Mutually-paced Knowledge Distillation for Cross-lingual Temporal Knowledge Graph Reasoning

Ruijie Wang*
Department of Computer Science, University of Illinois Urbana-Champaign
ruijiew2@illinois.edu

Zheng Li†
Amazon.com Inc
amzzhe@amazon.com

Jingfeng Yang
Amazon.com Inc
jingfe@amazon.com

Tianyu Cao
Amazon.com Inc
caoty@amazon.com

Chao Zhang
School of Computational Science and Engineering, Georgia Institute of Technology
chaozhang@gatech.edu

Bing Yin
Amazon.com Inc
alexbyin@amazon.com

Tarek Abdelzaher‡
Department of Computer Science, University of Illinois Urbana-Champaign
zaher@illinois.edu

ABSTRACT

This paper investigates cross-lingual temporal knowledge graph reasoning problem, which aims to facilitate reasoning on Temporal Knowledge Graphs (TKGs) in low-resource languages by transfering knowledge from TKGs in high-resource ones. The cross-lingual distillation ability across TKGs becomes increasingly crucial, in light of the unsatisfying performance of existing reasoning methods on those severely incomplete TKGs, especially in low-resource languages. However, it poses tremendous challenges in two aspects. First, the cross-lingual alignments, which serve as bridges for knowledge transfer, are usually too scarce to transfer sufficient knowledge between two TKGs. Second, temporal knowledge discrepancy of the aligned entities, especially when alignments are unreliable, can mislead the knowledge distillation process. We correspondingly propose a mutually-paced knowledge distillation model MP-KD, where a teacher network trained on a source TKG can guide the training of a student network on target TKGs with an alignment module. Concretely, to deal with the scarcity issue, MP-KD generates pseudo alignments between TKGs based on the temporal information extracted by our representation module. To maximize the efficacy of knowledge transfer and control the noise caused by the temporal knowledge discrepancy, we enhance MP-KD with a temporal cross-lingual attention mechanism to dynamically estimate the alignment strength. The two procedures are mutually paced along with model training. Extensive experiments on twelve cross-lingual TKG transfer tasks in the EventKG benchmark demonstrate the effectiveness of the proposed MP-KD method.

CCS CONCEPTS

• **Computing methodologies** → **Temporal reasoning**.

KEYWORDS

Temporal Knowledge Graph, Cross-lingual Transfer, Knowledge Distillation, Self-training

ACM Reference Format:
Ruijie Wang, Zheng Li, Jingfeng Yang, Tianyu Cao, Chao Zhang, Bing Yin, and Tarek Abdelzaher. 2023. Mutually-paced Knowledge Distillation for Cross-lingual Temporal Knowledge Graph Reasoning. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583407

1 INTRODUCTION

Temporal Knowledge Graphs (TKGs) [2, 15, 28, 42] characterize temporally evolving events, where each event, represented as (*subject*, *relation*, *object*), is associated with temporal information (*time*), e.g., (*Macron*, *reelected*, *French president*, *2022*). TKGs has facilitated a wide spectrum of knowledge-intensive Web applications with timeliness, such as question answering [31], product recommendation [39, 41, 54, 55], and social event forecasting [17, 30, 33, 38, 40].

As new events are continually emerging, modern TKGs are still far from being complete. Conventionally, the TKG construction process relies primarily on information extraction from unstructured corpus [8, 15, 28], which necessitates extensive manual annotations

*Part of work was done during internship at Amazon
†Corresponding authors.
‡Corresponding authors.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583407

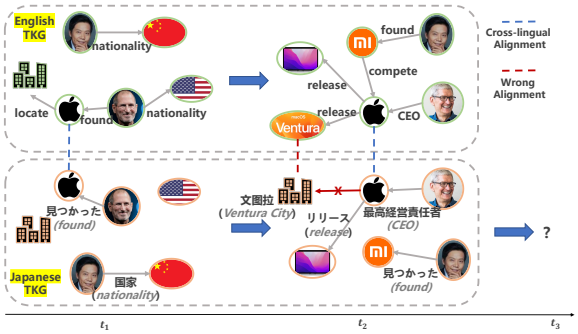


Figure 1: An illustrative example of cross-lingual reasoning on TKGs. 1) We aim to transfer knowledge from English TKG to Japanese TKG, where the English version provides more complete information; 2) Cross-lingual alignments only cover a small ratio of entities, e.g., Apple Inc; 3) Cross-lingual alignments can be noisy and misleading, e.g., A city called Ventura is linked to new macOS Ventura at t_2, introducing noise for reasoning in Japanese.

to keep up with changing events. For instance, the recent transition from Trump to Biden as the President of the United States has not been reflected in many TKGs, highlighting the need for timely updates. This spurs research on temporal knowledge graph reasoning to automate evolving events prediction over time [6, 13, 19, 37]. Unfortunately, the problem of TKG incompleteness is particularly pronounced in low-resource languages, where it is unable to collect enough corpus and annotations to support robust TKG construction. This results in suboptimal reasoning performance and distinctly unsatisfying accuracy in predicting recent and future events.

Inspired by the incompleteness issue facing low-resource languages in constructing TKGs, we introduce a novel task named Cross-Lingual Temporal Knowledge Graph Reasoning (as shown in Figure 1). This task aims to alleviate the reliance on supervision for TKGs in low-resource languages (referred to as the target language) by transferring temporal knowledge from high-resource languages (referred to as the source language) [1]. In contrast, all the existing efforts are either limited to reasoning in monolingual TKGs (usually high-resource languages, e.g., English) [6, 13, 19, 37], or multilingual static KGs [4, 12, 34]. To the best of our knowledge, cross-lingual TKG reasoning that transfers temporal knowledge between TKGs has not been investigated.

The fulfillment of this task poses tremendous challenges in two aspects: 1) **Scarcity of cross-lingual alignment**: as the informative bridge of two separate TKGs, cross-lingual alignment is imperative for cross-lingual knowledge transfer [4, 12, 34]. However, obtaining alignments between languages is a time-consuming and resource-intensive process that heavily relies on human annotations. The transfer of knowledge through a limited number of alignments is often insufficient to fully enhance the TKG in the target language. 2) **Temporal knowledge discrepancy**: the information associated with two aligned entities is not necessarily identical, especially with regards to temporal patterns. Utilizing a rough approach to equate the aligned entities at all times can result in the transfer of misleading knowledge and negatively impact performance. This becomes more pronounced when the alignments are noisy and unreliable. For example, at the time step t_2, a new event about operating system "*Ventura*" from Apple company occurs in the source English TKG, and meanwhile there is a noisy aligned entity "*Ventura city*" in the target Japanese TKG. Directly pulling those two entities at this point, can inevitably introduce noise and fail to predict a set of related events in the target TKG. Therefore, it is crucial to dynamically regulate the alignment strength of each local graph structure over time in order to maximize the effectiveness of cross-lingual knowledge distillation.

In this paper, we propose a novel Mutually-paced Knowledge Distillation (MP-KD) framework, where a teacher network learns more enriched temporal knowledge and reasoning skills from the source TKG to facilitate the learning of a student network in the low-data target one. The knowledge transfer is enabled via an alignment module, which estimates entity correspondence across languages based on temporal patterns. Firstly, to alleviate the limited language alignments (**Challenge #1**), such a knowledge distillation process is mutually paced over time. This means, on one hand, we encourage the mutually interactive learning between the teacher and student. Concretely, the alignment module between the teacher and the student learns to generate pseudo alignment between TKGs to maximally expand the upper bound of knowledge transfer. And subsequently, it empowers the student to encode more informative knowledge in target TKG, which can in turn boost the alignment module to explore more reasonable alignments as the bridge across TKGs. One the other hand, inspired by self-paced learning [14, 47], we make the generations as a progressively easy-to-hard process over time. We start from generating reliable pseudo data with high confidence. As time goes by, we then gradually increase the generation amount by relieving the restriction over time. Secondly, to inhibit the temporal knowledge mismatch (**Challenge #2**), the attention module can estimate the graph alignment strength distribution over time. This is achieved by a temporal cross-lingual attention in terms of the local graph structure and temporal-evolving patterns of aligned entities. As such, it can dynamically control the negative effect and suppress noise propagation from the source TKG. Moreover, we provide a theoretical convergence guarantee for the training objective on both initial ground-truth data and pseudo data. To evaluate MP-KD, we conduct extensive experiments of 12 cross-lingual TKG transfer tasks in multilingual EventKG dataset [8]. Our empirical results show that the MP-KD method outperforms state-of-the-art baselines in both with and without alignment noise settings, where only 20% of temporal events in the target KG and 10% of cross-lingual alignments are preserved.

To sum up, our contributions are three-fold:

- **Problem formulation**: We propose the cross-lingual temporal knowledge graph reasoning task, to boost the temporal reasoning performance in target TKG by transferring knowledge from source TKG;
- **Novel framework**: We propose a novel MP-KD framework, which enables the mutually-paced learning between the teacher and student networks, to promote both pseudo alignments and knowledge transfer reliability. Besides, MP-KD involves a

[1] In this paper, for the sake of brevity, we interchangeably use the terms high-resource/low-resource and source/target.

Table 1: Symbols and Notations.

