INTERPRETABLE MULTI-HOP REASONING FOR FORECASTING ON TEMPORAL KNOWLEDGE GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal knowledge graphs (KGs) have recently attracted growing attention. The temporal KG forecasting task, which plays a crucial role in applications such as event prediction, is predicting future links based on historical facts. The interpretability of the current temporal KG forecasting models is manifested in providing the reasoning paths. However, the comparison of reasoning paths is operated under the black box. Inspired by the observation that reasoning based on multi-hop paths is equivalent to answering questions step by step, this paper designs an Interpretable Multi-hop Reasoning (IMR) model for temporal KG forecasting. IMR transforms reasoning based on path searching into step-by-step question answering. Moreover, IMR designs three indicators according to the characteristics of temporal KGs and reasoning paths: question matching degree, answer completing level and path confidence. Unlike other models that can only utilize paths with a specified hop, IMR can effectively integrate paths of different hops; IMR can provide the reasoning paths like other interpretable models and further explain the basis for path comparison. While being more explainable, IMR has achieved state-of-the-art on four baseline datasets.

1 INTRODUCTION

Knowledge Graphs (KGs) are collections of triples, such as Freebase (Bordes et al., 2013), YAGO (Suchanek et al., 2008). Temporal KGs introduce a new dimension into static knowledge graphs (Li et al., 2021), i.e., a timestamp for every triple to form a quadruple. Although there are billions of triples in KGs, they are still incomplete. These incomplete knowledge bases will bring limitations to practical applications. Since temporal KGs involve the time dimension, the completion of temporal KGs can be divided into interpolation and forecasting. The former utilizes the facts of all timestamps to predict the triples at a particular moment; the latter employs historical facts to predict the future triples. Due to the importance of temporal KG forecasting in event prediction, it has recently attracted growing attention. This paper mainly focuses on temporal KG forecasting.

Most of the current researches on the temporal KG completion focus on interpolation (Jin et al., 2019; García-Durán et al., 2018; Xu et al., 2020a; Jung et al., 2021; Han et al., 2021; Wu et al., 2020). There have been recent attempts to investigate temporal KG forecasting (Jin et al., 2019; Han et al., 2021; Li et al., 2021; Zhu et al., 2020). According to the interpretability, researches on temporal KG forecasting can be divided into two categories. One type is the black-box model, which just design a unexplainable scoring function for evaluating the rationality of quadruples. The other type is interpretable approaches. CyGNet (Zhu et al., 2020) utilizes 1-hop repetitive facts to realize prediction. The performance of CyGNet is limited by lacking direct repetitive knowledge in historical moments. xERTR (Han et al., 2021), CluSTeR (Li et al., 2021) and TITer (Sun et al., 2021) are both path-based temporal KG forecasting models. xERTR (Han et al., 2021) adopts the inference subgraphs to aggregate local information around the query. CluSTeR (Li et al., 2021) and TITer (Sun et al., 2021) manipulate reinforcement learning for path search and improves the performance through temporal reasoning.

However, these models cannot effectively combine paths with different hops for reasoning. xERTR (Han et al., 2021) can only utilize the path with the specified hop. The experiments of CluSTeR (Li et al., 2021) illustrate that the performance with the maximum hop set to 2 is not as good as that with the maximum hop set to 1, which is not in line with common sense. Although these models can present the reasoning paths, they lack an explanation of the preference for various paths, i.e., the models cannot provide the basis for path comparison.

In practice, forecasting based on path searching is to find the appropriate multi-hop paths, the combination of whose relations is equivalent to the query's relation. As we observe, reasoning based on multi-hop paths is similar to answering questions step-by-step. Inspired by step-by-step question answering, this paper designs a new Interpretable Multi-hop Reasoning model - IMR, which can perform interpretable operations on reasoning and integrate paths of different hops simultaneously.

The primary route of IMR can be described as follows. IMR first transforms reasoning based on path searching into step-by-step question answering based on TransE (Bordes et al., 2013) and IRN (Zhou et al., 2018). When searching for multi-hop paths step by step, we calculate the rest part of the query for each path. Besides, IMR designs three indicators based on the remaining parts of the query and the reasoning tails: query matching degree, answer completing level and path confidence. Query matching degree, that is, the matching degree between the reasoning tails and the query, measures the rationality of the new quadruples. Answer completing level, that is, the matching degree between the relations of paths and that of the queries, measures the completeness of the answer. Path confidence, that is, the difference between the same entities with different timestamps, measures the reliability of the reasoning paths. IMR achieves the unified scoring of multi-hop paths and better explainable reasoning simultaneously with these indicators' combined effect.

The major contributions of this work are as follows. (1) This paper proposes a new interpretable multi-hop reasoning model (IMR) which can perform interpretable operations on the question. Furthermore, IMR designs three indicators: query matching degree, answer completing level and path confidence. (2) Unlike other models that can only process paths with a specified hop, IMR can measure paths of different hops equivalently and utilize paths with different hops for reasoning. (3) IMR can provide the reasoning path like other interpretable models and further explain the basis for path comparison. (4) Experiments show that IMR achieves state-of-the-art on four benchmark datasets.

2 Related Work

Static KG reasoning. Knowledge graph reasoning based on representation learning has been widely concerned by scholars. These approaches for reasoning can be categorized into geometric models (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015; Ji et al., 2015; Sun et al., 2019), tensor decomposition models (Yang et al., 2015; Nickel et al., 2011; Trouillon et al., 2016; Balazevic et al., 2019) and deep learning models (Dettmers et al., 2018; Nguyen et al., 2018; 2019). In recent years, some scholars have attempted to introduce GCN into knowledge graph reasoning (Vashishth et al., 2020), which can improve the performance of basic models. Some other scholars focus on multi-hop reasoning with symbolic inference rules learned from relation paths (Li & Cheng, 2019; Wang et al., 2019). The above methods are all designed for static KGs, challenging to deal with temporal knowledge graphs reasoning.

Temporal KG reasoning. Temporal KGs import the time dimension to static KGs, which makes the facts of a specific timestamp extremely sparse. The temporal KG reasoning task can be divided into two categories: reasoning about historical facts (Jin et al., 2019; García-Durán et al., 2018; Xu et al., 2020a; Jung et al., 2021; Han et al., 2021; Wu et al., 2020), i.e., interpolation on temporal KGs, and reasoning about future facts (Jin et al., 2019; Han et al., 2021; Li et al., 2021; Zhu et al., 2020), i.e., forecasting on temporal KGs. The former predicts the missing facts of a specific historical moment based on the facts of all moments, and the latter predicts future events based only on the past facts. There are many studies on the task of temporal KG interpolation. However, these studies are all black-box models, which cannot explain predictions. For temporal KG forecasting, most of the proposed models are also black box models. Recently, xERTR (Han et al., 2021), CluSTer (Li et al., 2021) and TITer (Sun et al., 2021) can explain predictions to some extent. These models can provides the reasoning paths for the predictions. However, both models can not integrate multi-hop paths. xERTR can only perform reasoning with a specified hop. Experiments show that CluSTeR performs worse on paths with multiple hops than on paths with only one hop.

In general, most of the current temporal KG forecasting models are black-box models. Only some models can provide reasoning paths for prediction. Moreover, none of them can explain how path comparisons work and none of them can combine reasoning paths with different hops effectively.

3 PRELIMINARIES

The task of temporal KG forecasting. Suppose that $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ represent the entity set, predicate set and timestamp set, respectively. The temporal KG is a collection of quadruples, which can be expressed as:$K = \{(e_s, r_q, e_o, t_q), e_s, e_o \in \mathcal{E}, r_q \in \mathcal{R}, t_q \in \mathcal{T}\}$. (e_s, r_q, e_o, t_q) denotes a quadruple, e_s and e_o represent the subject and object respectively, r_q represents the predicate, i.e. the relation, and t_q represents the time when the quadruple occurs. Suppose that the set of facts that happened before the time t_q can be expressed as $G_{t_q} = \{(e_i, r_i, e_j, t_i) \in K | t_i < t_q\}$. Temporal KG forecasting is to predict future links based on past facts, that is, the process of predicting e_o given a query $(e_s, r_q, ?, t_q)$ and the previous facts G_{t_q}. Similar to static KG reasoning, temporal KG forecasting ranks all entities of the specific moment and obtains the preference for prediction.

