Selective Temporal Knowledge Graph Reasoning

Zhongni Hou[1,2], **Xiaolong Jin**[1,2*], **Zixuan Li**[2*], **Long Bai**[2],
Jiafeng Guo[1,2], **Xueqi Cheng**[1,2]
[1]School of Computer Science and Technology, University of Chinese Academy of Sciences;
[2]Key Laboratory of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences.
{houzhongni18z, jinxiaolong, lizixuan,bailong18b,guojiafeng, cxq}@ict.ac.cn

Abstract

Temporal Knowledge Graph (TKG), which characterizes temporally evolving facts in the form of (subject, relation, object, timestamp), has attracted much attention recently. TKG reasoning aims to predict future facts based on given historical ones. However, existing TKG reasoning models are unable to abstain from predictions they are uncertain, which will inevitably bring risks in real-world applications. Thus, in this paper, we propose an abstention mechanism for TKG reasoning, which helps the existing models make selective, instead of indiscriminate, predictions. Specifically, we develop a confidence estimator, called Confidence Estimator with History (CEHis), to enable the existing TKG reasoning models to first estimate their confidence in making predictions, and then abstain from those with low confidence. To do so, CEHis takes two kinds of information into consideration, namely, the certainty of the current prediction and the accuracy of historical predictions. Experiments with representative TKG reasoning models on two benchmark datasets demonstrate the effectiveness of the proposed CEHis.

Keywords: Information Extraction, Knowledge Discovery/Representation, Question Answering

1. Introduction

Temporal Knowledge Graphs (TKGs), which store temporally evolving facts in the form of (subject, relation, object, timestamp) (Jin et al., 2019; Goel et al., 2020; Li et al., 2022a), have emerged as a very active research area over the last few years. Typically, a TKG can be denoted as a sequence of KG snapshots with timestamps, each of which contains all facts at the corresponding timestamp. The TKG reasoning task that aims to, given queries like (query entity, query relation, ?, future timestamp), conduct predictions about these future facts based on historical ones (Ding et al., 2021; Li et al., 2022c), has recently attracted more and more interest. It has also been increasingly used in various downstream time-sensitive applications, such as emerging event response (Muthiah et al., 2015; Phillips et al., 2017), policymaking (Deng et al., 2020) and disaster relief (Signorini et al., 2011).

Although existing models have achieved significant successes in the TKG reasoning task, they still inevitably make incorrect predictions due to the complex temporal dynamics in TKGs. The risk associated with incorrect predictions hinders the more extensive adoption of these models in real-world applications, especially some risk-sensitive applications such as disaster relief (Li et al., 2022a) and emergency response (Phillips et al., 2017). To better facilitate practical applications, it is necessary for the existing TKG reasoning models to have the ability to abstain from making uncertain, even incorrect, predictions.

* Corresponding authors.

This kind of ability to abstain from making certain predictions, also known as the selective prediction, has already been studied in the fields of image classification (Whitehead et al., 2022; Dancette et al., 2023) and text classification (Kuhn et al., 2023). To make selective predictions, existing studies equip the model with a confidence estimator, which estimates its confidence in the prediction and guides it to abstain from those with low confidence. Those existing studies estimate the confidence of the model mainly based on the final probability distribution of the current prediction. For instance, Geifman and El-Yaniv (2017) proposed SoftMax Response (SR) that utilizes the highest probability in the final probability distribution as the model's confidence score. Raina and Gales (2022) and Xin et al. (2021) utilized the entropy of the final probability distribution as the confidence score.

In TKG, besides the model's confidence in its current prediction, there usually exist some historical predictions that may also help the model decide whether or not to abstain. In fact, there are various historical queries that are relevant to the given query entity and query relation, even the same as the given query. Take the query (ISIS, Attack, ?, 2023-5-13) as an example, there are various related queries in the history, such as those (ISIS, Attack, ?, t) occurring before 2023-5-13. If the model can correctly make predictions for most of those related queries, it is very likely to make a correct prediction with high confidence for this query on 2023-5-13. This observation emphasizes the importance of leveraging the accuracy of predictions on historical queries to enhence the models' confidence in the

LREC-COLING 2024, pages 14555–14566

current prediction.

Motivated by the above issues, this paper studies for the first time the selective prediction setting for the TKG reasoning task. Specifically, we propose a confidence estimator, called Confidence Estimator with History (CEHis), for selective TKG reasoning. CEHis employs a certainty scorer to measure the certainty of the current prediction. It further uses a historical accuracy scorer to model the accuracy of historical predictions, by considering three types of related queries in the history, i.e., query entity related, query relation related, and both query entity and relation related, respectively. As intuitively the impact of the accuracy of the historical predictions on the confidence of the current prediction may decay over time, we employ the Hawkes process (Hawkes, 2018) to estimate this impact of long-term and short-term. Finally, CEHis leverages a ranking-based strategy to combine both the certainty score and the historical accuracy score to get the final confidence of the current prediction. Extensive experiments with representative TKG reasoning models on two benchmark datasets demonstrate the effectiveness of CEHis.

In summary, the contributions of this paper are as follows:

- To facilitate practical applications of TKG reasoning, it studies for the first time the selective TKG reasoning setting;
- It proposes a simple but effective confidence estimator for this task, which takes both the certainty of the current prediction and the accuracy of historical predictions on related queries into consideration;
- Experiments on two benchmark datasets demonstrate the necessity of the selective TKG reasoning setting and the superiority of the proposed confidence estimator.

2. Related Work

2.1. TKG Reasoning Methods

There are two different task settings for TKG reasoning, interpolation and extrapolation (Jin et al., 2020; Park et al., 2022; Cai et al., 2022; Messner et al., 2022; Liu et al., 2022). The interpolation setting aims to infer missing elements of facts at known timestamps in historical snapshots. In contrast, the extrapolation setting, which this paper focuses on, is to predict future facts.