Symbol	Definition
(e, r, e', t)	A quadruple in TKG.
$\mathcal{G}_s$, $\mathcal{G}_t$	Source TKG and Target TKG.
e_s, e_t	Entities in the source and target TKGs.
$\Gamma_{s\leftrightarrow t}$	Alignments between the source and target TKGs.
$\tilde{\mathcal{G}}_t$, $\tilde{\Gamma}_{s\leftrightarrow t}$	Incomplete target TKG and alignments.
$\tilde{\mathcal{G}}_t^{ST}$, $\tilde{\Gamma}_{s\leftrightarrow t}^{ST}$	Pseudo target TKG and pseudo alignment.
$f(\cdot;\Theta_s)$	Teacher network on the source TKG.
$f(\cdot;\Theta_t)$	Student network on the target TKG.
$g(e_s, e_t, t; \Phi)$	Alignment module measuring correspondence of (e_s, e_t) at t.
$\mathcal{L}_{\tilde{\mathcal{G}}_t}$, $\mathcal{L}_{\tilde{\mathcal{G}}_t^{ST}}$	Reasoning loss on groundtruth/pseudo target TKG.
$\mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}}$, $\mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}^{ST}}$	Alignment loss on groundtruth/pseudo alignment pairs.
$\mathcal{L}_{s\to t}$	Cross-lingual reasoning loss from source TKG to target TKG.
$\mathcal{L}_{s\to t}^{ST}$	Cross-lingual reasoning loss on both groundtruth and pseudo data.

dynamic alignment estimation across TKGs that inhibits the influence of temporal knowledge discrepancy.

- **Extensive evaluations**: Empirically, extensive experiments on 12 cross-lingual TKG transfer tasks in multilingual EventKG benchmark dataset demonstrate the effectiveness of MP-KD.

2 PRELIMINARIES AND NOTATIONS

In this section, we formally define the cross-lingual temporal knowledge graph reasoning task, and summarize the notations in Table 1. A temporal knowledge graph can be defined as follows:

*Definition 2.1 (***Temporal Knowledge Graph***)*. A temporal knowledge graph (TKG) is denoted as $\mathcal{G} = \{(e, r, e', t) | t \leq T\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$, where $\mathcal{E}$ denotes the entities set, $\mathcal{R}$ denotes the relation set, $\mathcal{T}$ denotes the timestamp set, and T denotes the latest update time. Each quadruple (e, r, e', t) refers to an event that a subject entity $e \in \mathcal{E}$ has a relation $r \in \mathcal{R}$ with an object entity $e' \in \mathcal{E}$ at timestamp $t \in \mathcal{T}$.

*Definition 2.2 (***Multilingual TKGs and Alignments***)*. To denote multilingual TKGs, we further utilize subscript to represent specific languages, i.e., $\mathcal{G}_s$ denotes TKG in the source language and $\mathcal{G}_t$ denotes TKG in the target language. The corresponding entities can be denoted as e_s and e_t respectively. Given two different languages s,t, we have the cross-lingual alignment set $\Gamma_{s\leftrightarrow t}$. To be more practical, we further assume the TKG in target language $\mathcal{G}_t$ and alignment set $\Gamma_{s\leftrightarrow t}$ are incomplete: $\tilde{\mathcal{G}}_t$, $\tilde{\Gamma}_{s\leftrightarrow t}$.

Based on the definition above, we formalize our cross-lingual reasoning task on TKGs as follows:

*Definition 2.3 (***Cross-lingual reasoning on TKGs***)*. Given the TKG $\mathcal{G}_s$ in the source language and the incomplete TKG $\tilde{\mathcal{G}}_t$ in the target language before the latest update time T, and the incomplete cross-lingual alignment $\tilde{\Gamma}_{s\leftrightarrow t}$, we aim to predict future events in the target TKG after time T. Concretely, we aim to predict missing entity in each future quadruple: $\{(e_t, r, ?, t)$ or $(?, r, e'_t, t) | t > T\}$ in the target TKG.

3 METHODOLOGY

In this section, we present the proposed MP-KD framework for the cross-lingual temporal knowledge graph reasoning task.

3.1 Overview

Figure 2 shows an overview of MP-KD. Given the TKGs in source language and target language, the teacher network and the student network first represent the source and target TKGs in a temporally evolving uni-space respectively. To facilitate training of the student, the knowledge distillation is enabled by a cross-lingual alignment module and an explicit temporal event transfer process. To deal with the scarcity issue of cross-lingual alignments, we propose a pseudo alignment generation technique to facilitate the knowledge distillation process, which is mutually-paced along with model training. To address the temporal knowledge discrepancy issue, the alignment module pulls the aligned entities close to each other based on the alignment strength which is dynamically adjusted.

Section 3.2 and Section 3.3 introduce our teacher/student network and the knowledge distillation respectively, followed by Section 3.4 which details how we generate pseudo alignments. Finally, Section 3.5 specifies our learning objective on both groundtruth data and pseudo data, and summarizes the training of MP-KD.

3.2 The Teacher/Student Network

We train two identical temporal representation modules on source and target TKGs with different parameters. The representation module $f(\cdot;\Theta)$ parameterized by Θ is designed to measure the plausibility of each quadruple, which represents each entity e into a low-dimensional latent space at each time: $\mathbf{h}_e(t) \in \mathbb{R}^d$. On a TKG $\mathcal{G}$, entities $e \in \mathcal{E}$ are evolving, as they interact with different entities over time. Such temporally interacted entities are defined as temporal neighbors. Therefore, we aim to model the temporal pattern of each entity e by encoding the changes of temporal neighbors.

Towards this goal, $f(\cdot;\Theta)$ first samples temporal neighbors $\mathcal{N}_e(t)$ from the TKG for each entity $e \in \mathcal{E}$. $\mathcal{N}_e(t)$ consists of a set of the most recently interacted entities at time t. Then $f(\cdot;\Theta)$ attentively aggregates information from the temporal neighbors. Specifically, given the temporal neighbor $\mathcal{N}_e(t)$, we represent the entity e as $\mathbf{h}_e(t)$ at time t:

$$\mathbf{h}_e^l(t) = \sigma\left(\sum_{(e_i, r_i, t_i) \in \mathcal{N}_e(t)} \alpha_{e,e_i}^l \left(\mathbf{h}_{e_i}^{l-1}(t_i)\mathbf{W}\right)\right), \tag{1}$$

where l denotes the layer number, $\sigma(\cdot)$ denotes the activation function *ReLU*, α_{e,e_i}^l denotes the attention weight of entity e_i to the represented entity e, and $\mathbf{W}$ is the trainable transformation matrix. To aggregate from history, α_{e,e_i}^l is supposed to be aware of entity feature, time delay and topology feature induced by relations. Thus, we design the attention weight α_{e,e_i}^l as follows:

$$\alpha_{e,e_i}^l = \frac{\exp(q_{e,e_i}^l)}{\sum_{(e_k, r_k, t_k) \in \mathcal{N}_e(t)} \exp(q_{e,e_k}^l)}, \; q_{e,e_k}^l = \mathbf{a}\left(\mathbf{h}_e^{l-1} \| \mathbf{h}_{e_k}^{l-1} \| \mathbf{h}_{r_k} \| \kappa(t - t_k)\right), \tag{2}$$

where q_{e,e_k}^l measures the pairwise importance by considering the entity embedding, relation embedding and time embedding, $\mathbf{a} \in \mathbb{R}^{4d}$ is the shared parameter in the attention mechanism. Following [5] we adopt random Fourier features as time encoding $\kappa(\Delta t)$ to reflect the time difference.

To measure plausibility of each possible quadruple, we utilize TransE [1] as the score function $f(e, r, e', t; \Theta) = -\|\mathbf{h}_e^l(t) + \mathbf{h}_r -$

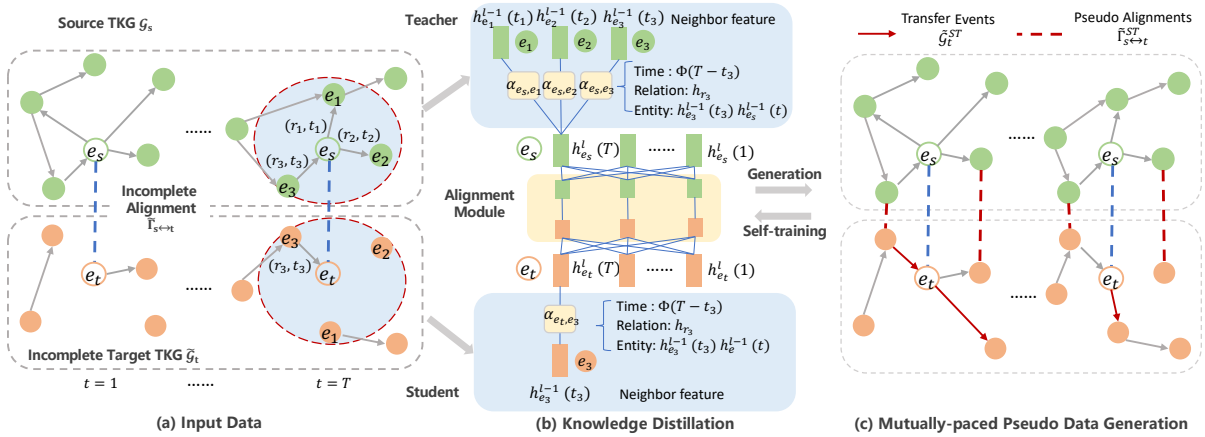


Figure 2: An overview of MP-KD. (a) The source TKG is more complete than the target TKG, and the cross-lingual alignments are also scarce; (b) A teacher/student representation module to represent source/target TKG, and an alignment module for knowledge transfer; (c) Mutually-paced knowledge distillation between knowledge transfer and pseudo alignment generation.