Temporal KG forecasting Based on Paths. Knowledge graph embedding associates the entities $e \in \mathcal{E}$ and relations $r \in \mathcal{R}$ with vectors $\mathbf{e}, \mathbf{r}$. Different from static KGs, the entities in temporal KGs contain time information. The entity may contain different attributes at different moments. In order to better characterize the entity in temporal KGs, we associate each entity e with a time label $t_q \in \mathcal{T}$, so the entity e can be depicted as e^{t_q} and its embedding can be denoted as $\mathbf{e^{t_q}}$. The set of quadruples directly associated with $e_s^{t_q}$, which means the 1-hop paths associated with $e_s^{t_q}$, can be expressed as: $P_{(e_s,t_q)} = \left\{(e_s, r_q, e_i, t_i) \,|\, (e_s, r_q, e_i, t_i) \in G_{t_q}, e_s, e_i \in \mathcal{E}, r_q \in \mathcal{R}, t_i < t_q \in \mathcal{T}\right\}$ [1]. The set of entities directly associated with $e_s^{t_q}$ in the path $P_{(e_s,t_q)}$, i.e., the 1-hop neighbors of ${e_s}^{t_q}$, can be denoted as $N_{(e_s,t_q)} = \left\{e_i^{t_h} \,|\, (e_s, r_k, e_i, t_h) \in P_{(e_s,t_q)}, e_s, e_i \in \mathcal{E}, r_k \in \mathcal{R}, t_h < t_q \in \mathcal{T}\right\}$. Given the query $(e_s, r_q, ?, t_q)$, the forecasting task can be depicted as requesting the entity e_o based on path searching. For example, we search the path with e_s as the starting point: $(e_s, r_{p1}, e_1, t_1), (e_1, r_{p2}, e_2, t_2), (e_2, r_{p3}, e_3, t_3), \ldots, (e_{i-1}, r_{pi}, e_i, t_i)$. So answers to the query may be $e_1, e_2, e_3, \ldots, e_i$, and the corresponding inference hop is $1, 2, 3, \ldots, i$ respectively.

Fact matching based on TransE. This paper is the first to design interpretable evaluation indicators from the perspective of actual semantics. In order to better illustrate the design route, we choose the basic embedding model–TransE as the basis of IMR. In TransE, relations are represented as translations in the embedding space: if the triple (e_s, r, e_o) holds in static KGs, TransE (Bordes et al., 2013) assumes that $|\mathbf{e_s} + \mathbf{r} - \mathbf{e_o}| = 0$. For each quadruples (e_s, r_q, e_o, t_q) in temporal KGs, the relation r_q can also be taken as the translations from the subject e_s to the object e_o, i.e. $\mathbf{e_s^{t_q}} + \mathbf{r_q} = \mathbf{e_o^{t_q}}$. We suppose that when the distance d of quadruples is smaller, the quadruple will be better matched. The distance of the quadruple (e_s, r_q, e_o, t_q) can be expressed as:

$$d = \left|\mathbf{e_s^{t_q}} + \mathbf{r_q} - \mathbf{e_o^{t_q}}\right| \tag{1}$$

The relations indicate the translations between entities, whose specific design determines the complexity of indicators designed by IMR. The design route of IMR starts from the perspective of reasoning from actual semantics, which is not limited to specific basic models. All the specific formulas of IMR in this paper are based on TransE, which will not be explained below.

4 OUR MODEL

We introduce Interpretable Multi-hop Reasoning (IMR) in this section. We first provide an overview of IMR in Section 4.1. IMR comprises three modules: path searching module, query updating module, and path scoring module. The path searching module searches related paths hop by hop from the subjects of questions, involving path sampling and entity clipping, which will be introduced in Section 4.2. The query update module calculates the remaining questions hop-by-hop for each path, involving the update of the subject and relations, which will be introduced in Section 4.3. The scoring module designs three indicators: question matching degree, answer completing level, and path confidence. This module combines three indicators to score each path, which will be introduced in Section 4.4. We will introduce training strategies and the regularizations on state continuity in Section 4.5.

[1] We reverse all the quadruple. Add $\left(e_j, r_q^{-1}, e_i, t_k\right)$ for each (e_i, r_q, e_j, t_k). In this way, $P_{(e_s,t_q)}$ can represent all associated quadruples.

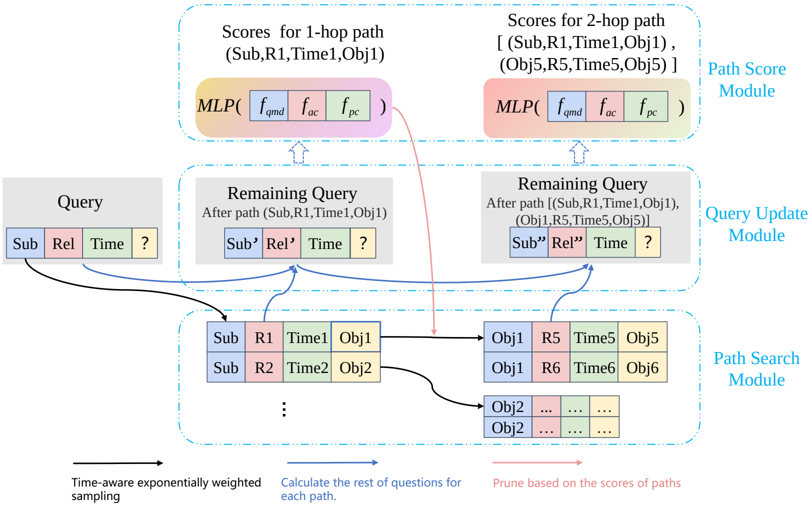


Figure 1: Model Architecture. We take the 2-hop path searching as an example. The black arrow means time-aware exponentially weighted sampling. The blue arrows denote calculating the rest of questions for each path. The red arrow represents pruning based on the scores of paths. We take **(Sub, Rel, ?, Time)** as the original question ,which can be denoted as $(e_s, r_q, ?, t_q)$. The searched two paths are **[(Sub,R1,Obj1,Time1)]** and **[(Sub,R1,Obj1,Time1),(Obj1,R5,Obj5,Time5)]**, which can be denoted as $[(e_s, r_{p1}, e_1, t_1)]$ and $[(e_s, r_{p1}, e_1, t_1), (e_1, r_{p2}, e_2, t_2)]$ respectively. **(Sub', Rel', ?, Time)** and **(Sub", Rel", ?, Time)** denote the remaining question after 1-hop and 2-hop path, wihch can be taken as $(e_{s\_1}, r_{q\_1}, ?, t_q)$, $(e_{s\_2}, r_{q\_2}, ?, t_q)$ respectively.

4.1 Model Overview

We observe that predicting unknown facts based on paths is similar to question answering, i.e., the question can be answered directly via finding triples with the equivalence relation or gradually by utilizing a multi-hop equivalence path. Inspired by this observation, we view the task of link prediction as to question answering. IMR primarily consists of searching for paths hop by hop, updating the remaining questions for each path, and filtering the best answers based on three indicators: question matching degree, answer completing level, and path confidence.