Under the interpolation setting, HyTE (Dasgupta et al., 2018) proposes to conduct TKG reasoning task based on projected-time translation. DE-DistMult (Goel et al., 2020) and DE-SimplE (Goel et al., 2020) both utilize a diachronic embedding to generate entity representations at any given time. However, most interpolated TKG reasoning models perform worse when predicting future temporal facts. Under the extrapolation setting, RE-Net (Jin et al., 2020) and REGCN (Li et al., 2021b) both utilize the recurrent mechanism to capture the complex evolutional patterns among the facts in history. Besides, CyGNet (Zhu et al., 2020) utilizes a copy-generation mechanism to capture recurrence patterns of temporal facts. Considering that most TKG reasoning methods are black-box, TITer (Sun et al., 2021) and Cluster (Li et al., 2021a) further employ RL to adaptively find history paths, in order to provide interpretations for a specific prediction. More recently, TiRGN (Li et al., 2022a) and HisMatch (Li et al., 2022c) both design multi-encoders to model different characteristics of historical facts. CENET (Xu et al., 2023) further utilizes contrastive learning to identify significant entities from both historical and non-historical dependency. All these methods are encouraged to make predictions even wrong, leading to uncontrollable risks. Different from all these methods, we focus on making selective, instead of indiscriminate, predictions to control the risk of TKG reasoning in this work.

2.2. Selective Prediction

The selective prediction setting gives a model an option to abstain from generating certain predictions, which has been explored in different scenarios (Bartlett and Wegkamp, 2008; Grandvalet et al., 2008; Gal and Ghahramani, 2016; Cortes et al., 2016). When conducting selective prediction, a typical technique is to set a threshold over a confidence score derived from a pre-trained Neural Network. In 2017, Geifman and El-Yaniv (2017) proposed to selectively output using the well-known SR and MC-Dropout (Gal and Ghahramani, 2016) as selection strategies. Recently, SelectiveNet (Geifman and El-Yaniv, 2019) calculates a confidence score via an additional selection head to determine whether to abstain or not. Similarly, Deep Gamblers (Liu et al., 2019) and SAT (Huang et al., 2020) introduce an extra abstention class, the corresponding logit of this class determines whether a query is selected to predict or not. More recently, Feng et al. (2022) achieves better results via using the classification scores outputted by selective models with architectural change. Most previous work on selective prediction is mainly applied to CV tasks and the NLP field and mainly focuses on the probability of the current prediction. To the best of our knowledge, we are the first to utilize the characteristics of TKGs and apply selective prediction to the TKG reasoning scenario.

3. Problem Formulation

3.1. Formulation of TKG Reasoning

A TKG can be formalized as a sequence of KGs with timestamps, i.e., $\{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_t, ...\}$. The KG at timestamp t can be denoted as $\mathcal{G}_t = \{\mathcal{V}, \mathcal{R}, \mathcal{E}_t\}$, where $\mathcal{V}$, $\mathcal{R}$, $\mathcal{E}_t$ are the sets of entities, relations, and temporal facts occurring at timestamp t, respectively. Each fact in $\mathcal{E}_t$ is denoted as (s, r, o, t), where $s, o \in \mathcal{V}$ are the subject and object entities involved in this fact, $r \in \mathcal{R}$ is the relation between s and o. The TKG reasoning task aims to predict future facts based on given historical facts, which can be divided into two subtasks, namely, entity reasoning and relation reasoning. The former aims to predict the object (or subject) entity for a given query $q = (s_q, r_q, ?, t_q)$ (or $q = (?, r_q, o_q, t_q)$) based on the corresponding history before t_q, i.e., $\mathcal{G}_q = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_{t_q-1}\}$. The latter aims to predict the relation for a given query $q = (s_q, ?, o_q, t_q)$ based on $\mathcal{G}_q$.

3.2. Formulation of Selective TKG Reasoning

Typically, a selective TKG reasoning model f_s consists of three parts: a basic TKG reasoning model f, a confidence estimator g for evaluating the confidence of each prediction, and a threshold γ to determine if a prediction should be abstained based on the corresponding confidence. When a prediction is trustable, f_s outputs a ranked entity list generated by f, otherwise an empty list, i.e., $\emptyset$.

The basic TKG reasoning model. Given an input $x = (q, \mathcal{G}_q) \in \mathcal{X}$, where q denotes the query, $\mathcal{G}_q$ represents the corresponding history and $\mathcal{X}$ is the input space, the model f calculates the probability of the candidate $y \in \mathcal{Y}$ to be the correct answer, i.e., $p(y|x)$, where $\mathcal{Y}$ is the candidate answer space. Note that, for entity reasoning, $\mathcal{Y}$ is the set of entities $\mathcal{V}$, whilst for relation reasoning, $\mathcal{Y}$ is the set of relations $\mathcal{R}$. Finally, f outputs a list of all candidates, ranked in descending order according to their corresponding prediction probability.

The confidence estimator. The confidence estimator $g(x)$ is a positive real-valued function, which evaluates how confident the model $f(x)$ is on its prediction for a given input x. Ideally, g should obtain high values when f makes correct predictions, and otherwise, low values.

The threshold. Selective prediction seeks to make the trade-off between making correct predictions with high-confidence and abstaining from making low-confidence predictions to control potential risks. Therefore, f_s is equipped with a threshold γ to determine whether or not a prediction made by f should be trusted, and further control the overall level of abstention.

Above all, the selective TKG reasoning can be formulated as follows:

$$f_s(x) = \begin{cases} f(x), & if\ g(x) > \gamma, \\ \emptyset, & if\ g(x) \leq \gamma. \end{cases} \quad (1)$$

Obviously, the key in selective TKG reasoning is to assess the confidence of a prediction. Therefore, in this paper we develop a universal confidence estimator that can be readily integrated with existing TKG reasoning models, to estimate their predictions (see Section 4).

3.3. Evaluation Metrics of Selective TKG Reasoning

For selective prediction, *coverage*, *risk* and *effective reliability* are three widely adopted metrics (Geifman and El-Yaniv, 2017; Whitehead et al., 2022). In what follows, we formulate these metrics for the selective TKG reasoning task. Let $\mathcal{D} = \{(x_i, y_i^*)\}_{i=1}^{|\mathcal{D}|} \subseteq \mathcal{X} \times \mathcal{Y}$ be a set of inputs and their corresponding ground truth answers, i.e., y_i^* is the ground truth corresponding to the input $x_i = (q_i, \mathcal{G}_{q_i})$.

