. Directly inferring Eq. (2) is intractable. But we can resort to amortised variational inference to learn model parameters. In what follows, we present our proposed Bayesian learning framework built on translational models for event temporal relation extraction, called **Bayesian-Trans**, with its architecture shown in Figure 2.

In particular, the context S in which the two events occur is the input to our Bayesian-Trans. First, we encode S via a pre-trained language model generating the contextual embeddings e_h and e_t for the triggers of the head and tail events,

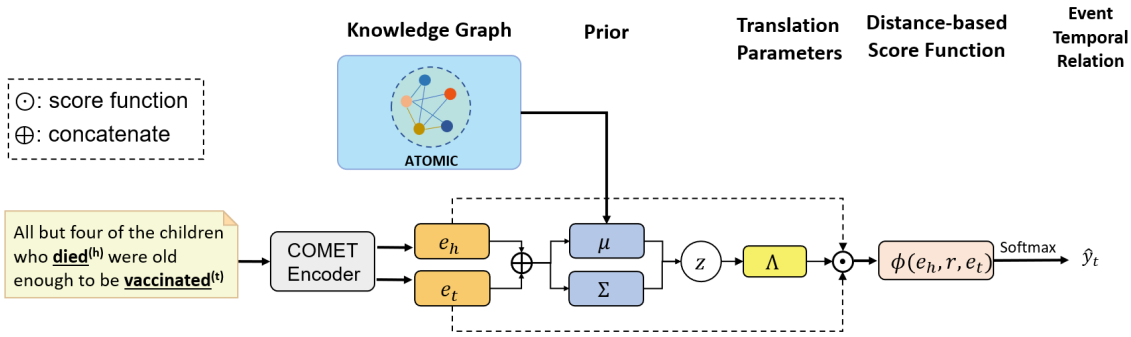

Figure 2: The network structure of Bayesian-Trans. Context sentences are first fed into a COMET encoder to generate event representations. With MLP layers, the event representations are mapped to generate a variational distribution of relation representations which is guided by KG priors. The relation representations are then used in the translational model to generate prediction scores.

respectively. The contextualised event trigger representations, e_h and e_t, are fed as input into a Bayesian translational module. This module, by means of variational inference, determines the parameters of the translational model, encoding the posterior distribution of the temporal relations conditioned upon the input events. Finally, we use a score function on the translated head and tail triggers to predict their temporal relation. We provide a more detailed description in the following.

3.1 Contextual Encoder

The proposed model uses COMET-BART (Hwang et al., 2021) as the context encoder. COMET-BART is a BART pre-trained language model (Lewis et al., 2020) fine-tunned on ATOMIC (Bosselut et al., 2019; Hwang et al., 2021), which is an event-centric knowledge graph encoding inferential knowledge about entities and events, including event temporal relations. The COMET-BART is able to generate consequence events given the antecedent event and a relation with good accuracy thus is regarded encodes knowledge well. Following the approach adopted in previous works (Ning et al., 2019; Wang et al., 2020b; Tan et al., 2021), we use the representation of the first token of an event trigger as the contextual embedding of that event[2], $\mathrm{e}_h, \mathrm{e}_t = \text{COMET-BART}(x_h, x_t)$, where $\mathrm{e}_h, \mathrm{e}_t \in \mathbb{R}^d$. The event representations are then concatenated together and fed through MLPs to generate the parameters of the variational distribution, from which the latent event-pair representation z is sampled. z is then mapped to the parameter space of the translational model as $\mathbf{\Lambda}$.

3.2 Incorporating Knowledge via Bayesian Learning

The proposed model utilizes relation embeddings for classifying event relation in a similar manner as the translational models in knowledge graph embedding, such as TransE (Bordes et al., 2013). If the embedding of the tail event is close enough to the embedding of head event after applying a series of relation-specific transformation, the relation stands, and vice versa. A wide range of translational models typically proposed for learning knowledge graph embeddings can be adopted in the proposed Bayesian-Trans. Additionally, to incorporate prior knowledge, we extend translational models to operate within the Bayesian inference framework. We proceed with introducing a standard translational model in the context of temporal relations, and describe how we extend it to work in the Bayesian framework.