We show a toy example in Figure 1. Given a question $(e_s, r_q, ?, t_q)$ and the previous facts G_{t_q}, the task of forecasting is predicting the missing object e_o. The steps of IMR are as follows. **Step 1**: Starting from the subject e_s, find out the associated quadruples $P_{(e_s,t_q)}$, namely 1-hop paths. We temporally bias the neighborhood sampling using an exponential distribution for the neighbors. The distribution negatively correlates with the time difference between node e_s and its neighbor $N_{(e_s,t_q)}$. Then, calculate the remaining questions (the remaining subject e_{s_1} and the remaining relation r_{q_1}) for each sampled path. Finally, IMR scores 1-hop paths based on three indicators, which will be discussed in Section 4.4. **Step 2**: To prevent the path searching from exploding, the model samples the tails of 1-hop paths for the 2-hop path searching. As shown by the pink arrow in Figure 1, tails of 1-hop paths are clipped according to scores of 1-hop paths. For the 2-hop paths searched from the clipped tails, IMR samples the paths negatively correlated with time distances. Then, IMR calculates the remaining questions for each 2-hop path (the remaining subject e_{s_2} and the remaining relation r_{q_2}) and scores the 2-hop paths based on three indicators. **Step 3**: Rank the scores of all paths to obtain the preference answer.

4.2 Path Searching Module

Path sampling. For the path searching from the starting subject e_s, the number of triples in G_{t_q} may be very large. To prevent the path searching from exploding, we sample a subset of the paths. In fact, the attributes of entities in temporal KGs may change over time. Consider the observation that when t_1 is closer to t_q, the attributes of $e_s^{t_1}$ should be more similar to those of $e_s^{t_q}$. Therefore, we are more prone to sample nodes whose time is closer to t_q. In this paper, we employ time-aware exponentially weighted sampling in xERTR (Han et al., 2021). xERTR (Han et al., 2021) temporally biases the neighborhood sampling using an exponential distribution of time distance.

Entity pruning. The search for next-hop paths is based on the tails of previous hop paths. Besides, the number of paths is increased by $O(k^n)$. In order to avoid the explosion of next-hop path searching, only a few tails are selected for further path searching. We directly sort the scores of the previous hop and select the top-K entities for the next-hop search.

4.3 Query Updating Module

Given a question $(e_s, r_q, ?, t_q)$, there may be few direct equivalent relations with r_q in the temporal KGs for the task of link prediction. More questions need to go through multi-hop paths to reason about the result. In question answering, a complex question can be decomposed into multiple sub-questions, and one sub-question is answered at each step. So the reasoning based on the multi-hop path is equivalent to answering complex questions step by step. For the part that has already been answered, we need to remove the part resolved so that we can focus on the remaining questions. Therefore, we need to update the question according to the last hop of the path, focusing on finding the unsolved parts. The embedding of entities is first introduced in this subsection, followed by the query updating module.

Entity representation. The attributes contained in the entity may change over time. This paper divides the entity embeddings of each timestamp into static representation and dynamic representation.

$$\mathbf{e} = act\,(MLP([\ \mathbf{e_{static}}\ ||\ \mathbf{e_{dynamic}}\ ])) \tag{2}$$

Here, the vector $\mathbf{e_{static}} \in \mathbf{R^d}$ denotes the static embedding, which captures time-invariant features and global dependencies over the temporal KGs. The vector $\mathbf{e_{dynamic}} \in \mathbf{R^{2d}}$ represents the dynamic embedding for each entity that changes over time. $MLP(\cdot)$ denotes the multilayer perceptron (MLP). $act(\cdot)$ means the activation function and we take LeakyReLU as the activation function in this paper. We will provide more details about the each component A.1.

Question updating. Each path has a different set of relations. After each hop, the question needs to discard the processed semantic, i.e., to obtain the remaining subject and relation of the question. As shown in Figure 1, the subject and relation of the question after the i-th hop path are updated based on Eq.1 as follows.

$$\mathbf{e_{q\_i}} = \mathbf{e_{q\_i-1}} + \mathbf{r_{pi}} \tag{3}$$

$$\mathbf{r_{q\_i}} = \mathbf{r_{q\_i-1}} - \mathbf{r_{pi}} \tag{4}$$

where the embedding $\mathbf{e_{q\_i}}$ and $\mathbf{r_{q\_i}}$ represents the remaining subject and relation of the question after the i-hop path respectively. Besides, $\mathbf{r_{pi}}$ denotes the relation of i-th hop path and i is the number of hops for each path. $\mathbf{e_{q\_0}} = \mathbf{e_q}$, $\mathbf{r_{q\_0}} = \mathbf{r_q}$.

4.4 Path Scoring Module

We evaluate the path searching from three perspectives. First, the searched tails should match the original question, which means that the correct tails searched by paths and the question should satisfy TransE. Secondly, the ideal path should be the search of equivalent semantics for relations, not just the search for the correct tails. It is necessary to ensure the correctness of semantic equivalence, i.e., the path is semantically equivalent to the relation of the question. Finally, considering the particularity of the temporal KGs, the attributes of the same entity may change over time. The current sampling strategy for path searching is to sample adjacent timestamp triples of the same entity. When the attribute value of the entity changes significantly over time, it is inappropriate to perform this sampling strategy for the next hop. We need to ensure that the same entity with different timestamps has similar properties in the same path. In this way, three indicators have been developed by IMR to measure the rationality of the reasoning path respectively: question matching degree, answer completing level, and path confidence.

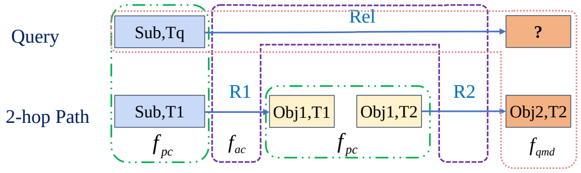


Figure 2: A brief illustration of the path scoring module. For the query **(Sub, Rel, Tq,?)**, we search the 2-hop path **(Sub, R1, T1, Obj1),(Obj1, R2, T2, Obj2)**. The pink box indicates that the original question and the tail of the path are formed a quadruple to measure the rationality of the inference tail as the answer, that is, Question Matching Degree f_{qmd}. The purple box represents the comparison between the question's relation and the path relations to measure the semantics equivalence between the question and the path, that is, Answer Completing Level f_{ac}. These green boxes compare the attributes of the same entities with different timestamps to measure the reliability of the search path, that is, Path Confidence f_{pc}.

Question Matching Degree. For the tails found by the path searching, we need to measure the matching degree between the tails and the question, that is, the Question Matching Degree. In fact, the scoring function applied by some traditional reinforcement learning methods is a kind of question matching degree. As shown in the yellow box in Figure 2, for the entity $e_i^{t_i}$ searched by the paths, we combine the entity $e_i^{t_i}$ and the question $(e_s, r_q, ?, t_q)$ into a new quadruple $\left(e_s, r_q, e_i^{t_i}, t_q\right)$. Question matching degree f_{qmd} calculates the distance of the constructed quadruple based on TransE (Bordes et al., 2013). The better the entity matches the query, the smaller the distance of quadruples will be. The calculation of f_{qmd} for ith-hop path is as follows.

$$f_{qmd}^{i} = \left\|\mathbf{e_s^{t_q}} + \mathbf{r_q} - \mathbf{e_i^{t_i}}\right\|_p \tag{5}$$

where the p-norm of a complex vector V is defined as $\|V\|_p = \sqrt[p]{|V_i|^p}$. We use L1-norm for all indicators in the following.