Coverage and Risk. The *coverage* (Geifman and El-Yaniv, 2017) of f_s on $\mathcal{D}$ is the proportion of predictions that are not abstained on the entire dataset, namely,

$$C(f_s, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i^*) \in \mathcal{D}} \mathbb{1}[g(x_i) > \gamma], \quad (2)$$

where $\mathbb{1}(P)$ is the indicator function that obtains 1 if P is true, otherwise 0.

As for the *risk* metric, let's first define that on a specific prediction. Obviously, for a given query, the higher the position of the corresponding ground truth in the final ranked list of candidate answers, the lower the risk of the model. Therefore, for a given input $x = (q, \mathcal{G}_q)$ and its corresponding ground truth y^*, its corresponding *risk* is formulated in this paper as follows:

$$risk_x(y^*) = \alpha(1 - \frac{1}{R_x(y^*)}), \quad (3)$$

where α is the risk parameter and $\alpha \geq 1$. Accordingly, the risk of f_s on $\mathcal{D}$ can thus be defined as:

$$R(f_s, \mathcal{D}) = \frac{\sum_{(x_i, y_i^*) \in \mathcal{D}} risk_{x_i}(y_i^*) \cdot \mathbb{1}[g(x_i) > \gamma]}{C(f_s, \mathcal{D})}. \quad (4)$$

Based on *coverage* and *risk*, the overall performance of f_s on $\mathcal{D}$ can be measured by the Area Under risk-coverage Curve (AUC), which plots risk against coverage (Geifman and El-Yaniv, 2017). And given a threshold γ, the lower the value of AUC, the better the performance as it represents lower average risk.

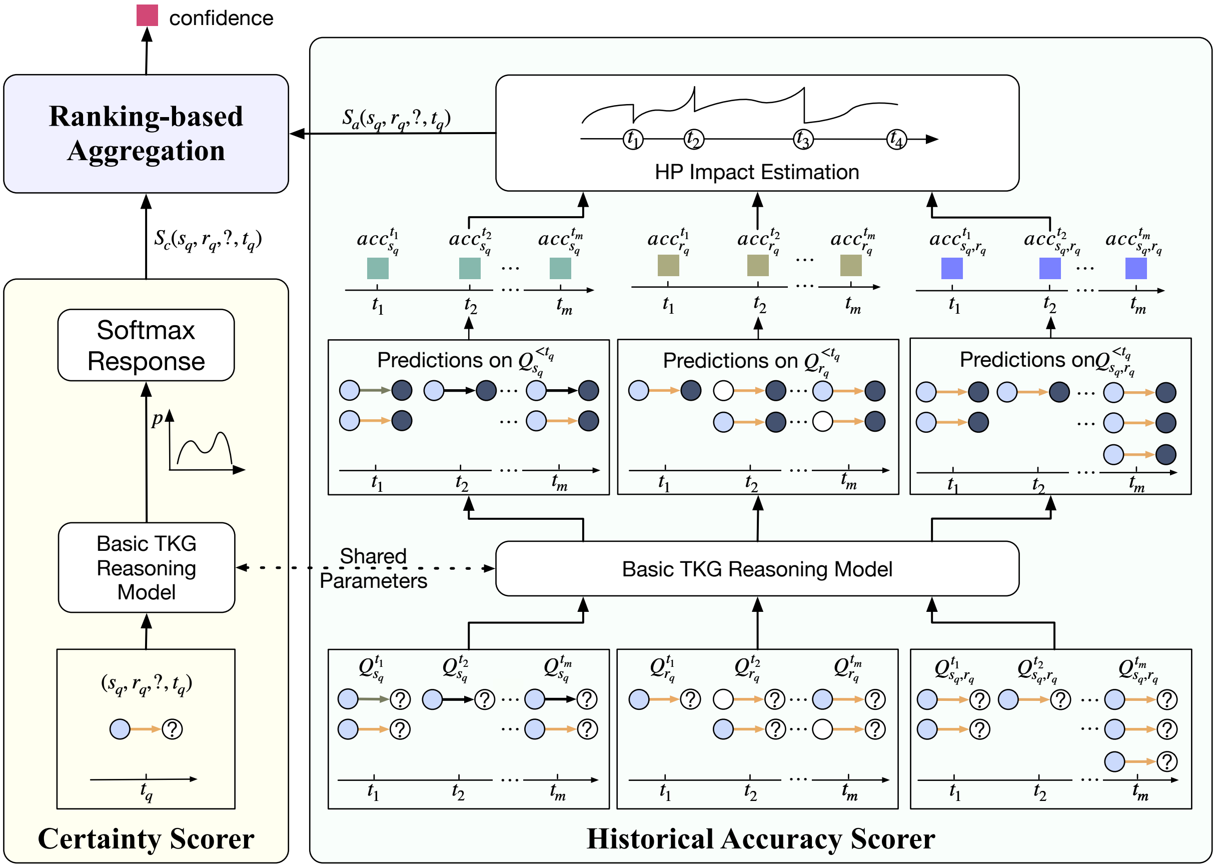


Figure 1: An illustrative diagram of the proposed confidence estimator, CEHis, for selective entity reasoning. For the sake of brevity, the corresponding history $\mathcal{G}_q$ paired with each query q is not explicitly given.

Effective Reliability. This metric is first proposed in Whitehead et al. (2022) to measure the effectiveness of a selective Visual Question Answering (VQA) model via assigning a penalty to the model when wrong predictions are outputted. Unlike VQA, the size of the candidate answer space in TKGs is relatively large, which increases the difficulty of making precise predictions. Motivated by this, in selective TKG reasoning, for a given input x and the corresponding ranked candidate list generated by the basic TKG reasoning model, we assign a penalty to f_s if the ground truth y^* is not in the top N positions, a reward to f_s if the ground truth y^* is in the top N positions, and no reward if f_s abstains from making a prediction. Here, N can be seen as the tolerance of the model. Formally, the *effective reliability* of f_s on a given input x in TKGs can be defined by:

$$\phi_{c,N}(f_s,x)=\begin{cases}\frac{1}{R_x(y^*)}, & if\ g(x)>\gamma\ \&\ R_x(y^*)< N,\\ -c, & if\ g(x)>\gamma\ \&\ R_x(y^*)\geq N,\\ 0, & if\ g(x)\leq\gamma,\end{cases} \tag{5}$$

where c is the penalty, and $\frac{1}{R_x(y^*)}$ denotes the corresponding reward. The *effective reliability* of f_s on the entire dataset $\mathcal{D}$ can thus be obtained as

$$\Phi_{c,N}(f_s,\mathcal{D})=\frac{1}{|\mathcal{D}|}\sum_{(x_i,y_i^*)\in\mathcal{D}}\phi_{c,N}(f_s,x_i). \tag{6}$$

Obviously, the higher the effective reliability, the better the performance of the TKG reasoning model.