$\mathbf{h}^l_{e'}(t)\|^2$, where true quadruples should have higher scores. To optimize the parameter Θ on a TKG $\mathcal{G}$, we set the objective to rank the scores of true quadruples higher than all other false quadruples produced by negative sampling:

$$\mathcal{L}_{\mathcal{G}} = \mathop{\mathbb{E}}_{(e,r,e',t)\in\mathcal{G}} \left[\max(0, \lambda_1 - f(e,r,e',t;\Theta) + f(e,r,e^-,t;\Theta))\right], \quad (3)$$

where (e,r,e^-,t) is negative samples with object e' replaced by e^-, λ_1 is the margin to distinguish positive and negative quadruples.

3.3 Knowledge Distillation

The incomplete target TKG, $\tilde{\mathcal{G}}t$, can be used to train the corresponding parameter Θ_t through minimization of $\mathcal{L}\tilde{\mathcal{G}}_t$. However, the low-resource nature of the target language often results in an incomplete target TKG, leading to suboptimal Θ_t. In light of this, we propose a knowledge distillation approach to transfer temporal knowledge from the source TKG to the target TKG. The proposed approach consists of two components: an alignment module that enhances Θ_t using the more informative Θ_s learned from the source TKG, and an explicit temporal event transfer based on the improved parameters. This integrated approach aims to improve the completeness and quality of the target TKG by leveraging the knowledge contained in the source TKG.

The Alignment Module. In general, the source parameters Θ_s provide a more informative representation of each entity $e \in \mathcal{E}$ compared to the target parameters Θ_t. To take advantage of this, we utilize Θ_s to guide the optimization of Θ_t through the alignment module $g(\cdot;\Phi)$, which measures the correspondence between each pair of entities and is parameterized by Φ.

Directly pulling embeddings of aligned entities at all time steps can transfer misleading knowledge due to the temporal knowledge discrepancy. Therefore, the alignment module first utilizes a temporal attention layer to integrate information of each entity from history in both source and target TKGs, i.e., $\mathbf{H}^s_e(t), \mathbf{H}^t_e(t) \in \mathbb{R}^d$, then it pulls such integration $\mathbf{H}^s_e(t)$ close to $\mathbf{H}^t_e(t)$ instead of the initial $\mathbf{h}^s_e(t)$ and $\mathbf{h}^t_e(t)$. Moreover, the temporal integration $\mathbf{H}^s_e(t)$ and $\mathbf{H}^t_e(t)$ also encode the temporal evolution information for each entity, which can be utilized to estimate the adaptive alignment strength at different time to improve the alignment module. Concretely, the temporal integration is learned by:

$$\begin{aligned}\mathbf{H}^s_e(t) &= \text{Temporal-Attn}(\mathbf{h}^s_e(1), \mathbf{h}^s_e(2), \cdots, \mathbf{h}^s_e(t)),\\ \mathbf{H}^t_e(t) &= \text{Temporal-Attn}(\mathbf{h}^t_e(1), \mathbf{h}^t_e(2), \cdots, \mathbf{h}^t_e(t)),\end{aligned} \quad (4)$$

$$g(e_s, e_t, t; \Phi) = \frac{\mathbf{H}^s_e(t) \cdot \mathbf{H}^t_e(t)}{\|\mathbf{H}^s_e(t)\|_2 \cdot \|\mathbf{H}^t_e(t)\|_2}, \quad (5)$$

where Temporal-Attn is the temporal attention network designed to integrate information on the temporal domain. The correspondence between each pair of entities (e_s, e_t) across source and target languages at time t is measured by $g(e_s, e_t, t; \Phi)$. As the temporal knowledge for aligned entities is not identical, the alignment strength between them should vary across time t. The alignment strength is strong when the two entities share similar information, and weak when the information is dissimilar or the alignment is unreliable. This variability is achieved through the design of a trainable weight $\beta_{e,t}$ to adjust the alignment strength for different entities at different times, which is generated by a cross-lingual attention layer:

$$\beta_{e,t} = \text{Cross-Attn}(key = \mathbf{H}^t_e(1:T), query = \mathbf{H}^s_e(1:T))_{tt}. \quad (6)$$

Due the page limitation, we refer readers to Appendix A.1 for the detailed implementation of Temporal-Attn$(\cdot)$ and Cross-Attn$(\cdot)$.

To optimize the parameter Φ on the incomplete alignments $\tilde{\Gamma}_{s\leftrightarrow t}$, we set the objective in order to rank the correspondence of true alignments higher than false alignments:

$$\mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}} = \mathop{\mathbb{E}}_{\tilde{\Gamma}_{s\leftrightarrow t}} \left[\mathop{\mathbb{E}}_{t\in\mathcal{T}} \left[\beta_{e,t} \cdot \max(0, \lambda_2 - g(e_s, e_t, t; \Phi) + g(e_s, e^-_t, t; \Phi))\right]\right], \quad (7)$$

where the entity pair $(e_s, e_t) \in \tilde{\Gamma}_{s\leftrightarrow t}$ is the aligned entities across languages, (e_s, e^-_t) is the negative samples, λ_2 is the margin value.

Temporal Event Transfer. Cross-lingual alignments offer the potential to directly transfer temporal events towards the progressive completion of the target TKG. This is based on the premise that entities that are reliably aligned are likely to experience similar temporal events across languages, with the same relations.

Given an aligned pair (e_s, e_t), the temporal event $(e_t, r, e^?_t, t)$ or $(e^?_t, r, e_t, t)$ is added to the target TKG if the corresponding event

$(e_s, r, e_s^?, t)$ or $(e_s^?, r, e_s, t)$ exists in the source TKG $\mathcal{G}_s$. To determine the missing entity $e_t^?$, we first verify if $(e_s^?, e_t^?)$ is present in the alignment set. If so, the temporal event is directly added to the target TKG. Otherwise, the updated student network $f(\cdot;\Theta_t)$ is utilized to predict the missing entity and the top-1 entity is utilized to complete the temporal event. We define the set of transferred temporal events in the target TKG as $\tilde{\mathcal{G}}_t^{ST}$ for ease of discussion.

3.4 Generating Pseudo Alignments

The limited amount of cross-lingual alignments negatively constrain the effect of the knowledge distillation process. In this section, we introduce how to generate pseudo alignments $\tilde{\Gamma}_{s\leftrightarrow t}^{ST}$ with high confidence to boost cross-lingual transfer effectiveness.

To expand the range of alignments used in the knowledge transfer process, we generate pseudo-alignments with high confidence scores and incorporate them into the training data. The confidence score for each pair of entities (e_s, e_t) is calculated as the average cosine similarity: $sim(e_s, e_t) = \mathop{\mathbb{E}}_{t}[g(e_s, e_t, t; \Phi)]$. While pair-wise similarity comparison is computationally intensive, we improve efficiency by first adding alignments for entities that are neighbors of already aligned entities $\tilde{\mathcal{E}}t = \{e_t | (e_s, e_t) \in \tilde{\Gamma}_{s\leftrightarrow t}\}$ in the target TKG, as they are likely to be represented well to produce reliable alignment. Following [36], we formulate the generation process as solving the following optimization problem:

$$\begin{aligned} \max \sum_{e_t \in \mathcal{N}(\tilde{\mathcal{E}}_t)} \sum_{e_s \in \mathcal{E}_s} sim(e_s, e_t) \cdot \phi(e_s, e_t), \\ s.t. \sum_{e_s \in \mathcal{E}_s} \phi(e_s, e_t) = 1, \quad \sum_{e_t \in \mathcal{N}(\tilde{\mathcal{E}}_t)} \phi(e_s, e_t) = 1, \end{aligned} \tag{8}$$

where $\phi(e_s, e_t)$ is a binary indicator of whether to add (e_s, e_t) as pseudo alignment, $\phi(e_s, e_t) = 1$ if we choose to add this pair, otherwise $\phi(e_s, e_t) = 0$. The two constrains can guarantee each entity $e_t \in \mathcal{N}(\tilde{\mathcal{E}}_t)$ is aligned to at most one entity in source language $e_s \in \mathcal{E}_s$. Finally, all pairs that satisfying $\phi(e_s, e_t) = 1$ can be viewed as candidates to be added into alignment data. We further select the top ones in terms of $sim(e_s, e_t)$ to control the total size of pseudo alignments. Notably, in each generation, the target entities to be aligned can already have the alignment, i.e., $\mathcal{N}(\tilde{\mathcal{E}}_t) \bigcup \tilde{\mathcal{E}}_t \neq \emptyset$. In this case, we can update the existing alignments with the pseudo ones to eliminate the possible alignment noise.

3.5 Mutually-paced Optimization

Learning Objective. Given a source TKG $\mathcal{G}_s$, the incomplete target TKG $\tilde{\mathcal{G}}_t$, and the incomplete cross-lingual alignment $\tilde{\Gamma}_{s\leftrightarrow t}$, the objective of cross-lingual temporal knowledge graph reasoning $\mathcal{L}_{s\to t}$ can be summarized as follows:

$$\mathcal{L}_{s\to t} = \mathcal{L}_{\tilde{G}_t} + \mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}}, \tag{9}$$

where $\mathcal{L}_{\tilde{G}_t}$ denotes knowledge graph reasoning loss which measures the correctness of each quadruple, $\mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}}$ denotes the alignment loss which measures the distance of aligned entities in the uni-space. To enlarge the knowledge distillation effect, we progressively transfer temporal events $\tilde{\mathcal{G}}_t$ and generate high-quality pseudo alignment $\tilde{\Gamma}_{s\leftrightarrow t}^{ST}$. Therefore, the training objective on both ground-truth data and pseudo data $\mathcal{L}_{s\to t}^{ST}$ becomes:

$$\begin{aligned} \mathcal{L}_{s\to t}^{ST} = \frac{|\tilde{\mathcal{G}}_t|}{|\tilde{\mathcal{G}}_t| + |\tilde{\mathcal{G}}_t^{ST}|} \cdot \mathcal{L}_{\tilde{\mathcal{G}}_t} + \frac{|\tilde{\mathcal{G}}_t^{ST}|}{|\tilde{\mathcal{G}}_t| + |\tilde{\mathcal{G}}_t^{ST}|} \cdot \mathcal{L}_{\tilde{\mathcal{G}}_t^{ST}} \\ + \frac{|\tilde{\Gamma}_{s\leftrightarrow t}|}{|\tilde{\Gamma}_{s\leftrightarrow t}| + |\tilde{\Gamma}_{s\leftrightarrow t}^{ST}|} \cdot \mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}} + \frac{|\tilde{\Gamma}_{s\leftrightarrow t}^{ST}|}{|\tilde{\Gamma}_{s\leftrightarrow t}| + |\tilde{\Gamma}_{s\leftrightarrow t}^{ST}|} \cdot \mathcal{L}_{\tilde{\Gamma}_{s\leftrightarrow t}^{ST}}, \end{aligned} \tag{10}$$

where $|\cdot|$ denotes the set size. Eq. (10) formulates the learning objective on both scarce data and pseudo data for cross-lingual temporal knowledge graph reasoning in target languages. We give the convergence analysis in the following theorem:

Algorithm 1: The optimization process for MP-KD.