Translational Model Generally speaking, a translational model uses *relation representations* $\mathbf{\Lambda}_r$ to perform "translation" for relation r on the head and tail events. Then, the transformed head and tail event embeddings are compared using a *distance-based score function*, whose score is indicative of the temporal relation between the events. The score function $\phi(\cdot)$ takes the general form:

$$\phi(\mathrm{e}_h, r, \mathrm{e}_t) = -d(\mathcal{T}^h_{\mathbf{\Lambda}_r}(\mathrm{e}_h), \mathcal{T}^t_{\mathbf{\Lambda}_r}(\mathrm{e}_t)) \quad (3)$$

where r is a relation type, $\mathcal{T}_{\mathbf{\Lambda}_r}(\cdot)$ is a function depending on the parameters $\mathbf{\Lambda}_r$ of relation r to transform the event embeddings e_h and e_t, and $d(\cdot)$ is any distance metrics (e.g., Euclidean distance).

[2]We conducted some exploratory experiments adopting the last token or the average representation, but results showed that the first token was still the best option in this context.

We explored several models with different translation functions and distance metrics in the context of temporal relations, including TransE (Bordes et al., 2013), AttH (Chami et al., 2020b), MuRE (Balazevic et al., 2019) and MuRP (Balazevic et al., 2019), and based on our preliminary results[3], we eventually adopted MuRE as it strikes a good balance of training efficiency and accuracy of temporal relation classification. We define the scoring function in the proposed model as follows:

$$\phi(\mathrm{e}_h, r, \mathrm{e}_t) = -\|\mathbf{W}_r \mathrm{e}_h + \mathrm{t}_r - \mathrm{e}_t\|_2^2 \quad (4)$$

where $\mathbf{W}_r \in \mathbb{R}^{d\times d}$ is a diagonal relation matrix and $t_r \in \mathbb{R}^d$ a translation vector of relation r, $\mathbf{\Lambda}_r = \{\mathbf{W}_r, \mathrm{t}_r\}, r \in \mathcal{R}$.

Although the number of parameters to train is rather low, the number of annotated samples is usually small compared to the wide range of linguistic expressions capturing temporal relations. We thus extend the MuRE model into a Bayesian framework to enhance its scalability by treating the translational parameters $\mathbf{\Lambda}$ as latent variables. The proposed framework enhances generalization by defining a variational inference process that optimizes the regularization and leverages the additional information injected via the prior distributions.

Bayesian Inference As shown in the inference equation 2, the prior is derived from an external knowledge graph, such as ATOMIC, as a means to inject prior information about events and temporal relations. In particular, $\mathbf{\Lambda}$ is assumed to follow a Gaussian distribution with unit variance and with mean determined by the relation representations trained on the knowledge graph. The probability function is formulated as a softmax function over a pre-defined scoring function:

$$p(y_r|\mathrm{e}_h, \mathrm{e}_t, \mathbf{\Lambda}) = \frac{\exp\Big(\phi(\mathrm{e}_h, r, \mathrm{e}_t)\Big)}{\sum_{r'\in\mathcal{R}} \exp\Big(\phi(\mathrm{e}_h, r', \mathrm{e}_t)\Big)} \quad (5)$$

with e_h and e_t denoting the embedding for the head and the tail events, respectively.

Yet, Eq. (2) is intractable and cannot be inferred directly. Thus, we resort to amortized variational inference by introducing a variational posterior $q_\theta(\mathbf{\Lambda}|x_h, x_t)$, which follows the isotropic Gaussian distribution and can be modeled as:

$$\begin{aligned} \mu = f_\mu(\mathrm{e}_h;\mathrm{e}_t) \quad \Sigma = \mathrm{diag}\Big(f_\Sigma(\mathrm{e}_h;\mathrm{e}_t)\Big) \\ q_\theta(\mathbf{\Lambda}|\mathrm{e}_h,\mathrm{e}_t) = \mathcal{N}(\mathbf{\Lambda}|\mu,\Sigma), \end{aligned} \quad (6)$$

where f_μ and f_Σ are both fully connected layers that map the event pair representation into the parameters of the variational distribution.