Answer completing level. Among the paths to the right tails, some paths are not related to the semantics of the question. Although these paths can infer the tail, these paths are not valid for being unrelated to the question in semantic. Therefore, IMR designs an index to measure the semantic relevance between the path and the question. Answer completing level f_{ac} indicates whether the combination of path relations can reflect the relation of the question in semantic. IMR takes the remaining relations of the question as the answer completing level, which is calculated based on the distance between the relations of paths $r_{p1}, r_{p2}, ...$ and the relation r_q. In general, the fewer the relation of a query remains, the more complete answer the combination of path relations will give.The calculation of f_{ac} for ith-hop path is as follows.

$$\begin{aligned} f_{ac}^{i} &= \|\mathbf{r_q} - \mathbf{r_{p1}} - \mathbf{r_{p2}} - \mathbf{r_{p3}} - ... - \mathbf{r_{pi}}\|_p \\ &= \left\|\mathbf{r_{q\_1}} - \mathbf{r_{p2}} - \mathbf{r_{p3}} - ... - \mathbf{r_{pi}}\right\|_p \\ &= \left\|\mathbf{r_{q\_2}} - \mathbf{r_{p3}} - ... - \mathbf{r_{pi}}\right\|_p \\ &= \left\|\mathbf{r_{q\_i}}\right\|_p \end{aligned} \tag{6}$$

Path confidence. Path searching is the process of searching for the next-hop paths based on the tail of the previous hop. When searching for a path, the current sampling strategy is to sample adjacent timestamp triples of the same entity. There are deviations between the same entities with different timestamps in temporal KGs. The premise of this sampling strategy is that only when entities have similar attributes under different timestamps, the path searching is valid. When the entity's attributes change significantly over time, performing an effective next path search is inappropriate. The reasoning path is more reliable when the deviations between entities are smaller. IMR designs Path Confidence f_{pc}, i.e., the error between the subject of the updated question e_{q_i} and the tails of the path $e_i^{t_i}$. The calculation of f_{pc} for ith-hop path is as follows.

$$f_{pc}^{i} = \left\| \mathbf{e_{q\_i}} - \mathbf{e_i^{t_i}} \right\|_p \tag{7}$$

Where e_{q_i} represents the remaining subject of the question updated by paths of the length i, and $e_i^{t_i}$ represents the tail reasoned by the i-hop paths.

Combination of scores. IMR merges the scores of paths with multilayer perceptron (MLP) to obtain the final score f of each path .

$$f = MLP\left(\left[\; f_{pc} \;||\; f_{ac} \;||\; f_{qmd} \;\right] \right) \tag{8}$$

The temporal KG forecasting is to sort all entities with the same timestamp, that is, IMR needs to combine scores of entities with different timestamps. Entities with the same timestamp may be inferred from different paths. Considering only one path matches the query the most, IMR employs max aggregation for the score of paths inferring same entities with the same timestamp. In this case, the entity has only one unique score per timestamp. Besides, certain paths may infer the same entity with a different timestamp. In order to make better use of the path information of different timestamps, IMR performs average aggregation for the scores of entities with different timestamps. Finally, IMR obtains the score of each entity at the query timestamp.

4.5 LEARNING

We utilize the binary cross-entropy as the loss function, which is defined as:

$$L = -\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{\left|\varepsilon_q^{path}\right|}\sum_{e_i\in\varepsilon_q^{path}}\left(y_{e_i,q}\log\left(\frac{f_{e_i,q}}{\sum_{e_i\in\varepsilon_q^{path}} f_{e_i,q}}\right) + (1-y_{e_i,q})\log\left(1-\frac{f_{e_i,q}}{\sum_{e_i\in\varepsilon_q^{path}} f_{e_i,q}}\right)\right) \tag{9}$$

where ε_q^{path} represents the set of entities reasoned by selected paths. $y_{e_i,q}$ represents the binary label that indicate whether e_i is the answer for q and Q represents the training set. $f_{e_i,q}$ denotes the score obtained by Eq.8 for each path. We jointly learn the embeddings and other model parameters by end-to-end training.

Regularization. For the same entity with different timestamps, the closer the entity's time distance is, the closer its dynamic embedding is. IMR implements the regularization on continuity for the dynamic vectors of entities.

$$reg = \left\| \mathbf{e_i^{t_j}} - \mathbf{e_i^{t_{j-1}}} \right\|_p + \left\| \mathbf{e_i^{t_j}} - \mathbf{e_i^{t_{j+1}}} \right\|_p \tag{10}$$

where, $\mathbf{e_i^{t_j}}$ denotes the dynamic embedding of the i-th entity at the j-th timestamp. $\mathbf{e_i^{t_{j-1}}}, \mathbf{e_i^{t_{j+1}}}$ denotes the dynamic embedding of the previous and later timestamp against $e_i^{t_j}$ respectively. $\|\cdot\|_p$ denotes the p norm of the vectors and we take p as 1 in this paper.

5 EXPERIMENTS

5.1 DATASETS AND BASELINES

In order to evaluate the proposed module, we consider two standard temporal KG datasets Integrated Crisis Early Warning System (ICEWS) (Boschee et al., 2015), WIKI (Leblay & Chekol, 2018b) and YAGO (Mahdisoltani et al., 2015). The ICEWS dataset contains information about political events with time annotations. We select three subsets of the ICEWS dataset, i.e., ICEWS14 and ICEWS18, containing event facts in 2014 and 2018, respectively. WIKI and YAGO is a temporal KG that fuses information from Wikipedia with WordNet (Miller, 1995). Following the experimental settings of HyTE (Dasgupta et al., 2018), we deal with year-level granularity by dropping the month and date information. We compare IMR and baseline methods by performing the temporal KGs forecasting task on the ICEWS14, ICEWS18, WIKI, and YAGO. Details of these datasets are listed in Table 1. We adopt the same dataset split strategy as in (Jin et al., 2020).

Dataset	entity	relation	timestamp	training	validation	test
ICEWS14	7128	230	365	63685	13823	13222
ICEWS18	23033	256	304	373018	45995	49545
WIKI	12554	24	232	539286	67538	63110
YAGO	10623	10	189	161540	19523	20026

Table 1: The number of entities, relations, timestamps and observed triples for four benchmark datasets.

	ICEWS14				ICEWS18				WIKI				YAGO			
	MRR	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**
TTransE	13.43	3.11	17.32	34.55	8.31	1.92	8.56	21.89	29.27	21.67	34.43	42.39	31.19	18.12	40.91	51.21
TA-DistMult	26.47	17.09	30.22	45.41	16.75	8.61	18.41	33.59	44.53	39.92	48.73	51.71	54.92	48.15	59.61	66.71
DE-SimplE	32.67	24.43	35.69	49.11	19.30	11.53	21.86	34.80	45.43	42.6	47.71	49.55	54.91	51.64	57.30	60.17
TNTComplEx	32.12	23.35	36.03	49.13	27.54	19.52	30.80	42.86	45.03	40.04	49.31	52.03	57.98	52.92	61.33	66.69
CyGNet	32.73	23.69	36.31	50.67	24.93	15.90	28.28	42.61	33.89	29.06	36.10	41.86	52.07	45.36	56.12	63.77
RE-NET	38.28	28.68	41.34	54.52	28.81	19.05	32.44	47.51	49.66	46.88	51.19	53.48	58.02	53.06	61.08	66.29
xERTE	40.79	32.70	45.67	57.30	29.31	21.03	33.51	46.48	71.14	68.05	76.11	79.01	84.19	80.09	88.02	89.78
TANGO-Tucker	–	–	–	–	28.68	19.35	32.17	47.04	50.43	48.52	51.47	53.58	57.83	53.05	60.78	65.85
TANGO-DistMult	–	–	–	–	26.75	17.92	30.08	44.09	51.15	49.66	52.16	53.35	62.70	59.18	60.31	67.90
TITer	41.73	32.74	46.46	58.44	29.98	22.05	33.46	44.83	75.50	72.96	77.49	79.02	87.47	84.89	89.96	90.27
IMR	**44.76**	**35.64**	**49.49**	**62.30**	**32.45**	**22.97**	**36.05**	**49.36**	**80.41**	**76.04**	**84.91**	**85.95**	**90.24**	**87.91**	**92.65**	**92.77**

Table 2: Results comparison on four datasets. Compared metrics are time-aware filtered MRR(%) and Hits@1/3/10 (%), which are multiplied by 100. The best results among all models are in bold.

We compare the performance of IMR against the temporal KG reasoning models, including TTransE (Leblay & Chekol, 2018a), TA-DistMult/TA-TransE (García-Durán et al., 2018) , DE-SimplE (Goel et al., 2020), TNTComplEx (Lacroix et al., 2020), CyGNet (Zhu et al., 2020), RE-Net (Jin et al., 2020), TANGO (Ding et al., 2021), TITer (Sun et al., 2021) and xERTR (Han et al., 2021).