Input: Source TKG $\mathcal{G}_s$, incomplete target TKG $\tilde{\mathcal{G}}_t$, incomplete alignment $\tilde{\Gamma}_{s\leftrightarrow t}$.
Output: Student Model Parameter Θ_t for target TKG.

```
Optimize Θ_s by minimizing L_{G_s} on source TKG;
Initialize Θ_t ← Θ_s for target TKG;
while model not converged do
    Optimize alignment module g(·;Φ):
    Minimize L^{ST}_{s→t} in Eq. (10) w.r.t. alignment parameter Φ;
    Transfer temporal events G̃^{ST}_t based on updated Θ_t;
    Optimize student representation module f_t(·;Θ_t):
    for Each time step T_i during training do
        Prepare training data {(e_t, r, e'_t, t) | (e_t, r, e'_t, t) ∈
          G̃_t ⋃ G̃^{ST}_t and T_i < t < T_{i+1}}
        Update Θ_t by minimizing L^{ST}_{s→t} in Eq. (10);
    end
    Generate pseudo alignments Γ̃^{ST}_{s↔t} based on updated Φ;
end
```

Theorem 3.1. *Let N denote the number of negative samples for optimization, ϵ denotes the portion of correct pseudo data, β denotes the proportion of pseudo data to the initial ground-truth data. As the number of negative samples $N \to \infty$, the $\mathcal{L}_{s\to t}^{ST}$ converges to its limit with an absolute deviation decaying in $O(\frac{1+\epsilon}{1+\beta} \cdot N^{-2/3})$.*

Proof. Refer to Appendix A.2. □

Mutually-paced Optimization and Generation. The MP-KD framework can be optimized by minimizing Eq. (10) w.r.t. Θ and Φ alternatively. The generated pseudo alignments can help the training of the representation modules by the knowledge distillation, and in turn transferring temporal events in the target TKG can improves alignment module by providing high-quality representations. In light of this, we propose a mutually-paced optimization and generation procedure. Generally speaking, we iteratively generate pseudo alignments and update the representation module and alignment module respectively. To be concrete, as shown in Algorithm 1, we first update the alignment module $g(\cdot;\Phi)$ and transfer temporal events to transfer knowledge from source to target. Then we divide the time span into several time steps to update the student representation module from recent time step to far away ones. Finally, we generate the pseudo alignments, as the optimization Θ_t on all temporal events can improve the entity feature quality, which is beneficial for alignment prediction. Algorithm 1 summarizes the training procedure.

Table 2: Statistics of the datasets.

Languages	Entity	Relation	Quadruple	Train/Val/Test	Time
English (EN)	34,416	105	602K	602K/0K/0K	28
French (FR)	32,546	105	580K	580K/0K/0K	28
Spanish (ES)	31,808	105	316K	114K/136K/66K	40
German (DE)	27,657	105	268K	97K/114K/56K	40
Italian (IT)	23,734	94	236K	84K/100K/51K	40
Danish (DA)	15,710	94	125K	48K/50K/26K	40
Slovene (SL)	13,250	94	55K	24K/21K/10K	40
Bulgarian (BG)	3,508	105	23K	8K/9K/6K	40

4 EXPERIMENT

We evaluate MP-KD on EventKG data [8] including 2 source languages and 6 target languages, and we aim to answer the following research questions:

- **RQ1**: How does MP-KD perform compared with state-of-the-art models on the low-resource target languages?
- **RQ2**: How do reliability of alignment information (with various noise ratio) affect model performances?
- **RQ3**: How do each component and important parameters affect MP-KD performance?

4.1 Datasets

We evaluate MP-KD by 12 cross-lingual TKG transfer tasks on EventKG data [8], which is a multilingual TKG including 2 source languages and 6 target languages. For each language, we collect events during 1980 to 2022 and split the time span into 40 time steps for training, validation and testing (28/4/8). Table 2 shows the dataset statistics. We describe the dataset details in Appendix A.3.

4.2 Experimental Setup

Baselines. We compare ten state-of-the-art baselines from three related areas. We describe the baseline details in Appendix A.4.

- Static KG embedding methods: **TransE** [1], **TransR** [23], **DistMult** [51], **RotatE** [35];
- Temporal KG embedding methods: **TA-DistMult** [6], **RE-NET** [13], **RE-GCN** [19];
- Multilingual KG embedding methods: **KEnS** [4]; **AlignKGC** [34]; **SS-AGA** [12].

Evaluation Protocol and Metrics. For each prediction $(e, r, ?, t)$ or $(?, r, e, t)$, we rank missing entities to evaluate the performance. Following [19], we adopt raw mean reciprocal rank (*MRR*) and raw Hits at 10 (*H@10*) as evaluation metrics. To quantitatively compare how well the transferred knowledge from the source languages can improve predictions on the low-resource languages, we adopt Transfer Ratio (*T.R.*) to evaluate the average improvement of each method over the best baseline without knowledge transferring, i.e.:

$$T.R.(t_i) = \frac{1}{|S|} \sum_{s_i \in \mathcal{S}} \frac{\text{Model}(s_i \rightarrow t_i)}{\text{BestBaseline}(t_i)} \tag{11}$$

where t_i denotes each target language, $s_i \in \mathcal{S}$ denotes each source language, and $\text{BestBaseline}(t_i)$ denotes the best baseline performance on the target language t_i without any knowledge transferring, i.e., *RE-GCN w/o source*.

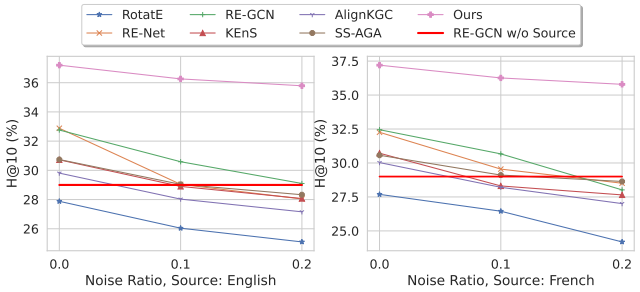


Figure 3: Experimental results under various alignment noise ratios. Average H@10 on 6 target languages are reported. MP-KD achieves relatively robust results, with only 3.7% relative drop, others have over 10% drop.

Implementation. To simulate scarce setting, by default, we utilize 10% alignments and 20% events of target TKG by random selection. For static/temporal KG embedding methods, we merge source graph and target graph by adding one new type of relation (alignment). For multilingual baselines, we train them on 1-to-1 knowledge transferring (instead of the original setting) for fair comparison. We introduce implementation details of baseline models and MP-KD in Appendix A.5. Code and data are open-source and available at https://github.com/amzn/mpkd-thewebconf-2023.

4.3 Experiments on Cross-lingual Reasoning (RQ1)

We first evaluate the model performance with incomplete cross-lingual alignments, where we randomly preserve 10% alignments of the target entities for distilling knowledge. Table 3 reports the overall results for the cross-lingual experiments. By utilizing only 10% cross-lingual alignments, MP-KD achieves 33% (*MRR*) and 30% (*H@10*) relative improvement over best baseline without the knowledge transferring (*RE-GCN w/o source*) on average, demonstrating the effectiveness of MP-KD in modeling alignments for knowledge transferring. Compared with ten baselines using alignments, MP-KD still achieves relative 14% relative improvements over the second best results. Specifically, we have the following observations:

- Static baseline (*TransE, TransR, DistMult, RotatE*) fail to beat *RE-GCN w/o source*, although using alignments, due the insufficient modeling of temporal information. Similarly, multilingual methods (*KEnS, AlignKGC, SS-AGA*) also produce unsatisfying results;
- All temporal baselines (*TA-DistMult, RE-Net, RE-GCN*) manage to beat *RE-GCN w/o source*, as the modeling of both temporal evolution and cross-lingual alignment can facilitate the representation learning of target entities. But the improvements are marginal compared with our model, as the effect of knowledge distillation is constrained by the limited amount of cross-lingual alignments;
- Our model consistently achieves the best performance. Through 10% alignments, MP-KD can progressively transfer temporal knowledge and generate pseudo alignments with high confidence to boost the effect and range of the knowledge distillation;
- We also notice the uneven improvements across languages, (e.g., 40% improvements for German, 20% for Italian). We hypothesize it is because of various language dependencies with source languages.