Following the amortized variational inference, we maximize the evidence lower bound (ELBO) $\mathcal{L}_e$, defined in Eq. (7), and approximated by a Monte Carlo estimation with sample size N, as described in Eq. (8):

$$\begin{aligned} \mathcal{L}_e = \mathbb{E}_{q_\theta(\mathbf{\Lambda}|x_h,x_t),\{x_h,x_t\}\in\mathcal{D}}\Big[\log p_\theta(y|x_h,x_t,\mathbf{\Lambda})\Big] - \\ \mathrm{Reg}\Big(q_\theta(\mathbf{\Lambda}|x_h,x_t,\mathcal{G})||p(\mathbf{\Lambda}|\mathcal{G})\Big) \quad (7) \\ \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{\{x_h,x_t\}\in\mathcal{D}}\Big[\log p_\theta(y|x_h,x_t,\mathbf{\Lambda}^{(n)}) - \\ \mathrm{Reg}\Big(q_\theta(\mathbf{\Lambda}^{(n)}|x_h,x_t,\mathcal{G})||p(\mathbf{\Lambda}^{(n)}|\mathcal{G})\Big)\Big] \quad (8) \end{aligned}$$

where $\mathrm{Reg}(\cdot)$ is a regularization term which will be discussed in 3.3. To train end-to-end a fully differentiable model, we adopt the reparameterization trick (Kingma and Welling, 2014b).

3.3 Prior Distribution and Regularization

We proceed to discuss how the Bayesian framework enabled the incorporation of prior acquired from an external knowledge source. Then, we provide the details of how we compute the regularization term to induce a more stable training.

Prior Distribution One of the main advantages of the Bayesian inference framework is the possibility to inject commonsense knowledge into the model through the prior distribution of the latent variables, i.e., $p(\mathbf{\Lambda}|\mathcal{G})$ in Eq. (2), where $\mathbf{\Lambda}$ are the translational parameters and $\mathcal{G}$ denotes an external knowledge graph, in our case, the ATOMIC knowledge graph (Hwang et al., 2021). ATOMIC is a commonsense knowledge graph containing inferential knowledge tuples about entities and events encoding social and physical aspects of human everyday experiences. For our task of event temporal relation extraction, we are only interested in the events linked via temporal relations, such as 'IsBefore' (23,208 triples) or 'IsAfter' (22,453 triples). By conducting link prediction on these links, we use relation embeddings learnt using an RGCN (Schlichtkrull et al., 2018) as the mean of the prior distribution for the translational latent variables. For the relations in the experiment

[3]Experimental results using different translational models are shown in Table A1.

dataset that do not have applicable counterparts in ATOMIC (e.g., VAGUE), we set their priors to standard Gaussian. The variance of the priors is defined as the identity matrix.

Specifically, we use COMET-BART to encode the event nodes from ATOMIC, then use their context embeddings as the node features in the RGCN. In our preliminary experiment, we also found that RGCN cannot train well on the commonsense graph with only the event-event relation links. The graph is too sparse which makes the information difficult to propagate through the nodes. Thus, we added semantic similarity links based on the cosine similarity of the event context embeddings. During the training of the RGCN, the node embeddings are kept frozen. After the training of the link prediction task, we extract the relation embeddings of the RGCN.

Regularization Term To mitigate the posterior collapse problem (Lucas et al., 2019) and have a stable inference process, we adopt the Maximum Mean Discrepancy (MMD)[4] which is an estimation of Wasserstein distance (Tolstikhin et al., 2018) as the regularization term (Eq. 8).

4 Experimental Setup

Datasets We evaluated the proposed Bayesian-Trans model on three event temporal relation datasets: MATRES (Ning et al., 2018c), Temporal and Causal Reasoning (TCR) (Ning et al., 2018a), and TimeBank-Dense (TBD) (Cassidy et al., 2014). TimeBank-Dense is a densely annotated dataset focusing on the most salient events and providing 6 event temporal relations. MATRES follows a new annotation scheme which focuses on main time axes, with the temporal relations between events determined by their endpoints, resulting in a consistent inter-annotator agreement (IAA) on the event annotations (Ning et al., 2018c). TCR follows the same annotation scheme, yet with a much smaller number of event relation pairs than in MATRES. Table 1 shows the statistics of the datasets.

Baselines We compare the proposed Bayesian-Trans[5] with the following baselines:

CogCompTime (Ning et al., 2018d) is a multi-step system which detect temporal relation using semantic features and structured inference.

[4]MMD calculation can be found in Appendix A.

[5]Hyperparameter setting can be found in Appendix B.

Class	MATRES	TCR	TBD
BEFORE	6,852	1,780	2,590
AFTER	4,752	862	2,104
EQUAL/SIMULTANEOUS	448	4	215
VAGUE/NONE	1,425	N/A	5,910
INCLUDE	N/A	N/A	836
ISINCLUDED	N/A	N/A	1,060
Total	12,740	2,646	12,715

Table 1: The statistics of MATRES, TCR, and TBD.