4. Confidence Estimator with History

This section presents the proposed confidence estimator, i.e., CEHis, for selective TKG reasoning. As illustrated in Figure 1, CEHis mainly consists of two components, i.e., a certainty scorer and a historical accuracy scorer, to estimate the confidence score of a prediction generated by the basic TKG reasoning model f. The former measures the model's certainty of the current prediction, and the latter qualifies the impact of the accuracy of historical predictions on whether to abstain from the current prediction. Furthermore, it aggregates the above two kinds of information using a ranking-based strategy to determine the final confidence score, and subsequently, abstains from predictions with low confidence. In the following, we take selective entity reasoning as an example to illustrate how CEHis estimates the confidence of a prediction.

4.1. The Certainty Scorer

Given an input $x = (q, \mathcal{G}_q)$, existing TKG reasoning models usually utilize the softmax activation and finally output the probability of each entity to be the correct answer. Typically, the correct predictions tend to have greater maximum probabilities than those incorrect ones (Hendrycks and Gimpel, 2016). This characteristic can be utilized to estimate the level of certainty in the current prediction. As a result, in this paper, we adopt the widely used SR to measure the model's certainty of the current prediction, i.e., $S_c(x)$, as follows:

$$S_c(x) = \max_{o \in \mathcal{V}} p(o|x), \tag{7}$$

where $p(o|x)$ is the corresponding probability of the entity o to be the correct answer.

4.2. The Historical Accuracy Scorer

As mentioned above, a selective TKG reasoning model f_s abstains from the incorrect predictions made by the basic TKG reasoning model f to control the potential risks. However, whether a prediction on a query is correct or not is unknown, as the corresponding fact has not yet occurred. Typically, there are various historical queries that are relevant to the query entity and the query relation. The accuracy of the historical predictions regarding these related queries can reflect the difficulty of the current query, and further can serve as an indicator of whether to trust f's prediction on the current query. Motivated by this, the historical accuracy scorer estimates the accuracy of f's predictions based on the accuracy of historical predictions on three kinds of related queries, i.e., the subject related ones $Q_{s_q}^{<t_q}$, the relation related ones $Q_{r_q}^{<t_q}$, as well as the subject and relation related ones $Q_{s_q,r_q}^{<t_q}$. Considering that the accuracy of recent historical predictions is more important than older ones, the historical accuracy scorer utilizes the Hawkes process to model the time-varying impact of these historical predictions on the confidence of the current prediction, and finally calculates the historical accuracy score of the current prediction.

Specifically, the subject related queries at timestamp t_i consist of queries with s_q as the subject, and are denoted as $\mathcal{Q}_{s_q}^{t_i} = \{(s_q, -, ?, t_i)\}_{i=1}^{K_1}$. Here, "$-$" means that the corresponding element can be any relations, K_1 is the size of $\mathcal{Q}_{s_q}^{t_i}$, and $t_i < t_q$. Similarly, at timestamp t_i, the relation related queries are denoted as $\mathcal{Q}_{r_q}^{t_i} = \{(-, r_q, ?, t_i)\}_{i=1}^{K_2}$, the subject and relation related queries are denoted as $\mathcal{Q}_{s_q,r_q}^{t_i} = \{(s_q, r_q, ?, t_i)\}_{i=1}^{K_3}$. Taking $\mathcal{Q}_{s_q}^{<t_q}$ as an example, we illustrate how to calculate the impact of the historical predictions on the confidence of the current prediction, i.e., $S_{s_q}^{<t_q}$. Given a query $\tilde{q} = (s_q, r_i, ?, t_i) \in \mathcal{Q}_{s_q}^{t_i}$, CEHis first utilizes the basic TKG reasoning model f to make a prediction based on the corresponding history $\mathcal{G}_{\tilde{q}}$, and calculates the position of the ground truth $o_{\tilde{q}}^*$, i.e., $R_{\tilde{q},\mathcal{G}_{\tilde{q}}}(o_{\tilde{q}}^*)$, over the ranked entity list generated by f. At timestamp t_i, the accuracy of historical predictions on the subject related queries, i.e., $acc_{s_q}^{t_i}$, is calculated by:

$$acc_{s_q}^{t_i} = \frac{1}{|\mathcal{Q}_{s_q}^{t_i}|} \sum_{\tilde{q} \in \mathcal{Q}_{s_q}^{t_i}} \frac{1}{R_{\tilde{q},\mathcal{G}_{\tilde{q}}}(o_{\tilde{q}}^*)}. \tag{8}$$

After processing all historical snapshots, we can obtain a prediction accuracy sequence of $\mathcal{Q}_{s_q}^{<t_q}$, i.e., $Acc_{s_q}^{<t_q} = \{acc_{s_q}^{t_1}, ..., acc_{s_q}^{t_i}, ..., acc_{s_q}^{t_m}\}$, where $t_1 < ... < t_i < ... < t_m < t_q$.