Table 3: Overall Performance without alignment noise. Average results on 5 independent runs are reported. $*$ indicates the statistically significant improvements over the best baseline, with p-value smaller than 0.01.

Models	Target	ES		DE		IT		DA		BG		SL		Avg.	
	Source	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
RE-GCN w/o source	**NA**	14.31	31.85	16.32	34.19	14.59	31.64	14.19	31.24	10.27	23.44	9.33	21.63	13.17	29.00
Static KG embedding methods															
TransE [1]	**EN**	11.67	26.73	15.19	31.37	9.15	21.44	12.71	23.31	10.17	23.72	9.73	21.83	11.44	24.73
	FR	12.37	27.79	14.01	28.30	11.38	23.19	10.05	22.10	11.88	23.01	10.63	22.44	11.72	24.47
	T.R.	0.84	0.86	0.89	0.87	0.70	0.71	0.80	0.73	1.07	1.00	1.09	1.02	0.88	0.85
TransR [23]	**EN**	11.88	28.66	16.01	32.01	8.14	22.07	13.34	24.73	10.33	23.51	8.89	22.12	11.43	25.52
	FR	12.01	28.32	14.58	29.51	9.93	24.66	11.90	22.64	11.98	23.44	9.27	23.88	11.61	25.41
	T.R.	0.83	0.89	0.94	0.90	0.62	0.74	0.89	0.76	1.09	1.00	0.97	1.06	0.87	0.88
DistMult [51]	**EN**	13.66	29.77	17.46	33.19	11.63	26.63	14.63	25.91	9.97	22.92	9.08	20.44	12.74	26.48
	FR	12.58	28.73	16.03	31.81	12.12	27.76	11.64	22.97	9.01	23.77	10.13	21.07	11.92	26.02
	T.R.	0.92	0.92	1.03	0.95	0.81	0.86	0.93	0.78	0.92	1.00	1.03	0.96	0.94	0.91
RotatE [35]	**EN**	12.99	28.89	19.87	35.46	15.62	30.14	13.44	25.79	11.10	22.98	11.37	23.99	14.07	27.88
	FR	13.01	29.33	17.63	34.81	14.99	31.04	11.62	23.17	10.73	23.14	11.10	24.66	13.18	27.69
	T.R.	0.91	0.91	1.15	1.03	1.05	0.97	0.88	0.78	1.06	0.98	1.20	1.12	1.03	0.96
Temporal KG embedding methods															
TA-DistMult [6]	**EN**	15.83	34.77	18.99	37.46	14.98	29.99	14.97	30.01	9.02	21.10	8.74	17.76	13.75	28.51
	FR	16.61	35.83	17.81	37.96	15.58	31.21	13.21	28.58	9.63	22.91	9.03	18.83	13.65	29.22
	T.R.	1.13	1.11	1.13	1.10	1.05	0.97	0.99	0.94	0.91	0.94	0.95	0.85	1.04	1.00
RE-Net [13]	**EN**	17.58	37.97	19.03	39.46	15.88	33.69	15.03	34.77	12.01	25.72	11.07	25.64	15.10	32.88
	FR	17.01	36.79	18.32	38.07	15.47	34.83	15.63	33.86	12.31	25.03	11.79	24.97	15.09	32.26
	T.R.	1.21	1.17	1.14	1.13	1.07	1.08	1.08	1.10	1.18	1.08	1.23	1.17	1.15	1.12
RE-GCN [19]	**EN**	16.88	36.54	19.84	40.17	16.17	34.84	15.99	35.62	12.22	26.02	10.63	23.38	15.29	32.76
	FR	17.14	37.01	19.63	41.01	16.44	35.61	15.03	33.19	11.91	25.13	11.09	22.77	15.21	32.45
	T.R.	1.19	1.15	1.21	1.19	1.12	1.11	1.09	1.10	1.17	1.09	1.16	1.07	1.16	1.12
Multilingual KG embedding methods															
KEnS [4]	**EN**	15.98	33.91	17.33	37.62	14.41	31.44	14.47	29.61	12.88	26.77	11.03	24.99	14.35	30.72
	FR	17.02	34.07	16.61	37.99	15.57	33.82	13.62	30.24	12.03	24.32	10.51	23.86	14.23	30.72
	T.R.	1.15	1.07	1.04	1.11	1.03	1.03	0.99	0.96	1.21	1.09	1.15	1.13	1.09	1.06
AlignKGC [34]	**EN**	13.59	33.19	16.44	33.14	13.71	34.07	12.13	31.07	11.33	26.63	8.32	20.77	12.59	29.81
	FR	13.90	34.71	17.14	34.81	14.97	33.65	12.07	30.44	10.92	25.31	9.64	21.28	13.11	30.03
	T.R.	0.96	1.07	1.03	0.99	0.98	1.07	0.85	0.98	1.08	1.11	0.96	0.97	0.98	1.03
SS-AGA [12]	**EN**	15.11	32.19	16.49	36.14	14.83	33.31	12.27	30.68	12.99	27.03	11.55	25.07	13.87	30.74
	FR	16.54	33.99	18.32	37.19	15.02	32.99	11.73	29.98	11.13	25.62	11.01	23.64	13.96	30.57
	T.R.	1.11	1.04	1.07	1.07	1.02	1.05	0.85	0.97	1.17	1.12	1.21	1.13	1.06	1.06
MP-KD *	**EN**	**19.51**	41.55	**22.84**	**49.30**	17.18	37.62	**18.79**	**40.01**	**14.33**	**30.13**	**13.87**	**30.30**	**17.75**	**38.15**
	FR	19.05	**42.86**	21.67	46.57	**17.92**	**39.18**	17.95	37.95	13.85	29.27	12.54	27.36	17.16	37.20
	T.R.	1.35	1.33	1.36	1.40	1.20	1.21	1.29	1.25	1.37	1.27	1.42	1.33	1.33	1.3
Gains		11%	13%	15%	20%	9%	10%	18%	12%	10%	11%	18%	18%	16%	16%

4.4 Experiments under Alignment Noises (RQ2)

In reality, cross-lingual alignments can be obtained by human labeling or rule-based inference modules, which may introduce indispensable noises. We evaluate how the reliability of alignment information affects baseline models and MP-KD. In this experiment, we still utilize 10% alignments. To simulate unreliable alignments, we select a subset of alignments (measured by *Noise Ratio*) and randomly change the aligned target entity to another entity without alignment information.

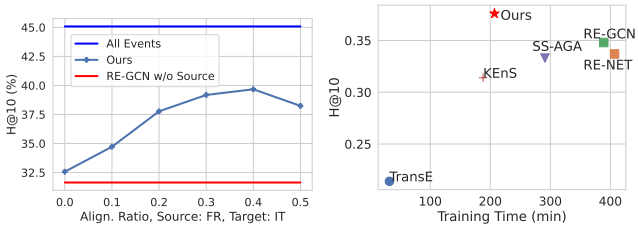
We vary the noise ratio from 0.0 to 0.2 to evaluate the models performance, as shown in Figure 3. We report the average H@10 on 6 target languages by utilizing English TKG and French TKG respectively. As expected, with the increase of noise ratio, the performances of all compared models degrade, as the wrong alignment links mislead the knowledge transfer process. Most baselines fail to beat *RE-GCN w/o Source* even with 10% noise, and all lose with 20%*noise*, which indicates that the quality of alignments significantly influences the model effectiveness in the cross-lingual TKG reasoning task. Notably, MP-KD achieves relatively robust results, with only 3.7% performance drop, while other strong baselines have over 10% drop. This is because during the generation of pseudo alignments, MP-KD can automatically replace those unreliable ones based on the confidence score. Also, in the alignment module, MP-KD can assign small alignment strength to unreliable alignments.

Table 4: Ablation Studies.

Ablations	Target	ES		SL		Avg.	
	Source	MRR	H@10	MRR	H@10	MRR	H@10
MP-KD w/o Align. Strength Control	**EN**	17.61	38.59	13.07	28.51	16.24	37.04
	FR	17.07	39.46	13.03	28.14	16.33	36.61
	T.R.	1.21	1.23	1.27	1.21	1.21	1.22
MP-KD w Pure Training	**EN**	17.09	37.03	11.89	26.25	14.81	31.90
	FR	16.99	37.10	11.78	25.91	14.97	32.15
	T.R.	1.19	1.16	1.15	1.11	1.13	1.09
MP-KD w/o Pseudo Align.	**EN**	17.97	38.04	12.31	27.83	15.98	34.83
	FR	17.55	38.45	12.19	26.32	15.79	35.27
	T.R.	1.24	1.20	1.19	1.16	1.22	1.19
MP-KD w/o Event Transfer	**EN**	18.79	39.03	13.07	28.07	16.03	36.74
	FR	18.83	39.88	12.95	28.79	15.99	36.95
	T.R.	1.31	1.24	1.27	1.21	1.27	1.24
MP-KD	**EN**	**19.51**	41.55	**14.33**	**30.13**	**17.75**	**38.15**
	FR	19.05	**42.86**	13.85	29.27	17.16	37.20
	T.R.	1.35	1.33	1.37	1.27	1.33	1.3

4.5 Model Analysis (RQ3)

Ablation Study. We evaluate performance improvements brought by the MP-KD framework by following ablations:



(a) Pseudo align. analysis. **(b) Efficiency analysis.**

Figure 4: Efficiency analysis and pseudo alignment analysis.

- **MP-KD w/o Align. Strength Control** uniformly set the alignment strength for all entities across all time steps;
- **MP-KD w Pure Training** optimizes the teacher-student framework without pseudo alignment generation and temporal event transfer;
- **MP-KD w/o Pseudo Align** eliminates the pseudo alignments generation process;
- **MP-KD w/o Event Transfer** eliminates the explicit transfer of temporal events.