In the experiments, the widely used Mean Reciprocal Rank (MRR) and Hits@1,3,10 are employed as the metrics. The filtered setting for static KGs is not suitable for the reasoning task under the exploration setting, as mentioned in xERTR (Han et al., 2021). The time-aware filtering scheme only filters out triples that are genuine at the query time.

5.2 Experimental results and ablation study

Result Comparison. Table 2 shows the comparison between IMR and other baseline models on the ICEWS, WIKI, and YAGO datasets [2]. Overall, IMR outperforms all baseline models in all metrics while being more interpretable, which convincingly verifies its effectiveness. Compared to the strongest baseline TITer (Sun et al., 2021), IMR obtains a relative improvement of 3.3% and 2.5% in MRR and Hits@1, which are averaged on ICEWS, WIKI, and YAGO. To assess the importance of each component, we further conduct several ablation studies.

Comparison of multi-hop paths. Figure 3 shows the performance of IMR on ICEWS, WIKI, and YAGO as the maximum hop of paths increases. The performance basically continues to rise with the increase of the maximum hop of paths. But when the maximum hop of paths increases, the performance of IMR on ICEWS18 hardly improves. The further analysis of ICEWS18 in (Li et al., 2021) explains that there are no strong dependencies between the relations of the question and the multi-hop paths. Thus, in this situation, longer paths provide little gain for inference [3]. Besides, as the max-hop of paths increases, the number of inference paths increases exponentially, most of which are invalid paths and will suppress the performance of IMR. In order to ensure that the performance of the model does not decrease, we strictly control the sampling number of next-hop paths to limit the number of multi-step paths and suppress the impact of noise samples. Here, we set the number of next-hop sampling to 5 in the experiments of this paper. In summary, experiments show that unified indicators designed by IMR measure the paths of different hops in the same space, allowing better reasoning based on paths with different hops, which is consistent with the claim in Section 4.4.

[2] Codes and datasets will be available at https://github.com/lfxx123/TKBC

[3] We leave it for future work to construct a more complex dataset for verifying the effectiveness of multi-hop paths.

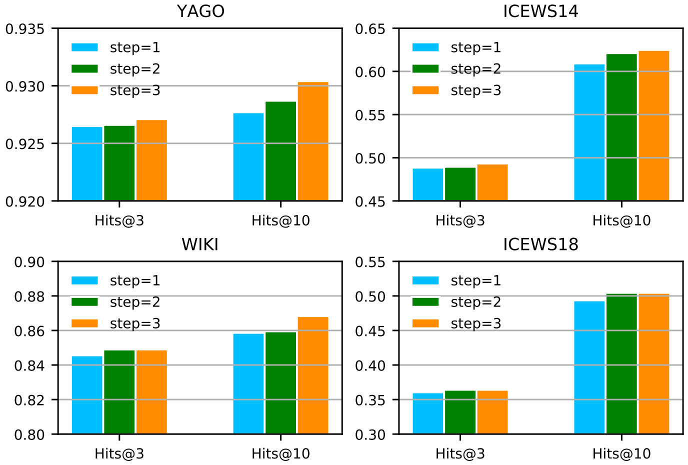


Figure 3: Comparison of the performance of paths with different maximun hop on four datasets. We average the output of four experiments with different random seeds and fixed hyperparameters.

Dataset	**YAGO**				**ICEWS14**				**WIKI**				**ICEWS18**			
Indicator	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**
f_{qmd}	87.32	92.53	92.76	89.87	22.61	39.20	55.32	33.48	70.75	83.39	85.87	77.12	12.76	26.47	43.75	22.66
f_{ac}	87.79	92.67	92.78	90.18	31.67	46.02	59.21	41.05	-	-	-	-	20.41	33.50	47.48	29.45
f_{pc}	87.74	92.67	92.77	90.15	25.65	43.03	58.25	36.63	70.72	83.35	85.31	77.00	14.92	29.00	45.58	24.82
f_{ac}, f_{qmd}	87.95	**92.67**	**92.77**	90.26	34.91	49.26	61.12	43.82	**76.12**	84.90	85.94	**80.46**	23.05	**36.20**	49.47	31.84
f_{pc}, f_{qmd}	87.74	**92.67**	92.75	90.15	25.64	43.16	58.30	36.63	73.85	84.12	85.65	78.99	13.10	26.38	43.27	22.75
f_{ac}, f_{pc}	87.91	92.65	**92.77**	90.24	34.81	49.02	**61.15**	43.74	76.04	84.91	85.95	80.41	23.04	36.10	49.46	31.83
f_{ac}, f_{pc}, f_{qmd}	**88.31**	92.66	**92.77**	**90.48**	**34.96**	**49.27**	61.09	**43.89**	76.09	**84.92**	**85.96**	80.44	**23.15**	36.12	**49.52**	**31.89**
Distance to the best	0	0.01	0	0	0	0	0.06	0	0.03	0	0	0.02	0	0.08	0	0

Table 3: The comparison of three indicators in different combinations. We average the output of ten experiments with different random seeds and fixed hyperparameters. All metrics are multiplied by 100.

Combination of indicators. The three indicators measure different aspects of the path: the matching degree between answers and the question, the completeness of relational equivalence, and the reliability of reasoning paths. We verify the performance of each metric through ablation experiments. As shown in Table 3, the first block displays the performance with only one indicator, the second block presents the performance with a combination of two parameters, and the last is a combination of three indicators. The bottom line shows the error between the combination of the three parameters and the best result. Since the distribution varies across two datasets, there are certain differences in performance when employing a single indicator to rank paths. The model's performance has been significantly improved after incorporating the three indicators in pairs, but few differences still remain. IMR can obtain the best inference performance in most datasets by combining three indicators. In summary, the experiment illustrates that the combination of three indicators designed by IMR can effectively measure the reasoning paths.

6 Conclusion

We proposed an Interpretable Multi-hop Reasoning approach (IMR) for forecasting future links on temporal KGs. IMR transforms reasoning based on path searching into step-by-step question answering. Moreover, IMR designs three indicators to measure the answer and reasoning paths. Extensive experiments on four benchmark datasets demonstrate the effectiveness of our method. In the future, we plan to enhance the prediction by integrating different paths reaching the same tail, which will be more effective and interpretable.

References

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Tucker: Tensor factorization for knowledge graph completion. *ArXiv*, abs/1901.09590, 2019.

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

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. *Icews coded event data.*, volume 12. Harvard Dataverse, 2015.

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha P. Talukdar. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *EMNLP*, pp. 2001–2011. Association for Computational Linguistics, 2018.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. In *AAAI*, 2018.

Zifeng Ding, Zhen Han, Yunpu Ma, and Volker Tresp. Temporal knowledge graph forecasting with neural ode. *ArXiv*, abs/2101.05151, 2021.

Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In *EMNLP*, pp. 4816–4821. Association for Computational Linguistics, 2018.

Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In *EMNLP*, 2018.

Rishab Goel, Seyed Mehran Kazemi, Marcus A. Brubaker, and P. Poupart. Diachronic embedding for temporal knowledge graph completion. *ArXiv*, abs/1907.03143, 2020.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *ICLR*. OpenReview.net, 2021.

Guoliang Ji, Shizhu He, L. Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, 2015.

Woojeong Jin, Changlin Zhang, Pedro A. Szekely, and Xiang Ren. Recurrent event network for reasoning over temporal knowledge graphs. *CoRR*, abs/1904.05530, 2019.

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

Jaehun Jung, Jinhong Jung, and U. Kang. Learning to walk across time for interpretable temporal knowledge graph completion. In *KDD*, pp. 786–795. ACM, 2021.

Timothée Lacroix, G. Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. *ArXiv*, abs/2004.04926, 2020.

J. Leblay and M. Chekol. Deriving validity time in knowledge graph. *Companion Proceedings of the The Web Conference 2018*, 2018a.

Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. *Companion Proceedings of the The Web Conference 2018*, 2018b.

Ruiping Li and Xiang Cheng. DIVINE: A generative adversarial imitation learning framework for knowledge graph reasoning. In *EMNLP/IJCNLP (1)*, pp. 2642–2651. Association for Computational Linguistics, 2019.