To precisely estimate whether the prediction made by f is correct or not, CEHis considers the impact of both long-term and short-term accuracy regarding historical predictions. Typically, the accuracy of the historical predictions on a recent timestamp is more important than that on an earlier timestamp. As a result, it is necessary to take the time information into consideration. Motivated by this, CEHis utilizes the Hawkes process ("HP Impact Estimation" in Figure 1; Cox and Isham (1980); Laub et al. (2015); Zuo et al. (2020)) to estimate the impact of the accuracy of historical predictions on whether to abstain or not as follows:

$$S_{s_q}^{<t_q} = \mu_{s_q}^{<t_q} + \sum_{h=0}^{l-1} k(t_q - t_{m-h}) acc_{s_q}^{t_{m-h}}, \tag{9}$$

where $\mu_{s_q}^{<t_q}$ represents the base prediction accuracy (the long-term accuracy) of the subject related queries, which is calculated by the mean of $Acc_{s_q}^{<t_q}$; l is a length hyperparameter which is used to truncate the prediction accuracy sequence $Acc_{s_q}^{<t_q}$. $k(\cdot)$ is a predefined decaying function, calculating the decaying impact of the historical accuracy:

$$k(t_q - t_h) = exp(-\delta(t_q - t_h)), \tag{10}$$

where $\delta \geq 0$ denotes the decay rate. Besides the above absolute time interval, we can also choose the relative time order information (Zhang et al., 2022), which can be seen as a normalization, to calculate the decaying impact of the historical accuracy. Obviously, the cumulative term describes that the historical prediction accuracy of the latest timestamps (the short-term accuracy) has a positive contribution to whether to trust f's current prediction. Similarly, at the query timestamp t_q, we can derive the impact of the accuracy of historical predictions regarding $\mathcal{Q}_{r_q}^{<t_q}$ and $\mathcal{Q}_{s_q,r_q}^{<t_q}$, namely, $S_{r_q}^{<t_q}$ and $S_{s_q,r_q}^{<t_q}$, respectively.

For a given input $x = (q, \mathcal{G}_q)$, the final historical accuracy score is calculated as:

$$S_a(x) = S_{s_q}^{<t_q} + S_{r_q}^{<t_q} + S_{s_q,r_q}^{<t_q}. \tag{11}$$

Datasets	#Train	#Valid	#Test	#Ent	#Rel
ICEWS14	74,845	8,514	73,71	6,869	230
ICEWS18	373,018	45,995	49,545	23,033	256

Table 1: Statistics of the datasets.

4.3. The Ranking-based Aggregation

To calculate the final confidence scores of predictions made by f, we need to aggregate two kinds of scores outputted by the above two scorers. However, the certainty score and the historical accuracy score are calculated based on different information and are on different scales. As a result, they cannot be directly combined using absolute values. To this end, CEHis utilizes a ranking-based aggregation strategy, which first ranks queries according to the above two different scores, and then calculates the final confidence score of two based on the results of the two rankings.

More specifically, let $\mathcal{Q}_{un} = \{(q_i, \mathcal{G}_{q_i})\}_{i=1}^{|\mathcal{Q}_{un}|}$ denote a set of query-history pairs that require predictions with unknown distribution. Given an input $x = (q, \mathcal{G}_q) \in \mathcal{Q}_{un}$, we can derive the model's certainty rank, i.e., $R_{c,\mathcal{Q}_{un}}(x)$, and the historical accuracy rank, i.e., $R_{a,\mathcal{Q}_{un}}(x)$, of its prediction among $\mathcal{Q}_{un}$:

$$R_{c,\mathcal{Q}_{un}}(x) = \sum_{x_i \in \mathcal{Q}_{un}} \mathbb{1}(S_c(x_i) > S_c(x)), \quad (12)$$

$$R_{a,\mathcal{Q}_{un}}(x) = \sum_{x_i \in \mathcal{Q}_{un}} \mathbb{1}(S_a(x_i) > S_a(x)). \quad (13)$$

The total confidence score generated by our confidence estimator g is defined as:

$$g(x) = \beta R_{c,\mathcal{Q}_{un}}(x) + (1-\beta) R_{a,\mathcal{Q}_{un}}(x), \quad (14)$$

where β is an aggregation weight that can be set using the validation dataset.

5. Experiments

5.1. Experiment Setting

Datasets. We conduct experiments on two widely used TKGs, namely, ICEWS14 (Li et al., 2022c) and ICEWS18 (Li et al., 2022a). They both, with the time granularity of 24 hours, are subsets of facts in the Integrated Crisis Early Warning System (ICEWS). Specifically, ICEWS14 and ICEWS18 contain facts that took place between 2014 and 2018, respectively. The statistics of these datasets is presented in Table 1.

Basic TKG Reasoning Models. As aforementioned, the proposed confidence estimator, CEHis, can be applied to a variety of existing TKG reasoning models. Here, we take the following three representative TKG reasoning models with varying architectures and performance, as the basic models: RENET (Jin et al., 2020), which employs the recurrent GCN to model the histories of queries; REGCN (Li et al., 2021b), which stacks GCN layers to mine evolutional patterns of each entity among facts occurring at the latest timestamps; TiRGN (Li et al., 2022a), which utilizes an additional encoder to capture the structure dependency among repetitive history and is the most relevant model to our confidence estimator.

Baselines. Since the selective TKG reasoning task has not been explored before, we adopt a few representative confidence estimators used in other tasks as our baselines, including SR (Geifman and El-Yaniv, 2017), entropy (EN; Geifman and El-Yaniv (2017)), SelectiveNet+SR (SNR; Feng et al. (2022), SAT+SR (SATR; Feng et al. (2022)) and Deep Sub-Ensembles (SE; Valdenegro-Toro (2023)). Specifically, SNR and SATR are built upon SelectiveNet (Geifman and El-Yaniv, 2019) and SAT (Huang et al., 2020), respectively. They both train a more general classifier by changing the model architectures, and utilize the maximum probability of the classifier itself to conduct selective prediction. SE ensembles only a selection of the model's layers that are close to the output, and estimates the confidence using the predictive uncertainty.

Implementation Details. We set both the decay rate δ and the aggregation weight β within [0,1], and α to be one. The long-term accuracy of historical predictions is calculated based on the latest 10 timestamps, and the short-term accuracy is calculated based on the latest 3 timestamps. All basic TKG reasoning models are trained using their reported optimal parameters.

5.2. Experimental Results

5.2.1. Results on Selective Entity Reasoning

To examine the effectiveness of CEHis on the selective entity reasoning task, we focus on measuring the max coverage at different risk levels, the AUC for the risk-coverage curve, and the effective reliability scores under different penalty and tolerance settings. The results for the former two metrics are presented in Table 2. Due to space limitation, for the latter metric, we only report the corresponding results on ICEWS14 in Figure 2.