BiLSTM is a basic relation prediction model built by Han et al. (2019b).
LSTM + knowledge (Ning et al., 2019) incorporates knowledge features learnt from an external source and optimize global consistency by ILP.
Deep Structured (Han et al., 2019a) adds a structured support vector machine on top of a BiLSTM.
Joint Constrained Learning (Wang et al., 2020b) constrains the training of a RoBERTa-based event pair classifier using predefined logic rules, while knowledge incorporation and global optimization are also included.
Poincaré Event Embedding (Tan et al., 2021) learns event embeddings based on a Poincaré ball and determines the temporal relation base on the relative position of events.
HGRU + knowledge (Tan et al., 2021) is a neural architecture processing temporal relations via hyperbolic recurrent units which also incorporates knowledge features like LSTM + knowledge.
Relative Event Time (Wen and Ji, 2021) is a neural network classifier combining an auxiliary task for relative time extraction over an event timeline.
UAST (Cao et al., 2021) is an uncertainty-aware self-training model. We show the result of the model which is trained on all the labeled data.

5 Experimental Results

Temporal Relation Classification We first compare Bayesian-Trans with the most recent approaches for temporal event classification in Table 2, including methods with or without commonsense knowledge injection. The results are obtained by training models on the MATRES training set and evaluated on both the MATRES test set and TCR. Table 3 shows results from the TBD dataset which are generated using the provided train, development, and test sets. We report F_1 score on MATRES and TCR following the definition in (Ning et al., 2019), and micro-F_1 on TimeBank-Dense. Compared with existing methods, the proposed

Model	MATRES			TCR		
	P	R	F_1	P	R	F_1
CogCompTime (Ning et al., 2018d)	61.6	72.5	66.6	-	-	70.7
Poincaré Event Embeddings (Tan et al., 2021)	74.1	84.3	78.9	85.0	86.0	85.5
Relative Event Time (Wen and Ji, 2021)	78.4	85.2	81.7	84.3	**86.8**	85.5
LSTM + knowledge (Ning et al., 2019)	71.3	82.1	76.3	-	-	78.6
Joint Constrainted Learning (Wang et al., 2020b)	73.4	85.0	78.8	83.9	83.4	83.7
HGRU + knowledge (Tan et al., 2021)	79.2	81.7	80.5	88.3	79.0	83.5
Bayesian-Trans	**79.6**	**86.0**	**82.7**	**89.8**	82.6	**86.1**

Table 2: Experimental results on MATRES and TCR. The first three lines contain methods without commonsense knowledge incorporation. The rest are methods which inject commonsense knowledge. The results of Wang et al. (2020b) and (Wen and Ji, 2021) on TCR are generated from our run of the source code provided by the authors since they are not available in their original papers. The others are taken from the cited papers.

Model	Micro-F_1
BiLSTM (Han et al., 2019b)	61.9
Deep Structured (Han et al., 2019a)	63.2
Relative Event Time (Wen and Ji, 2021)	63.2
UAST (Cao et al., 2021)	64.3
Bayesian-Trans	**65.0**

Table 3: Experimental results on TBD. All compared methods do not incorporate commonsense knowledge explicitly. The result of Wen and Ji (2021) is generated from our run of the source code provided by the authors since they are not available in their original paper. The others are taken from the cited papers.

Model	MATRES	TBD
(1) RoBERTa + MLP	81.5	62.8
(2) RoBERTa + Vanilla MuRE	80.4	60.5
(3) RoBERTa + Bayesian-Trans	82.2	63.0
(4) COMET-BART + MLP	81.8	63.2
(5) COMET-BART + Vanilla MuRE	81.8	62.6
(6) COMET-BART + Bayesian-Trans	**82.7**	**65.0**

Table 4: Ablation test results on MATRES and TBD.

Bayesian-Trans has generally better performance on all three datasets, with more noticeably improvements on MATRES. Bayesian-Trans has significant performance gains over previous methods with knowledge incorporation, which shows that it can utilize knowledge more extensively. Details of the per-class performance can be found in Table A2 and A3.