Zixuan Li, Xiaolong Jin, Saiping Guan, Wei Li, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng. Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs. In *ACL/IJCNLP (1)*, pp. 4732–4743. Association for Computational Linguistics, 2021.

Yankai Lin, Zhiyuan Liu, M. Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2015.

F. Mahdisoltani, J. Biega, and Fabian M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*, 2015.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

Dai Quoc Nguyen, T. Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. *ArXiv*, abs/1712.02121, 2018.

Dai Quoc Nguyen, Thanh Vu, T. Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A capsule network-based embedding model for knowledge graph completion and search personalization. *ArXiv*, abs/1808.04122, 2019.

Maximilian Nickel, Volker Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217, 2008.

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

Zhiqing Sun, Zhihong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *ArXiv*, abs/1902.10197, 2019.

Théo Trouillon, Johannes Welbl, S. Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, 2016.

Shikhar Vashishth, Soumya Sanyal, V. Nitin, and P. Talukdar. Composition-based multi-relational graph convolutional networks. *ArXiv*, abs/1911.03082, 2020.

Heng Wang, Shuangyin Li, Rong Pan, and Mingzhi Mao. Incorporating graph attention mechanism into knowledge graph reasoning based on deep reinforcement learning. In *EMNLP/IJCNLP (1)*, pp. 2623–2631. Association for Computational Linguistics, 2019.

Zhen Wang, J. Zhang, Jianlin Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

Jiapeng Wu, Meng Cao, Jackie Chi Kit Cheung, and William L. Hamilton. Temp: Temporal message passing for temporal knowledge graph completion. In *EMNLP (1)*, pp. 5730–5746. Association for Computational Linguistics, 2020.

Chenjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. Temporal knowledge graph completion based on time series gaussian embedding. In *ISWC (1)*, volume 12506 of *Lecture Notes in Computer Science*, pp. 654–671. Springer, 2020a.

Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *ArXiv*, abs/2002.07962, 2020b.

B. Yang, Wen tau Yih, X. He, Jianfeng Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *CoRR*, abs/1412.6575, 2015.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. An interpretable reasoning network for multi-relation question answering. In *COLING*, pp. 2010–2022. Association for Computational Linguistics, 2018.

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

A Appendix

A.1 The Embedding for Entities

This paper divides the entity embeddings of each timestamp into static representation and dynamic representation. The static embedding captures time-invariant attributes of the entity. We denote the static embedding of the entity e_i with $\mathbf{e_{i-static}} \in \mathbf{R^d}$, which is a d-dimensional vector independent of time. (Xu et al., 2020b) proposes a generic time encoding to generate the time-variant part of entity representations, which can be denoted as $\Phi\left(t\right)$. Employing this time-encoding, quadruples with the same subject, predicate, and object can have different attention scores. Specifically, quadruples that occurred recently tend to have higher attention scores. This makes the embedding more interpretable and effective.

$$\Phi\left(t\right)=\sqrt{\frac{1}{d}}\left[\cos\left(\omega_1 t+\phi_1\right),\ldots,\cos\left(\omega_d t+\phi_d\right)\right], \Phi\left(t\right)\in\mathbf{R^d} \tag{11}$$

We observe that the semantic attributes of entities determine the reasoning, and the attribute deviation caused by the time deviation is the only assumption obtained after statistics. In order to avoid being only affected by time factors, we propose a new time-specific entity representation $\mathbf{e_i^t}\in\mathbf{R^d}$, that is, each entity has a different representation at different timestamps. If each entity applies different representations at every moment, it will consume enormous resources. As most of the entities are only observed at limited timestamps, this paper characterizes the entities whose timestamps only appear in the training dataset. IMR utilizes the embedding of the separate entity when it last occurred in the training dataset to represent the embedding at the timestamps missing from the training dataset. Besides, to avoid over-fitting caused by too many parameters, we apply regularizations on time continuity. This regularization believes that the temporally continuous entities should have closer embeddings, as described in 4.5. Finally we combine $\Phi\left(t\right)$ and $\mathbf{e_i^t}$ to construct $\mathbf{e^t_{i-dynamic}}\in\mathbf{R^{2d}}$.

$$\mathbf{e^t_{i-dynamic}}=\left[\ \Phi\left(t\right)\ ||\ \mathbf{e_i^t}\ \right] \tag{12}$$

where $||$ denotes the operation of concatenation. In summary, the embedding for each entity e_i can be represented as :

$$\mathbf{e_i}=act\left(MLP(\left[\ \mathbf{e_{i-static}}\ ||\ \mathbf{e^t_{i-dynamic}}\ \right])\right) \tag{13}$$

where $MLP(\cdot)$ denotes the multilayer perceptron (MLP). $act(\cdot)$ means the activation function and we take LeakyReLU as the activation function in this paper.

A.2 Case Studies and Interpretability

For the query (John Kerry,Make a visit,?,2014-11-11), we extract part of the paths for the case study in Table 4. The lower the scores or indicators in Table 4, the better the performance of the path. We compare the paths based on the total score, analyze various aspects of the paths based on detailed indicators, and verify the interpretation of the model with actual semantics.

The first block of Table 4 selects reasoning paths with the same objects to analyze the answer completing level. First, we compare path 1-1 and path 1-2. The score of path 1-1 is lower than that of path 1-2. As we analyze the three indicators further, we find that the answer completing level of path 1-1 is smaller than that of path 1-2. The comparison of answer completing level indicates that the relation of path 1-1 should be closer to the relations of the query. Practically, path 1-1 has the same relation as the query, which is closer to the relation of query than path1-2. Actual semantics verifies the interpretation of the model. Comparing path 1-4 and path 1-5, we find that the total score of path 1-4 is lower than that of path 1-5, and the answer completing level of path 1-5 is higher than that of path 1-4. IMR shows that the combination of reasoning relations of path 1-4 is better than that of path 1-5. In fact, these two paths for inference do not seem to be particularly appropriate to the query. Nevertheless, the combination of relations **[Meet at a 'third' location + Make a visit]** is actually closer to the relation of the query **[Make a visit]** than the combination of relations **[Consult + Consult]**. To summarize, the first set of experiments shows that the answer completing level can effectively indicate how well the combination of path relations equals to the relation of the query, verifying the statement in section 4.4.

The second block of Table 4 selects the paths of the same reasoning relations to verify the path confidence and the question matching degree. Comparing path 2-1, 2-2, 2-3 and 2-4, we observe that the scores of the paths are increasing. Additionally, the path confidence of these three paths is also increasing. In fact, the time distance between the paths and the query is gradually increasing,