Table 4 reports the results measured by $H@10$. Each component leads to performance boost. MP-KD with uniform alignment strength largely degrades performance, due to temporal knowledge discrepancy. MP-KD without pseudo data generation achieves similar performance with temporal baselines *RE-Net, RE-GCN*, because of the limited amount of cross-lingual alignments. Bother generating pseudo alignments and explicitly transferring temporal events increase the performance, and combining them together in a mutually-paced procedure (in MP-KD) can achieve the best results.

The Effect of Pseudo Alignments Ratio. To investigate the effects of the pseudo alignments on the reasoning performance, we vary the amount of pseudo alignments during training period and compare the corresponding performance measured by *H@10*, as shown in Figure 4a. The blue line and red line show the performances of single model on complete target TKG and single model on 20% target TKG (our setting) respectively. From 0.1, MP-KD starts to generate and expand the initially available alignments. We observe a significant performance improvement, demonstrating the positive effects of the pseudo alignments. As expected, we find a performance decrease at 50%, as the added pseudo data with relatively low confidence start to introduce noise that hurt the performance.

4.6 Efficiency Comparison

To demonstrate the efficiency of MP-KD framework, we train MP-KD and baseline models from scratch on both target language and source language, and compare the training time. Figure 4b shows that MP-KD significantly outperforms baseline models with reasonable training time. More details are provided in Appendix A.5.3.

5 RELATED WORK

Knowledge Graph Reasoning. Knowledge graph reasoning aims to predict missing facts to automatically complete KGs [15, 16, 28, 42]. It is mostly formulated as measuring the correctness of factual samples and negative samples by specially designed score functions [1, 25, 26, 35]. Recently, reasoning on temporal KGs attracts a lot of interests from the community [6, 13, 19, 37]. Compared with static KG reasoning task, the main challenge lies in how to incorporate time information. Several embedding-based methods have been proposed. They encode time-dependent information of entities and relations by decoupling embeddings into static component and time-varying component [7, 43, 46], utilizing recurrent neural networks (RNNs) to adaptively learn the dynamic evolution from historical fact sequence [13, 19], or learning a sequence of evolving representations from discrete knowledge graph snapshots [10, 13, 19, 22]. However, all of the existing temporal KG reasoning models aim to extrapolate future facts based on relatively complete TKGs in high-resource languages, and how to boost reasoning performance for TKGs in low-resource languages through cross-lingual alignments is largely under-explored.

Multilingual KG Reasoning. Entity alignment methods on KGs [3, 27, 48–50] can automatically enlarge the alignments by predicting the correspondence between the two KGs. But most of them, if not all, require the relatively even completeness of two KGs to capture the structural similarities, which can not be satisfied in our case, as target TKGs are far from complete. Inspired by recent cross-lingual transfer advances [20, 21, 29, 52], some recent works start to study the multilingual KG reasoning on static graphs [4, 12, 34], which aim to extract knowledge from several source KGs to boost the reasoning performance in the target KG, while they still require a sufficient amount of seed alignments and totally ignore the temporal information in our task. We extend this line of works on TKGs, where transferring temporal knowledge is more complex.

Self-training. Self-training is one of the learning strategies that addresses data scarcity issue by fully utilizing abundant unlabeled data [11, 32, 53]. Recent works start to study self-training strategy for graph data, as GNNs typically require large amount of data labeling [18, 24, 45]. In summary, most efforts are put on node classification problem, where node labels are largely unavailable. We focus on utilizing self-training technique to deal with link scarcity (events + alignments), which is also a bottleneck for improving the performance on graphs.

6 CONCLUSION

In this paper, we studied a realistic but underexplored cross-lingual temporal knowledge graph reasoning problem, which aims at facilitating TKG reasoning in low-resource languages by distilling knowledge from a corresponding TKG in high-resource language through a small set of entity alignments as bridges. To this end, we proposed a novel mutually-paced teacher student framework, namely MP-KD. During training, MP-KD iteratively generates pseudo alignments to expand the cross-lingual connection, as well as transfers temporal facts to facilitate student model training in low-resource languages. Our alignment module is learned to adjust the alignment strength for different entities at different time, thereby maximizing the benefits of knowledge transferring. We empirically validated the effectiveness of MP-KD on 12 language pairs of EventKG data, on which the proposed framework significantly outperforms an extensive set of state-of-the-art baselines.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. Research reported in this paper was sponsored in part by DARPA award HR001121C0165, DARPA award HR00112290105, Basic Research Office award HQ00342110002, the Army Research Laboratory under Cooperative Agreement W911NF-17-20196, and Amazon.com Inc.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*.

[2] Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. ICEWS Coded Event Data. (2015).

[3] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual Knowledge Graph Embeddings for Cross-Lingual Knowledge Alignment. In *IJCAI'17*. 1511–1517.

[4] Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun, and Carlo Zaniolo. 2020. Multilingual Knowledge Graph Completion via Ensemble Knowledge Transfer. *CoRR* abs/2010.03158 (2020). arXiv:2010.03158 https://arxiv.org/abs/2010.03158

[5] da Xu, chuanwei ruan, evren korpeoglu, sushant kumar, and kannan achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*.

[6] Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. 2018. Learning Sequence Encoders for Temporal Knowledge Graph Completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[7] Rishab Goel, Seyed Mehran Kazemi, Marcus A. Brubaker, and Pascal Poupart. 2020. Diachronic Embedding for Temporal Knowledge Graph Completion. In *AAAI'20*.

[8] Simon Gottschalk and Elena Demidova. 2019. EventKG: the hub of event knowledge on the web and biographical timeline generation. *Semantic Web* (2019).

[9] M. Gutmann and A. Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS) (JMLR W & CP, Vol. 9)*, Y.W. Teh and M. Titterington (Eds.). 297–304.

[10] 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 *NeurIPS'21*.

[11] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting Self-Training for Neural Sequence Generation. *CoRR* abs/1909.13788 (2019). arXiv:1909.13788 http://arxiv.org/abs/1909.13788

[12] Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. 2022. Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment. In *ACL'22*.

[13] Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. ReNet: Autoregressive Structure Inference over Temporal Knowledge Graphs. In *EMNLP*.

[14] M. Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-Paced Learning for Latent Variable Models. In *NIPS'10*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Eds.), Vol. 23. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf

[15] Julien Leblay and Melisachew Wudage Chekol. 2018. Deriving Validity Time in Knowledge Graph. In *WWW '18*.

[16] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 167–195. http://jens-lehmann.org/files/2015/swj_dbpedia.pdf

[17] Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher. 2022. Unsupervised Belief Representation Learning with Information-Theoretic Variational Graph Auto-Encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*.

[18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. In *AAAI'18* (New Orleans, Louisiana, USA). Article 433, 8 pages.

[19] Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021. Temporal Knowledge Graph Reasoning Based on Evolutional Representation Learning. In *SIGIR*.

[20] Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020. Learn to Cross-lingual Transfer with Meta Graph Learning Across Heterogeneous Languages. In *EMNLP'20*. 2290–2301.

[21] Zheng Li, Danqing Zhang, Tianyu Cao, Ying Wei, Yiwei Song, and Bing Yin. 2021. Metats: Meta teacher-student network for multilingual sequence labeling with minimal supervision. In *EMNLP'21*. 3183–3196.

[22] Siyuan Liao, Shangsong Liang, Zaiqiao Meng, and Qiang Zhang. 2021. Learning Dynamic Embeddings for Temporal Knowledge Graphs. In *WSDM '21*.

[23] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *AAAI'15*.

[24] Hongrui Liu, Binbin Hu, Xiao Wang, Chuan Shi, Zhiqiang Zhang, and Jun Zhou. 2022. Confidence May Cheat: Self-Training on Graph Neural Networks under Distribution Shift. *CoRR* abs/2201.11349 (2022). arXiv:2201.11349 https://arxiv.org/abs/2201.11349

[25] Lihui Liu, Boxin Du, Yi Ren Fung, Heng Ji, Jiejun Xu, and Hanghang Tong. 2021. KompaRe: A Knowledge Graph Comparative Reasoning System. In *KDD '21* (Virtual Event, Singapore). Association for Computing Machinery, New York, NY, USA, 3308–3318. https://doi.org/10.1145/3447548.3467128

[26] Lihui Liu, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. 2022. Joint Knowledge Graph Completion and Question Answering. In *KDD '22* (Washington DC, USA). Association for Computing Machinery, New York, NY, USA, 1098–1108. https://doi.org/10.1145/3534678.3539289

[27] Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, and Jie Tang. 2022. SelfKG: Self-Supervised Entity Alignment in Knowledge Graphs. *CoRR* abs/2203.01044 (2022). arXiv:2203.01044 https://arxiv.org/abs/2203.01044

[28] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias.. In *CIDR*.

[29] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[30] Jianwei Qian, Xiang-Yang Li, Chunhong Zhang, Linlin Chen, Taeho Jung, and Junze Han. 2019. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. *IEEE Transactions on Dependable and Secure Computing* (2019), 679–692.

[31] Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question Answering Over Temporal Knowledge Graphs. In *ACL'21*.

[32] H. J. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theory* 11 (1965), 363–371.

[33] Huajie Shao, Shuochao Yao, Andong Jing, Shengzhong Liu, Dongxin Liu, Tianshi Wang, Jinyang Li, Chaoqi Yang, Ruijie Wang, and Tarek Abdelzaher. 2020. Misinformation Detection and Adversarial Attack Cost Analysis in Directional Social Networks. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. 1–11. https://doi.org/10.1109/ICCCN49398.2020.9209609

[34] Harkanwar Singh, Soumen Chakrabarti, PRACHI JAIN, Sharod Roy Choudhury, and Mausam. 2021. Multilingual Knowledge Graph Completion With Joint Relation and Entity Alignment. In *Conference on Automated Knowledge Base Construction*.