From Table 2, it can be observed that CEHis outperforms all baseline methods, in terms of AUC and the coverage under different risk levels in most cases. This is because all baseline methods do not model the accuracy of historical predictions on related queries. Since those related queries in the history are similar to the current one, their accuracy can serve as an indicator of the model's ability to make precise predictions for the current query.

Model	Confidence Estimator	ICEWS14				ICEWS18			
		coverage			AUC	coverage			AUC
		risk=0.1	risk=0.3	risk=0.5		risk=0.1	risk=0.3	risk=0.5	
RENET	EN	1.81	17.68	61.90	43.34	0.01	4.99	22.64	56.82
	SR	1.46	14.62	62.15	43.01	0.01	4.16	22.26	56.76
	SNR	1.24	17.16	62.76	42.84	**0.22**	3.05	21.99	56.95
	SATR	1.37	16.70	62.84	42.76	0.02	4.04	22.44	57.03
	SE	1.78	15.32	60.19	43.65	0.02	4.07	21.25	57.38
	CEHis	**3.48**	**24.97**	**63.75**	**40.94**	0.08	**7.43**	**27.05**	**55.14**
REGCN	EN	2.00	23.11	69.38	40.16	0.24	6.76	30.51	53.20
	SR	1.98	23.33	68.33	40.26	0.61	7.09	31.88	52.76
	SNR	1.57	24.87	68.92	40.04	0.52	6.09	31.05	53.37
	SATR	3.71	25.20	68.93	39.67	0.28	6.33	29.58	53.86
	SE	**5.14**	23.99	61.73	42.49	0.13	4.96	26.39	54.84
	CEHis	4.86	**27.51**	**69.49**	**38.76**	**0.75**	**8.90**	**33.58**	**51.92**
TiRGN	EN	3.79	26.18	75.34	38.41	0.00	5.96	29.76	52.82
	SR	4.15	29.83	76.56	37.41	0.00	6.72	31.81	52.27
	SNR	4.23	28.92	72.70	37.86	0.17	7.52	33.40	51.85
	SATR	4.30	26.84	72.56	38.36	0.06	6.75	30.47	52.76
	SE	5.14	23.99	72.19	39.42	0.01	5.79	30.23	52.97
	CEHis	**5.19**	**30.32**	**76.77**	**36.97**	**0.26**	**8.16**	**34.60**	**51.50**

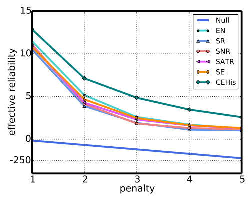
Table 2: Risk-coverage metrics results and AUC results of the selective entity reasoning task. The risk R is set to 0.1, 0.3, and 0.5.

Therefore, by modeling the impact of the accuracy of historical predictions on the confidence of the current prediction, CEHis outperforms existing confidence estimators.

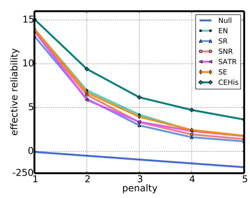
On ICEWS18, we notice that TiRGN with SR has lower coverage than REGCN with SR when the risk is 0.1 and 0.3. Considering that TiRGN has a higher parameter complexity of 13.98M (about 1.5x greater than that of REGCN), we guess TiRGN is overconfident in its incorrect predictions, which results in the above performance gap. The proposed CEHis alleviates the overconfidence problem in the complex TiRGN model, as the coverage of TiRGN with GEHis has been significantly improved when the risk level is 0.1 and 0.3.

Figure 2 shows that CEHis achieves the highest effective reliability scores across all penalty and tolerance levels. Also, it can be observed that when the penalty c increases from 1 to 5, the effective reliability scores decrease quickly. For instance, in Figure 2(a), the RENET with CEHis has $\Phi_{1,5} > 10$ and $\Phi_{2,5} < 5$. It suggests that when the penalty for a wrong answer is high, the selective model will be more cautious and abstain from more predictions.

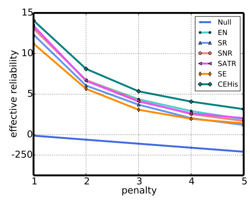
Additionally, we can observe that the performance of the selective TKG reasoning model is positively correlated with the model tolerance. For instance, in Figures 2(a) and 2(b), the RENET with CEHis has $\Phi_{1,5} < 15$ and $\Phi_{1,10} > 15$. A larger model tolerance N allows the model to make more predictions, thus increasing the number of correct predictions. Furthermore, it can be noticed that the models without a confidence estimator consistently perform poorly when compared with their selective model counterparts. As the penalty increases, the performance gap between them becomes greater, which illustrates the necessity of empowering the TKG reasoning model with the ability to abstain from making predictions.



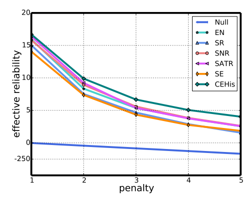
(a) RENET, N=5.



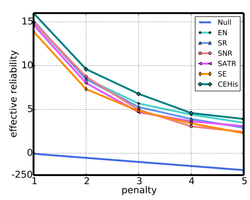
(b) RENET, N=10.



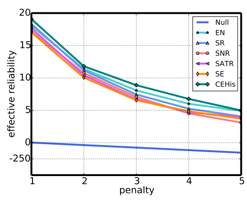
(c) REGCN, N=5.



(d) REGCN, N=10.



(e) TiRGN, N=5.



(f) TiRGN, N=10.

Figure 2: Effective reliability results of the selective entity reasoning task on ICEWS14. The penalty c is set to 1, 2, 3, 4 and 5, while the model's tolerance N is set to 5 and 10, respectively.

Model	EN	SR	SNR	SATR	SE	CEHis
RENET	43.37	41.33	41.79	43.61	43.89	**40.23**
REGCN	42.13	39.91	39.51	39.09	42.60	**38.68**
TiRGN	39.33	36.83	37.41	36.79	39.46	**36.60**

Table 3: AUC results of the selective relation reasoning task on ICEWS14. The risk R is set to 0.1, 0.3 and 0.5.