Query:	John Kerry	Make a visit	Oman	2014-11-11								
Path-ID	**Reasoning Path**								**Score**			
	e_s	r_2	e_2	t_2	e_2	r_3	e_3	t_3	f_{ac}	f_{qmd}	f_{pc}	**Combined Score**
path 1-1	John Kerry	Make a visit	Oman	2014-11-09	-	-	-	-	0	74	74	137
path 1-2	John Kerry	Express intent to meet or negotiate	Oman	2014-11-09	-	-	-	-	26	74	69	169
path 1-3	John Kerry	(Reversed) Host a visit	Oman	2014-11-09	-	-	-	-	27	74	76	178
path 1-4	John Kerry	Meet at a ’third’ location	Catherine Ashton	2014-11-10	Catherine Ashton	Make a visit	Oman	2014-11-09	38	74	90	206
path 1-5	John Kerry	Consult	Mohammad Javad Zarif	2014-11-10	Mohammad Javad Zarif	Consult	Oman	2014-11-09	73	74	107	254
path 2-1	John Kerry	Express intent to meet or negotiate	Oman	2014-11-10	-	-	-	-	26	47	41	119
path 2-2	John Kerry	Express intent to meet or negotiate	Oman	2014-11-09	-	-	-	-	26	74	69	170
path 2-3	John Kerry	Express intent to meet or negotiate	Oman	2014-11-05	-	-	-	-	26	89	83	196
path 2-4	John Kerry	Express intent to meet or negotiate	Oman	2014-11-02	-	-	-	-	26	90	82	197
path 2-5	John Kerry	Reversed Meet at a ’third’ location	Catherine Ashton	2014-11-10	Catherine Ashton	Express intent to meet or negotiate	Oman	2014-11-03	49	91	101	246
path 2-6	John Kerry	Reversed Meet at a ’third’ location	Catherine Ashton	2014-11-10	Catherine Ashton	Express intent to meet or negotiate	Oman	2014-11-05	49	89	100	242
path 2-7	John Kerry	Make a visit	China	2014-11-05	-	-	-	-	0	88	88	162
path 2-8	John Kerry	Make a visit	North Atlantic Treaty Organization	2014-06-25	-	-	-	-	0	87	87	160
path 2-9	John Kerry	Make a visit	Canada	2014-10-27	-	-	-	-	0	85	85	157
path 3-1	John Kerry	Reversed Meet at a ’third’ location	Catherine Ashton	2014-11-10	-	-	-	-	53	46	40	155
path 3-2	John Kerry	Express intent to meet or negotiate	Oman	2014-11-09	-	-	-	-	26	74	69	169
path 3-3	John Kerry	Make a visit	Afghanistan	2014-07-21	-	-	-	-	0	88	88	162
path 3-4	John Kerry	Make a visit	Afghanistan	2014-07-21	Afghanistan	Reversed Make statement	Barack Obama	2014-07-18	34	94	104	241
path 3-5	John Kerry	Make a visit	Angola	2014-08-05	Angola	(Reversed) Make statement	Anthony Foxx	2014-08-04	35	93	105	241
path 3-6	John Kerry	(Reversed) Make a visit	Catherine Ashton	2014-11-10	Catherine Ashton	Make a visit	Oman	2014-11-09	33	74	85	197

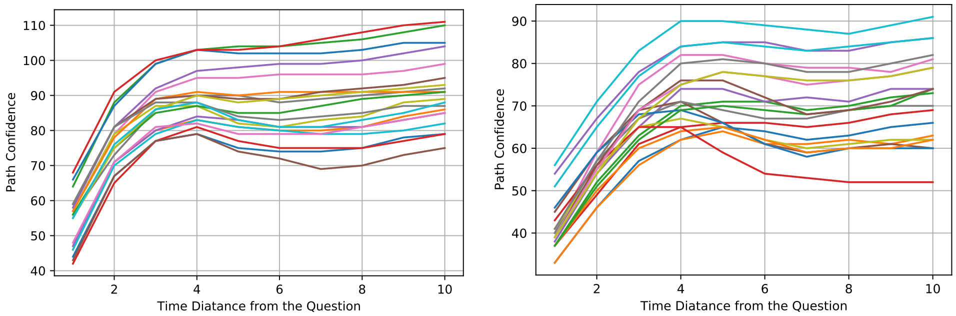
Table 4: Reasoning paths searched for the query **(John Kerry, Make a visit , ?, 2014-11-11)** and their scores respectively.

Query:	Citizen (Nigeria)	Use unconventional violence	Secretariat (Nigeria)	8016				
Path-ID	**Reasoning Path**				**Score**			
	e_s	r_2	e_2	t_2	f_{ac}	f_{qmd}	f_{pc}	**Combined Score**
path 4-1	Citizen (Nigeria)	Use unconventional violence	Militant (Nigeria)	7968	0	162	162	215
path 4-1	Citizen (Nigeria)	Use unconventional violence	Militant (Nigeria)	7728	0	185	185	245
path 5-2	Citizen (Nigeria)	Reversed Use unconventional violence	Terrorist (Boko Haram)	7824	72	204	199	359
path 5-3	Citizen (Nigeria)	Reversed Use unconventional violence	Terrorist (Boko Haram)	7776	72	174	168	319
path 5-4	Citizen (Nigeria)	Reversed Use unconventional violence	Militant (Boko Haram)	7872	72	206	202	363
path 5-5	Citizen (Nigeria)	Reversed Use unconventional violence	Militant (Boko Haram)	7776	72	173	166	317
path 5-6	Citizen (Nigeria)	Reversed Use unconventional violence	Militant (Boko Haram)	7752	72	175	167	319
Path 6-1	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7992	73	95	95	220
path 6-2	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7872	73	174	168	321
path 6-3	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7848	73	177	171	324
path 6-4	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7824	73	178	171	325
path 6-5	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7680	73	180	173	328
path 7-1	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7848	73	177	171	324
path 7-2	Citizen (Nigeria)	Make an appeal or request	Government (Nigeria)	7848	78	167	158	315
path 7-3	Citizen (Nigeria)	Reversed fight with small arms and light weapons	Boko Haram	7848	73	177	171	324
path 7-4	Citizen (Nigeria)	Reversed Make an appeal or request	Tony Momoh	7848	80	207	205	377
path 7-5	Citizen (Nigeria)	Reversed Express intent to meet or negotiate	South Africa	7848	85	169	165	330
path 7-6	Citizen (Nigeria)	Reversed Bring lawsuit against	Fessehaye Yohannes	7848	80	210	206	379

Table 5: Reasoning paths searched for the query **(Citizen (Nigeria), Use unconventional violence , ?, 8016)** and their scores respectively.

which means that the reliability of the paths gradually decreases. The reliability indicated by path confidence is consistent with the actual reliability. Similarly, we find that the path confidence of path 2-5 is higher than that of path 2-6, indicating that path 2-5 is less reliable. The actual situation is that the timestamp of path 2-5 (2014-11-03<2014-11-05) is farther from the timestamp of the query, which is consistent with the explanation. Comparing path 2-9 with path 2-7 and 2-8, respectively, the model further infers that the path confidence and question matching degree of path 2-9 are better than the other two paths. The actual situation is that the timestamp error with the query satisfies: path 2-7>path 2-9>path 2-8. This is because the question matching degree covers the path confidence. Considering the fact that the path confidence contains the error of the triple in the training dataset, the triple error covers the error caused by different timestamps, which makes path 2-9 more reliable than path 2-7. In general, the second set of experiments illustrates that the path confidence can effectively indicate the validity of each path.

In the third block of Table 4, we randomly select the paths, explain the paths based on these indicators, and verify them with the actual situation. We first sort three paths according to answer completing level : path 3-1 <path 3-2 <path 3-3. Therefore, the semantic similarity of relations between the three paths and the query should satisfy : path 3-3 >path 3-2 >path 3-1. The actual semantic similarity between the relations of paths and that of the query satisfies: **Make a visit** >**Express intent to meet or negotiate** >**Meet at a’third’ location**, which is consistent with the interpretation of IMR. Sort three paths by path confidence: path 3-1 <path 3-2 <path 3-3. The reliability of the three inference paths should satisfy: path 3-1 <path 3-2 <path 3-3. We observe that the time distance between the three paths and the query is gradually increasing, which verifies the explanation by path confidence. The analysis of path 3-4 to 3-6 is similar to the analysis of former paths. Case studies show that IMR can both provide reasoning paths as well as offer an valid basis for path comparison.



(a) YAGO

(b) WIKI

Figure 4: The relation between Path Confidence and time diatance.The questions and paths corresponding to each polyline is shown in Table 6 and Table 7.