Ablation Study We conducted an ablation study to highlight the impact of the different modules composing Bayesian-Trans. The results are shown in Table 4. In particular, we have the following variants: (1) RoBERTa+MLP, using RoBERTa to encode the context and then feeding representations of head and tail events to a multi-layer perceptron (MLP) for temporal relation classification; (2) RoBERTa+ Vanilla MuRE, using MuRE to extract temporal relations without modeling its parameters as latent variables; (3) RoBERTa+Bayesian-Trans, our proposed model by replacing COMERT-BART with RoBERTa as the text encoder; (4) COMET-BART+MLP, using COMET-BART as context encoder and an MLP for temporal relation classification; and (5) COMET-BART+ Vanilla MuRE, the proposed model without Bayesian learning or knowledge incorporation. The results demonstrate that COMET-BART is a better choice as the context encoder. Using MuRE for event temporal knowledge embedding learning does not bring any improvement compared to using a simple MLP layer for event temporal relation prediction (see (1) cf. (2), and (4) cf. (5)). Regardless of the contextual encoder used, the results of (3) and (6) show the benefit of employing Bayesian learning which naturally incorporates prior knowledge of event temporal relations learned from an external knowledge source for event temporal relation detection. With our proposed Bayesian translational model, we observe an improvement of $0.9-1.8\%$ in micro-F_1 on MATRES and $0.2-2.5\%$ in micro-F_1 on TimeBank-Dense compare to their non-Bayesian counterparts.

Effects of the Priors We further investigate the impact of different priors on the model performance. Inspired by the work on VAEs by Burda et al. (2016) and Truong et al. (2021), we employed

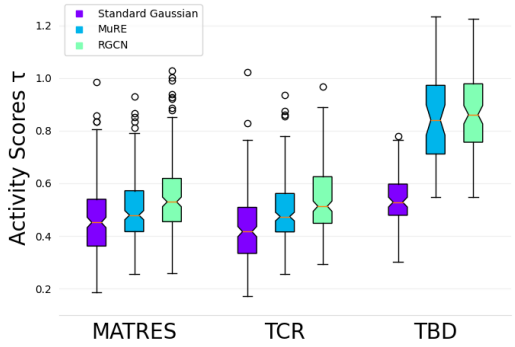

Figure 3: The box chart of the activity scores across all the dimensions of the latent encoding $\mathbf{\Lambda}$ with respect to the priors used in the model.

Dataset	Standard Gaussian	MuRE	RGCN
MATRES	81.2	81.8	82.7
TCR	84.3	85.4	86.1
TBD	63.6	64.6	65.0

Table 5: F_1 values based on different priors used in the proposed model.

an *'activity' score*, $\tau = Cov_{e_h,e_t}(\mathbb{E}_{q(\mathbf{\Lambda}|e_h,e_t)}[\mathbf{\Lambda}])$ to evaluate the quality and diversity of the latent encodings. The intuition behind the "activity" score is that if a latent dimension encodes relevant information and is not redundant, its value is expected to vary significantly over different inputs. By computing the score across all the test instances, every dimension of $\mathbf{\Lambda}$ is given an 'activity' value. Latent units with a higher value are considered more active and thus more informative. Figure 3 shows activity scores with respect to different prior distributions, including the standard Gaussian prior and priors learned on ATOMIC using MuRE or RGCN, in which the latent variables are the least active when using standard Gaussian as the prior distribution. The higher activation is obtained using the priors learnt on the external knowledge base. In particular, the prior based on RGCN and MuRE over ATOMIC displays the most active units, with RGCN showing the most active units on average. Table 5 shows the performance of the proposed model based on different priors. Two-sided Welch's t-test ($p < 0.05$) also supports that the RGCN-learned prior improves over standard Gaussian prior.

Uncertainty Quantification We present an analysis of uncertainty quantification of the Bayesian-Trans predictions. We adopted the uncertainty quantification methods as in Malinin and Gales (2018), computing the entropy (*total uncertainty*) and mutual information (*model uncertainty*) to visualize the predictive probabilities on a 2-simplex. Each forward pass on the same test instance is represented as a point on the simplex. For the sake of clarity of the visualization, we removed the EQUAL class, which is hardly ever predicted by the models.