[35] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

[36] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *IJCAI'18* (Stockholm, Sweden). 4396–4402.

[37] Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. 2017. Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs. In *ICML'17 (ICML'17)*. 3462–3471.

[38] Haiwen Wang, Ruijie Wang, Chuan Wen, Shuhao Li, Yuting Jia, Weinan Zhang, and Xinbing Wang. 2020. Author Name Disambiguation on Heterogeneous Information Network with Adversarial Representation Learning. In *AAAI '20*.

[39] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. In *CIKM '18*.

[40] Ruijie Wang, Zijie Huang, Shengzhong Liu, Huajie Shao, Dongxin Liu, Jinyang Li, Tianshi Wang, Dachun Sun, Shuochao Yao, and Tarek Abdelzaher. 2021. DyDiff-VAE: A Dynamic Variational Framework for Information Diffusion Prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*.

[41] Ruijie Wang, Zheng Li, Danqing Zhang, Qingyu Yin, Tong Zhao, Bing Yin, and Tarek Abdelzaher. 2022. RETE: Retrieval-Enhanced Temporal Event Forecasting on Unified Query Product Evolutionary Graph. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. 462–472.

[42] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. AceKG: A Large-Scale Knowledge Graph for Academic Data Mining. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*.

[43] Ruijie Wang, zheng li, Dachun Sun, Shengzhong Liu, Jinning Li, Bing Yin, and Tarek Abdelzaher. 2022. Learning to Sample and Aggregate: Few-shot Reasoning over Temporal Knowledge Graphs. In *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=1LmgISIDZJ

[44] Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *CoRR* abs/2005.10242 (2020). arXiv:2005.10242 https://arxiv.org/abs/2005.10242

[45] Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. 2021. Be Confident! Towards Trustworthy Graph Neural Networks via Confidence Calibration. In *NeurIPS'21*.

[46] Chenjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Yazdi, and Jens Lehmann. 2020. Temporal Knowledge Graph Completion Based on Time Series Gaussian Embedding. In *The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I*. 654–671.

[47] Miao Xu, Bingcong Li, Gang Niu, Bo Han, and Masashi Sugiyama. 2019. Revisiting Sample Selection Approach to Positive-Unlabeled Learning: Turning Unlabeled Data into Positive rather than Negative. *CoRR* abs/1901.10155 (2019). arXiv:1901.10155 http://arxiv.org/abs/1901.10155

[48] Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic Knowledge Graph Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35.

[49] Yuchen Yan, Si Zhang, and Hanghang Tong. 2021. BRIGHT: A Bridging Algorithm for Network Alignment. In *Proceedings of the Web Conference 2021*.

[50] Yuchen Yan, Qinghai Zhou, Jinning Li, Tarek Abdelzaher, and Hanghang Tong. 2022. Dissecting Cross-Layer Dependency Inference on Multi-Layered Inter-Dependent Networks. In *CIKM '22*.

[51] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and li Deng. 2014. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.

[52] Juan Zha, Zheng Li, Ying Wei, and Yu Zhang. 2022. Disentangling Task Relations for Few-shot Text Classification via Self-Supervised Hierarchical Task Clustering. In *EMNLP'22*.

[53] Danqing Zhang, Zheng Li, Tianyu Cao, Chen Luo, Tony Wu, Hanqing Lu, Yiwei Song, Bing Yin, Tuo Zhao, and Qiang Yang. [n. d.]. Queaco: Borrowing treasures from weakly-labeled behavior data for query attribute value extraction. In *CIKM'21*.

[54] Yuyue Zhao, Xiang Wang, Jiawei Chen, Yashen Wang, Wei Tang, Xiangnan He, and Haiyong Xie. 2022. Time-Aware Path Reasoning on Knowledge Graph for Recommendation. *ACM Trans. Inf. Syst.* (2022).

[55] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph Based Semantic Fusion. In *KDD '20 (KDD '20)*. 1006–1014.

A APPENDIX

A.1 Model Description

In this section, we introduce the temporal attention layer and cross-lingual attention layer for entity alignments utilized in Section 3.3. We first introduce the general attention mechanism we utilized, then specify the two layers respectively.

Given two representation sequence from temporal domain: key sequence $\boldsymbol{H}_K = \{\boldsymbol{h}_K^1, \boldsymbol{h}_K^2, \cdots, \boldsymbol{h}_K^T\}$ and query sequence $\boldsymbol{H}_Q = \{\boldsymbol{h}_Q^1, \boldsymbol{h}_Q^2, \cdots, \boldsymbol{h}_Q^T\}$ from all time steps, we propose the following attention to calculate the pairwise importance:

$$\beta = \text{Attn}(key = \boldsymbol{H}_K, query = \boldsymbol{H}_Q) = \text{softmax}\left(\frac{\boldsymbol{H}_Q \boldsymbol{W}^Q (\boldsymbol{H}_K \boldsymbol{W}^K)^T}{\sqrt{d}} + \boldsymbol{M}\right), \tag{12}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^L$ are trainable temporal parameters, β is learned temporal weight indicating pairwise importance, d denotes dimension of input representations, and $\boldsymbol{M}$ is added to ensure auto-regressive setting, i.e., preventing future information affecting current state. We define $\boldsymbol{M}_{ij} = 0$ if $i \le j$, otherwise $-\infty$.

For *temporal attention* layer, we use $\boldsymbol{h}_e = \{\boldsymbol{h}_e(1), \boldsymbol{h}_e(2), \cdots, \boldsymbol{h}_e(T)\}$ for both query and key sequence to obtain the temporal attention weights β:

$$\beta = \text{Attn}(key = \boldsymbol{h}_e, query = \boldsymbol{h}_e), \tag{13}$$

then the desired $\boldsymbol{H}_e(t)$ is leaned as the combination of input sequence, where $\boldsymbol{W}^V$ is a trainable matrix.:

$$\begin{aligned} \boldsymbol{H}_e(t) &= \text{Temporal-Attn}(\boldsymbol{h}_e(1), \cdots, \boldsymbol{h}_e(t)) \\ &= \sum_{i=1}^{t} \beta_{it} \boldsymbol{h}_e(i) \boldsymbol{W}^V \end{aligned} \tag{14}$$

For *cross-lingual attention* layer, we use $\boldsymbol{H}_e^s = \{\boldsymbol{H}_e^s(1), \cdots, \boldsymbol{H}_e^s(T)\}$ in source language as query sequence and $\boldsymbol{H}_e^t = \{\boldsymbol{H}_e^t(1), \cdots, \boldsymbol{H}_e^t(T)\}$ in target language as key sequence to obtain the attention weights β:

$$\beta_{e,t} = \text{Attn}(key = \boldsymbol{H}_e^t, query = \boldsymbol{H}_e^s)_{tt}, \tag{15}$$

where $\beta_{e,t}$ is trainable weight to adjust the alignment strength of different entities at different time.

A.2 Theorem Proof

Theorem A.1. *Let N denote the number of negative samples for optimization, ϵ denotes the ratio of correct pseudo data, β denotes the ratio of pseudo data amount to the initial groundtruth data amount. As the number of negative samples $N \to \infty$, the $\mathcal{L}_{s\to t}^{ST}$ converges to its limit with an absolute deviation decaying in $O(\frac{1+\epsilon}{1+\beta} \cdot N^{-2/3})$.*

Proof. In representation learning, the margin loss has been widely adopted as the similarity metric. Without loss of generality, they can be expressed in the form of Noise Contrastive Estimation (NCE) [9]. Therefore, we express $\mathcal{L}_{\mathcal{G}}$ and $\mathcal{L}_{\Gamma_{s\leftrightarrow t}}$ in the form of Noise Contrastive Estimation (NCE) by introducing the negative sampling:

$$\mathcal{L}_{\mathcal{G}} \triangleq \mathbb{E}\left[-\log \frac{e^{f(\cdot;\Theta)/\tau}}{e^{f(\cdot;\Theta)/\tau} + \sum_{-} e^{(f^-(\cdot;\Theta))/\tau}}\right], \tag{16}$$

$$\mathcal{L}_{\Gamma_{s\leftrightarrow t}} \triangleq \mathbb{E}\left[-\log \frac{e^{g(\cdot;\Phi))/\tau}}{e^{g(\cdot;\Phi)/\tau} + \sum_{-} e^{g^-(\cdot;\Phi)/\tau}}\right], \tag{17}$$

for simplicity, $f^-(\cdot;\Theta)$ denotes the score for negative quadruple, and $g^-(\cdot;\Phi)$ denotes score for negative alignment pair.