Question	Paths
Matt_Mills , playsFor , Swindon_Town_F.C. , 186	Matt_Mills , playsFor , Swindon_Town_F.C.
Bolton_Wanderers_F.C. , Reversed playsFor , Ariza_Makukula , 183	Bolton_Wanderers_F.C. , Reversed playsFor , Ulish_Booker
Marin_Raykov , isAffiliatedTo , Bulgarian_Communist_Party , 186	Marin_Raykov , isAffiliatedTo , Bulgarian_Communist_Party
Stuart_Lewis , playsFor , England_national_under , 17_football_team , 187	Stuart_Lewis , playsFor , England_national_under , 17_football_team
Ipswich_Town_F.C. , Reversed playsFor , Giovani_dos_Santos , 186	Ipswich_Town_F.C. , Reversed playsFor , Mark_Burchill
Sean_Thornton , playsFor , Blackpool_F.C. , 185	Sean_Thornton , playsFor , Republic_of_Ireland_national_under , 17_football_team
Bangkok_Dock_Company , Reversed owns , Ministry_of_Finance_(Thailand) , 187	Bangkok_Dock_Company , Reversed owns , Ministry_of_Finance_(Thailand)
Alan_Howard , isMarriedTo , Sally_Beauman , 184	Alan_Howard , Reversed isMarriedTo , Sally_Beauman
Denise_Kandel , isMarriedTo , Eric_Kandel , 187	Denise_Kandel , Reversed isMarriedTo , Eric_Kandel
Nîmes_Olympique , Reversed playsFor , Djibril_Cissé , 183	Nîmes_Olympique , Reversed playsFor , Milko_Djurovski
WebMD , owns , MedicineNet , 184 ,	WebMD , owns , RxList
WLBJ , LP , Reversed owns , Knights_of_Columbus , 185	WLBJ , LP , Reversed owns , Knights_of_Columbus
Kevin_Corrigan , isMarriedTo , Elizabeth_Berridge_(actress) , 185	Kevin_Corrigan , Reversed isMarriedTo , Elizabeth_Berridge_(actress)
Scottish_Opera , owns , Theatre_Royal , _Glasgow , 185	Scottish_Opera , owns , Theatre_Royal , _Glasgow
Ronald_Stuart_Burt , worksAt , University_of_Chicago , 187	Ronald_Stuart_Burt , worksAt , University_of_Chicago
Lela_Rochon , Reversed isMarriedTo , Shabba_Doo , 183	Lela_Rochon , isMarriedTo , Shabba_Doo
England_national_under , 18_football_team , Reversed playsFor , Terry_Dunfield , 186	England_national_under , 18_football_team , Reversed playsFor , Terry_Dunfield
Cha_In , pyo , isMarriedTo , Shin_Ae , ra , 187	Cha_In , pyo , Reversed isMarriedTo , Shin_Ae , ra
Michael_Shanks , isMarriedTo , Lexa_Doig , 186	Michael_Shanks , Reversed isMarriedTo , Lexa_Doig
Amara_(singer) , isMarriedTo , Frans_Mohede , 186	Amara_(singer) , Reversed isMarriedTo , Frans_Mohede
Hélder_Barbosa , playsFor , Portugal_national_under , 16_football_team , 184	Hélder_Barbosa , playsFor , Portugal_national_under , 16_football_team
Walter_Richardson_(politician) , isAffiliatedTo , Australian_Labor_Party , 183	Walter_Richardson_(politician) , isAffiliatedTo , Australian_Labor_Party
University_of_Basel , Reversed worksAt , Alfred_Rittmann , 186	University_of_Basel , Reversed worksAt , Alfred_Rittmann
William_Webster_(Australian_politician) , isAffiliatedTo , Nationalist_Party_of_Australia , 185	William_Webster_(Australian_politician) , isAffiliatedTo , Australian_Labor_Party

Table 6: 20 questions and the corresponding paths selected in YAGO.

A.3 Correlation between Path confidence and Time Distance

The current sampling strategy believes that the greater the time distance of the same entity, the greater the deviation of its semantic properties. Therefore, IMR adopts a time negative sampling strategy to search for more effective paths. Path reliability is affected by semantic similarity, and negative time-aware correlation is a general situation or statistical result. IMR proposes path reliability to better measure the reliability of the searched path. Here we utilize Path Confidence of the same path with different timestamps to analyze the changes in semantic similarity over time. For the same problem, we find the same path with various timestamps. We randomly select 20 questions for path search, and each question selects the same path containing ten different timestamps to calculate Path Confidence. The randomly selected questions and the corresponding search paths are shown in Table 6 and Table 7. Figure 4 shows how the Path Confidence of each path changes with time and distance.

Figure 4 shows that as the time distance between the paths and questions increases, the score of Path Confidence gradually increases, indicating that its confidence is gradually decreasing. Experiments show that the semantic deviation of the same entity increases as the time distance increases, which verifies the rationality of time-aware negative exponentially sampling.

A.4 The Property of Offset in The Question Updating

In order to infer the correct tails, the Query Update Module should satisfy that the question still matches the same tail entity even after the updating. As shown in Eq.14, the Query Updating Module satisfies the property.

Question	Paths
Q1647 ,Reversed P190 ,Q2948 ,231	Q1647 ,Reversed P190 ,Q52981
Q780378 ,P39 ,Q13653224 ,225	Q780378 ,P166 ,Q987080
Q45178 ,P1435 ,Q20747146 ,229	Q45178 ,P1435 ,Q20747146
Q450442 ,Reversed P26 ,Q1054316 ,230	Q450442 ,Reversed P26 ,Q1054316
Q358052 ,P1435 ,Q624232 ,222	Q358052 ,P1435 ,Q624232
Q840499 ,P31 ,Q484170 ,228	Q840499 ,P31 ,Q484170
Q774064 ,P131 ,Q12626 ,227	Q774064 ,P131 ,Q12626
Q80985 ,P190 ,Q31487 ,228	Q80985 ,P190 ,Q2044
Q649 ,Reversed P551 ,Q440996 ,226	Q649 ,P31 ,Q515
Q17397691 ,Reversed P579 ,Q184196 ,228	Q17397691 ,Reversed P579 ,Q220373
Q1170125 ,P31 ,Q484170 ,226	Q1170125 ,P31 ,Q484170
Q61306 ,P190 ,Q1489 ,229	Q61306 ,P190 ,Q5836
Q911052 ,P1435 ,Q7934314 ,229	Q911052 ,P1435 ,Q7934314
Q693208 ,P31 ,Q667509 ,226	Q693208 ,P131 ,Q854043
Q622321 ,P463 ,Q253414 ,230	Q622321 ,P463 ,Q83276
Q3120 ,P131 ,Q16394 ,223	Q3120 ,P131 ,Q16394
Q838714 ,P131 ,Q1726339 ,224	Q838714 ,P131 ,Q1726339
Q21300 ,P131 ,Q12717 ,223	Q21300 ,P131 ,Q12717
Q4974 ,P131 ,Q16227 ,230	Q4974 ,P131 ,Q16227
Q337531 ,Reversed P463 ,Q432786 ,231	Q337531 ,Reversed P463 ,Q401557

Table 7: 20 questions and the corresponding paths selected in WIKI.

Datasets	MRR	H@1	H@3	H@10
ICEWS14	44.76±0.17	35.64±0.10	49.49±0.05	62.30±0.04
ICEWS18	32.45±0.13	22.97±0.07	36.05±0.05	49.36±0.03
WIKI	80.41±0.15	76.04±0.09	84.91±0.06	85.95±0.05
YAGO	90.24±0.18	87.91±0.12	92.65±0.09	92.77±0.08

Table 8: Mean and standard deviation of IMR across ten runs on four datasets.

$$\begin{aligned}
\mathbf{e_{q\_i}} + \mathbf{r_{q\_i}} &= \mathbf{e_{q\_i-1}} + \mathbf{r_{pi}} + \mathbf{r_{q\_i-1}} - \mathbf{r_{pi}} \\
&= \mathbf{e_{q\_i-1}} + \mathbf{r_{q\_i-1}} \\
&= \mathbf{e_{q\_0}} + \mathbf{r_{q\_0}} \\
&= \mathbf{e_q} + \mathbf{r_q} \\
&= \mathbf{e_o}
\end{aligned} \tag{14}$$

This cancellation of relation guarantees that the answer to questions will not change along with the paths. In addition, the offset will not appear in the calculation of the indicator. Only the subject of the question is applied in the calculation of Path Confidence, and only the relation in the question is used in the calculation of Answer Completing Level.

A.5 The Robustness of IMR

We experiment on all datasets ten times by using ten different random seeds with fixed hyperparameters. Table 8 shows the mean and standard deviation of IMR on these datasets. It shows that IMR demonstrates a small standard deviation, which indicates its robustness.