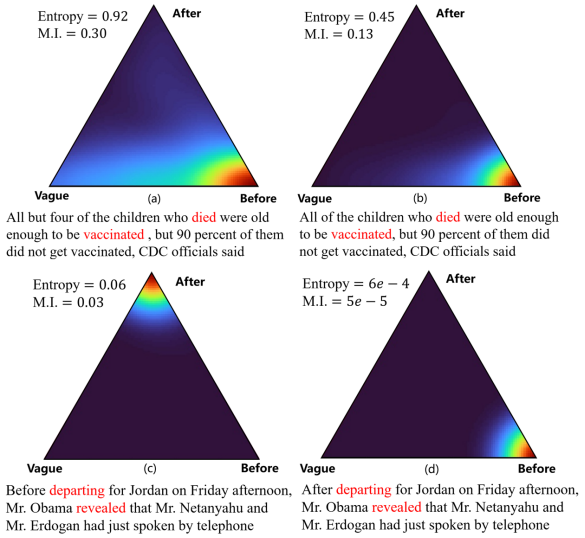

Figure 4: Examples of temporal relations in text and uncertainty quantification (entropy and mutual information) for the Bayesian-Trans model. Examples (a),(b) show how simplifying the linguistic structure without altering the temporal relation increases the model confidence. While examples (c),(d) illustrate the model's detection of temporal linguistic hints and its confidence.

In one of the test cases (Figure 4(a)), the true label is "*die*" BEFORE "*vaccinate*". This example exhibits a rather complex linguistic structure, as such, the model exhibits some uncertainty. Most of the predictions located at the corner are associated with BEFORE, but there also are several predictions scattered around it. We then simplified the sentence structure by removing "*but four*", and fed the modified sentence to the same model. This time, the model predicted the right temporal relation with much lower uncertainty (Figure 4(b)).

In another case study (Figure 4(c)), the true label is "*depart*" AFTER "*reveal*". This test case is rather straightforward, because of the explicit temporal word "before". The model predicted AFTER with high confidence, as shown by the predictive probabilities cluster at the top of the simplex. To show the impact of the temporal description, we swapped it from "*before*" to "*after*" and fed it to the same model. The model recognized the reversed meaning and correctly predicted BEFORE with low uncertainty (Figure 4(d)). The above cases demon-

strate that the proposed model reacts to different inputs with reasonable uncertainty, on both the total and model uncertainty scores.

6 Conclusion

We propose a principled approach to incorporate knowledge for event temporal relation extraction named Bayesian-Trans, which models the relation representations in the MuRE translational model, as latent variables. The latent variables are inferred through variational inference, during which commonsense knowledge is incorporated in the form of the prior distribution. The experiments on MATRES, TCR, and TBD show that Bayesian-Trans achieves state-of-the-art performance. Comprehensive analyses of the experimental results also demonstrate the characteristics and benefits of the proposed model.

Limitations

Our approach takes an event pair as input for the prediction of their temporal relation. We observe that if two events reside in different sentences, the error rate increases by 19%. A promising future direction is to construct a global event graph where temporal relations of any two events are refined with the consideration of global consistency constraints, for example, no temporal relation loop is allowed in a set of events. Our current work only deals with even temporal relations, it could be extended to consider other event semantic relations such as causal, hierarchical or entailment relations. The event temporal knowledge in this paper is derived from ATOMIC which can possibly be extended to more sources. Bayesian learning could also be extended to life-long learning. But we need to explore approaches to address the problem of catastrophic forgetting. We didn't exhaustively investigate all the translational models due to the large volume of work in that area. There might be a translational model which can achieve better performance, but the core idea of the proposed framework stays the same.

Ethical Considerations

The goal of the proposed method is to understand the temporal relation between events based on the descriptions in the given text. What the method can achieve in the most optimistic scenario is no more than giving the same text to a human reader and letting him or her explain the event relations. Therefore, the ethical concerns only come from the data collection. In this paper, we only use publicly available datasets which have already been widely used in the research field. As for potential application, as long as the user collects the training data legally, the proposed method does not have the potential to have a direct harmful impact.

Acknowledgements

This work was supported in part by 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). This work was conducted on the UKRI/EPSRC HPC platform, Avon, hosted in the University of Warwick's Scientific Computing Group. XT was partially supported by the Research Development Fund (RDF) 2022/23 (University of Warwick): '*An Event-Centric Dialogue System for Second Language Learners*'.

References

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Multi-relational poincaré graph embeddings. In *NeurIPS*.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2016. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859 – 877.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. 2016. Importance weighted autoencoders. In *International Conference on Learning Representations*.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.