For our training objective $\mathcal{L}_{s\to t}^{ST}$, we show the convergence analysis of four terms one by one, then prove the overall convergence results. First of all, following [27, 44], let N denote the number of negative samples per each quadruple, and we have:

$$\begin{aligned} &\lim_{N\to\infty}\left[\mathcal{L}_{\hat{\mathcal{G}}_t} - \log M\right] \\ &= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(e_t,r,e'_t,t)\in\hat{\mathcal{G}}_t} [f(\cdot;\Theta)] \\ &+ \lim_{N\to\infty} \mathop{\mathbb{E}}_{\substack{(e_t,r,e'_t,t)\in\hat{\mathcal{G}}_t \\ e_t^-\in\mathcal{E}_t}} \left[\log\left(\frac{\lambda}{N} e^{f(\cdot;\Theta)/\tau} + \frac{1}{N}\sum_{-} e^{f^-(\cdot;\Theta)/\tau}\right)\right] \\ &= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(e_t,r,e'_t,t)\in\hat{\mathcal{G}}_t} [f(\cdot;\Theta)] + \mathop{\mathbb{E}}_{(e_t,r,e'_t,t)\in\hat{\mathcal{G}}_t} \left[\log \mathop{\mathbb{E}}_{e_t^-\in\mathcal{E}_t} \left[e^{f^-(\cdot;\Theta)}\right]\right], \end{aligned} \tag{18}$$

where λ denotes the duplicate quadruples co-existing in both incomplete $\tilde{\mathcal{G}}_t$ and negative samples. The convergence speed is derived as follows:

For one side:

$$\mathcal{L}_{\hat{\mathcal{G}}_t} - \log N - \lim_{N\to\infty}\left[\mathcal{L}_{\hat{\mathcal{G}}_t} - \log N\right] \le \frac{\lambda}{N} e^{\frac{2}{\tau}}. \tag{19}$$

For another side:

$$\lim_{N\to\infty}\left[\mathcal{L}_{\hat{\mathcal{G}}_t} - \log N\right] - \left[\mathcal{L}_{\hat{\mathcal{G}}_t} - \log N\right] \le \frac{\lambda}{N} e^{2/\tau} + \frac{5}{4} N^{-\frac{2}{3}} e^{\frac{1}{\tau}} (e^{\frac{1}{\tau}} - e^{-\frac{1}{\tau}}). \tag{20}$$

We then generalize the above results to the loss term on pseudo data. Suppose ϵ of the pseudo data are correct. Then we can have the following two inequality. For one side:

$$\begin{aligned} &\mathcal{L}_{\tilde{\mathcal{G}}_t^{ST}} - \log N - \lim_{N\to\infty}\left[\mathcal{L}_{\tilde{\mathcal{G}}_t^{ST}} - \log N\right] \\ &\le \epsilon \mathop{\mathbb{E}}_{e\in\mathcal{E}_t} \left[\log \frac{e^{-\in\mathcal{E}_t}\left[\frac{\lambda}{N} e^{1/\tau} + e^{f^-(\cdot;\Theta_t)/\tau}\right]}{\mathop{\mathbb{E}}_{e^-\in\mathcal{E}_t} e^{f^-(\cdot;\Theta_t)/\tau}}\right] \le \epsilon \frac{\lambda}{N} e^{\frac{2}{\tau}} \end{aligned} \tag{21}$$

Therefore, for $\mathcal{L}_{s\to t}^{ST}$ in this side, we have:

$$\mathcal{L}_{s\to t}^{ST} - \log N - \lim_{N\to\infty}\left[\mathcal{L}_{s\to t}^{ST} - \log N\right] \le \frac{1+\epsilon}{1+\beta} \frac{\lambda}{N} e^{\frac{2}{\tau}}, \tag{22}$$

where β is the ratio of pseudo data amount to groundtruth data amount during training.

Similarly, for another side, we have:

$$\begin{aligned} \lim_{N\to\infty}\left[\mathcal{L}_{s\to t}^{ST} - \log N\right] - \left[\mathcal{L}_{s\to t}^{ST} - \log N\right] &\le \frac{1+\epsilon}{1+\beta} \frac{\lambda}{N} e^{2/\tau} \\ &+ \frac{1+\epsilon}{1+\beta} \frac{5}{4} N^{-\frac{2}{3}} e^{\frac{1}{\tau}} (e^{\frac{1}{\tau}} - e^{-\frac{1}{\tau}}). \end{aligned} \tag{23}$$

Therefore, we conclude that the $\mathcal{L}_{s\to t}^{ST}$ converges to its limit with an absolute deviation decaying in $O(\frac{1+\epsilon}{1+\beta} \cdot N^{-2/3})$

□

A.3 Datasets

Dataset Information. The commonly utilized benchmark TKGs are divided into two categories: temporal event graphs [2] and knowledge graphs where temporally associated facts have valid periods [15, 16, 28]. In this paper, we mainly evaluate MP-KD on the EventKG [8], which is a multilingual resource incorporating event-centric information extracted from several large-scale knowledge graphs such as Wikidata [15], DBpedia [16] and YAGO [28]. Each temporal event is organized as (e, r, e', t_s, t_e), where each piece of data is attached with a valid time period from start time t_s to end time t_e. Following [13], we preprocess the format such that each fact is converted to a sequence $\{(e, r, e', t_s), (e, r, e', t_s +$

1), $\cdots$, $(e, r, e', t_e)\}$ from t_s to t_e, with the minimum time unit as one step.

Splitting Scheme. We collect events during 1980 to 2022, and noisy events of early years are removed. To construct multilingual TKGs, we first preserve important entities and relations by excluding infrequent ones that have less than 20 events in each language. Then we collect the events and cross-lingual alignments.To guarantee the relation match, we only preserve relations appearing in English TKG. We split the time span into 40 equal time steps for training, validation and testing (28/4/8), where each time step roughly lasts for one year. To focus on the prediction on existing entities during training period, and eliminate the negative effects possibly caused by the randomly appearning new entities in val/test period, we only preserve entities having events during training period, following [19]. Table 2 shows the dataset statistics, including 2 source languages and 6 target languages. We purposefully choose 6 different target languages with diverse characteristics in term of the TKG size, which can evaluate MP-KD from different data granularity. It is worth noting that to simulate the scarcity issue in target TKGs, the training quadruples presented in Table 2 are randomly selected from original TKGs, with random ratio 20%.

A.4 Baselines

We describe the baselines utilized in the experiments in detail:

- **TransE** [1] is a translation-based embedding model, where both entities and relations are represented as vectors in the latent space. The relation is utilized as a translation operation between the subject and the object entity;
- **TransR** [23] advances TransE by optimizing modeling of n-n relations, where each entity embedding can be projected to hyperplanes defined by relations;
- **DistMult** [51] is a general framework with bilinear objective for multi-relational learning that unifies most multi-relational embedding models;
- **RotatE** [35] represents entities as complex vectors and relations as rotation operations in a complex vector space;
- **TA-DistMult** [6] is a temporal knowledge graph reansoing method aiming at predicting missing events in history. We utilize it for predicting future events;
- **RE-NET** [13] is a generative model to predict future facts on temporal knowledge graphs, which employs a recurrent neural network to model the entity evolution, and utilizes a neighborhood aggregator to consider the connection of facts at the same time intervals;
- **RE-GCN** [19] learns the temporal representations of both entities and relations by modeling the KG sequence recurrently;
- **KEnS** [4] starts to directly improve KGR performance on static KGs given a set of seed alignment, and proposes an ensemble-based approach for the task;
- **AlignKGC** [34] jointly optimizes entity alignment loss and knowledge graph reasoning loss to improve the performance;
- **SS-AGA** [12] views alignments as new edge type and employ a relation-aware GNN with learnable attention weight to model the influence of the aligned entities.

A.5 Reproducibility

A.5.1 Baseline Setup. For static knowledge graph reasoning methods, i.e., TransE, TransR, DistMult, and RotatE, we ignore all time information in quadruples, and view temporal knowledge graphs as static, cumulative ones. For static/temporal KG embedding methods, we merge source graph and target graph by adding one new type of relation (alignment), as they do not explicitly model cross-lingual entity alignment. For multilingual baselines, we train them on 1-to-1 knowledge transferring (instead of the original setting) for fair comparison. For static baselines, we utilize the static embeddings for predictions in all time steps. For fair comparisons, we keep the dimension of all embeddings as 128, we feed pre-trained TransE embeddings on the merge graph including both source and target TKGs to those that require initial entity/relation embeddings. We tune learning rate of baselines based on *MRR* on validation set, and we train all baseline models and MP-KD on same GPUs (Nvidia A100) and CPUs (Intel(R) Xeon(R) Platinum 8275CL).

A.5.2 MP-KD Setup. We first utilize the source TKG to train the teacher representation module. Then we initialize the student module with the parameters of the teacher. During the training procedure, we first optimize the objective without generating pseudo data in the first 10 epoch. After that, we start to generate high-quality pseudo data. For the generation in each epoch, we gradually increase the amount of pseudo alignments from 10% to 40%, and transfer all temporal events that meet the requirement. During evaluation, we tune hyperparameters based on *MRR* on validation set, and report the performance on the test set. Next, we report the choices of hyperparameters. For model training, we utilize Adam optimizer, and set maximum number of epochs as 50. We set batch size as 256, the dimension of all embeddings as 128, and dropout rate as 0.5. For the sake of efficiency, we set number of temporal neighbors b as 8, and employ 1 neighborhood aggregation layer in temporal encoder. For TKG reasoning, we set negative sampling factor as 10. For entity alignment, we set negative sampling factor as 50. For temporal generation process, We divide time span into 4 time intervals. For model training, we mainly tune margin value λ_1, λ_2 in score functions in range $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, learning rate in range $\{0.02, 0.01, 0.005, 0.001, 0.0005\}$.

A.5.3 Efficiency Comparison. To demonstrate the efficiency of MP-KD framework, we train MP-KD and baseline models from scratch on both target language and source language, and compare the training time. We train all baseline models and MP-KD on same GPUs (Nvidia A100) and CPUs (Intel(R) Xeon(R) Platinum 8275CL). Figure 4b shows that MP-KD significantly outperforms baseline models with reasonable training time. Notably, we include the pseudo data generation time. Compared with slow temporal models *RE-NET, RE-GCN* for knowledge graph reasoning, MP-KD is more efficient because our temporal encoder can learn temporal entity embeddings via sampled temporal neighbors at each time without using RNNs.