5.2.2. Results on Selective Relation Reasoning

To verify the effectiveness of CEHis on the selective relation reasoning task, we compare it with other confidence estimators. Note that, for the selective relation reasoning task, CEHis focuses on the following three kinds of related queries, i.e., subject related, object related, and both subject and object related. Due to space limitation, we only report the AUC results on ICEWS14 in Table 3. We can see that CEHis performs better than baselines with different basic TKG reasoning models, which demonstrates again that modeling the accuracy of historical predictions is helpful for obtaining more accurate confidence scores in the selective TKG reasoning task.

5.3. Ablation Study

To understand the behavior of CEHis with different basic TKG reasoning models on the selective TKG entity reasoning task, we take both RENET and TiRGN as the basic TKG reasoning models and conduct ablation studies on the ICEWS14 dataset. The corresponding AUC results are presented in Figure 3. In the following, we take RENET as an example, and analyze how each part of CEHis contributes to its performance with RENET.

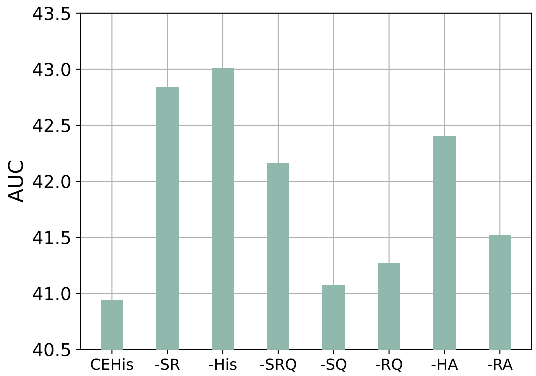
From Figure 3(a), we can observe that without considering the certainty of the current prediction (denoted as -SR) causes a drastic AUC increase, which indicates that the probability outputted by the basic TKG reasoning model is important, and should be taken into consideration when conducting the selective prediction task.

The result denoted as -His demonstrates the performance of CEHis without modeling the accuracy of historical predictions on related queries. It can be seen that, when employing RENET as the basic model, -His generates a higher AUC on ICEWS14, which justifies the necessity of modeling the accuracy of historical predictions.

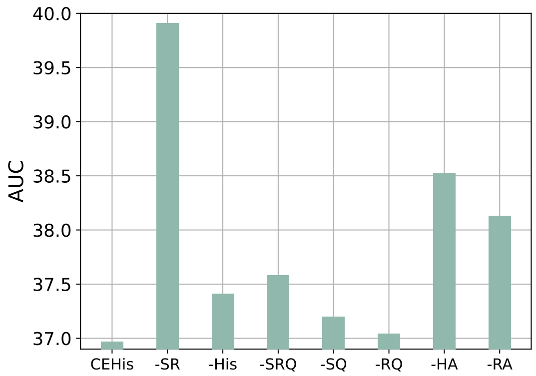
To further analyze the importance of three kinds of related queries, we ignore the accuracy of historical predictions on both subject and relation related queries, subject related queries and relation related queries, denoted as -SRQ, -SQ and -RQ, respectively. As shown in Figure 3(a), -SRQ brings the most significant performance drop when compared with -SQ and -RQ. This can be attributed to the fact that both subject and relation related queries contain the most useful information, which is also verified by existing TKG reasoning models (Zhu et al., 2020).

In addition, to verify the effectiveness of the Hawkes process (-HA in Figure 3(a)), we use the mean operation over the historical accuracy sequence of each kind of related queries. It can be seen that removing this part yields worse results compared to CEHis, demonstrating the necessity of utilizing the Hawkes process to model both the long-term and short-term impact of historical prediction accuracy.

To verify the effectiveness of the ranking-based aggregation (-RA in Figure 3(a)), we simply add the absolute value of the model's certainty of the current prediction and the accuracy of historical predictions on related queries. It can be seen that -RA results in worse performance compared to CEHis, demonstrating that the ranking-based strategy can



(a) Results by different variants of CEHis with RENET.



(b) Results by different variants of CEHis with TiRGN.

Figure 3: Comparison of variant models of CEHis with different basic TKG reasoning models on ICEWS14.

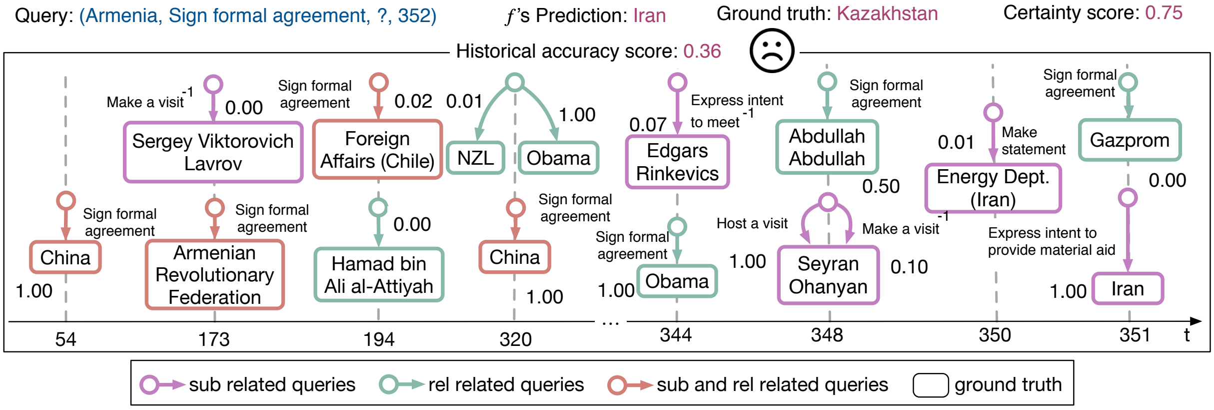


Figure 4: Case study on the necessity of modeling the accuracy of historical predictions. Each number represents how precise f's historical prediction is on the corresponding query.

help better aggregate the scores outputted by the certainty scorer and the historical accuracy scorer.

When taking TiRGN as the basic reasoning model, we can derive the same conclusion on different variants of CEHis, except for -His. From Figure 3(a) and Figure 3(b), we observe that -His has a reduced influence on the final AUC result when employing TiRGN as the basic model compared to utilizing RENET. This is because TiRGN captures valuable information within repeated historical facts. This information can help TiRGN precisely predict queries with repeated ones in the history. As a result, the impact of the historical accuracy scorer on TiRGN is weakened.

5.4. Case Study

In order to further show the necessity of modeling the accuracy of historical predictions, we present a case study in Figure 4 where the basic TKG reasoning model f makes a wrong prediction. It can be observed that SR assigns a high confidence score (0.75) to the current prediction. However, the historical predictions on related queries are of low accuracy, which indicates that trusting the current prediction made by the basic TKG reasoning model may bring risk. CEHis utilizes the historical accuracy scorer to capture the accuracy of historical predictions, and guide the selective TKG reasoning model to abstain from making the current prediction. As a result, the risk brought by incorrect predictions can be controlled.

6. Conclusions

In this paper, we introduced the selection prediction setting for TKG reasoning, where a model is allowed to abstain in order to avoid making incorrect predictions. We further proposed a confidence estimator, called CEHis, to conduct the selective TKG reasoning task. CEHis considers both the certainty of the current prediction and the accuracy of historical predictions on related queries, and employs the Hawkes process to model the time-varying impact of the accuracy of historical predictions. Finally, we demonstrated the effectiveness of CEHis upon comparison with other confidence estimators by applying them to existing TKG reasoning models.

7. Acknowledgments

The work is supported by the National Natural Science Foundation of China under grant 62306299, the National Key Research and Development Project of China, Beijing Academy of Artificial Intelligence under grant BAAI2019ZD0306, the KGJ Project under grant JCKY2022130C039, and the Lenovo-CAS Joint Lab Youth Scientist Project. We thank anonymous reviewers for their insightful comments and suggestions.

8. Bibliographical References

Peter L Bartlett and Marten H Wegkamp. 2008. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8).

Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, and Jianxin Li. 2022. Temporal knowledge graph completion: A survey. *arXiv preprint arXiv:2201.08236*.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29.

David Roxbee Cox and Valerie Isham. 1980. *Point processes*, volume 12. CRC Press.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. 2023. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059.

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2011.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2019. Learning dynamic context graphs for predicting social events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1007–1016.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2020. Dynamic knowledge graph based multi-event forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1585–1595.

Zifeng Ding, Zhen Han, Yunpu Ma, and Volker Tresp. 2021. Temporal knowledge graph forecasting with neural ode. *arXiv preprint arXiv:2101.05151*.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H Abdi. 2022. Towards better selective classification. In *The Eleventh International Conference on Learning Representations*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.

Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. 2020. Diachronic embedding for temporal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3988–3995.

Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. 2008. Support vector machines with a reject option. *Advances in neural information processing systems*, 21.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2020a. Dyernie: Dynamic evolution of riemannian manifold embeddings for temporal knowledge graph completion. *arXiv preprint arXiv:2011.03984*.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2020b. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *International Conference on Learning Representations*.

Zhen Han, Zifeng Ding, Yunpu Ma, Yujia Gu, and Volker Tresp. 2021. Learning neural ordinary equations for forecasting future links on temporal knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8352–8364.

Zhen Han, Yunpu Ma, Yuyi Wang, Stephan Günnemann, and Volker Tresp. 2020c. Graph hawkes neural network for forecasting on temporal knowledge graphs. *8th Automated Knowledge Base Construction (AKBC)*.

Alan G Hawkes. 2018. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Lang Huang, Chao Zhang, and Hongyang Zhang. 2020. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33:19365–19376.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. In *EMNLP*.

Woojeong Jin, Changlin Zhang, Pedro Szekely, and Xiang Ren. 2019. Recurrent event network for reasoning over temporal knowledge graphs. *arXiv preprint arXiv:1904.05530*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Patrick J Laub, Thomas Taimre, and Philip K Pollett. 2015. Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Julien Leblay, Melisachew Wudage Chekol, and Xin Liu. 2020. Towards temporal knowledge graph embeddings with arbitrary time precision. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 685–694.

Yujia Li, Shiliang Sun, and Jing Zhao. 2022a. Tirgn: time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2152–2158. ijcai. org.

Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022b. Complex evolutional pattern learning for temporal knowledge graph reasoning. *arXiv preprint arXiv:2203.07782*.

Zixuan Li, Zhongni Hou, Saiping Guan, Xiaolong Jin, Weihua Peng, Long Bai, Yajuan Lyu, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022c. Hismatch: Historical structure matching based temporal knowledge graph reasoning. *arXiv preprint arXiv:2210.09708*.

Zixuan Li, Xiaolong Jin, Saiping Guan, Wei Li, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng. 2021a. Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs. *arXiv preprint arXiv:2106.00327*.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021b. Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 408–417.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. 2022. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4120–4127.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2019. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32.

Johannes Messner, Ralph Abboud, and Ismail Ilkan Ceylan. 2022. Temporal knowledge graph completion using box embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7779–7787.

Fatemeh Sadat Minaye Hashemi, Chaiya Prasittichai, and Stacey F Bent. 2015. Self-correcting process for high quality patterning by atomic layer deposition. *Acs Nano*, 9(9):8710–8717.

Sathappan Muthiah, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, and Naren Ramakrishnan. 2015. Planned protest modeling in news and social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pages 3920–3927.

Namyong Park, Fuchen Liu, Purvanshi Mehta, Dana Cristofor, Christos Faloutsos, and Yuxiao Dong. 2022. Evokg: Jointly modeling event time and network structure for reasoning over temporal knowledge graphs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 794–803.

Lawrence Phillips, Chase Dowling, Kyle Shaffer, Nathan Hodas, and Svitlana Volkova. 2017. Using social media to predict the future: a systematic literature review. *arXiv preprint arXiv:1706.06134*.

Vatsal Raina and Mark Gales. 2022. Answer uncertainty and unanswerability in multiple-choice machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1020–1034.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. 2021. Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. *arXiv preprint arXiv:2109.04101*.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2018. Dyrep: Learning representations over dynamic graphs.

Matias Valdenegro-Toro. 2023. Sub-ensembles for fast uncertainty estimation in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4119–4127.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In

Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 950–958.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. 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 1040–1051.

Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. 2023. Temporal knowledge graph reasoning with historical contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4765–4773.

Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. 2022. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4741–4753.

Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhan. 2020. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. *arXiv preprint arXiv:2012.08492*.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR.