# TeMP: Temporal Message Passing for Temporal Knowledge Graph Completion

**Jiapeng Wu** **Meng Cao** **Jackie Chi Kit Cheung** **William L. Hamilton**
School of Computer Science, McGill University, Montreal, QC, Canada
MILA, Montreal, QC, Canada

{jiapeng.wu@mail, meng.cao@mail, jcheung@cs, wlh@cs}.mcgill.ca

## Abstract

Inferring missing facts in temporal knowledge graphs (TKGs) is a fundamental and challenging task. Previous works have approached this problem by augmenting methods for static knowledge graphs to leverage time-dependent representations. However, these methods do not explicitly leverage multi-hop structural information and temporal facts from recent time steps to enhance their predictions. Additionally, prior work does not explicitly address the temporal *sparsity* and *variability* of entity distributions in TKGs. We propose the **Te**mporal **M**essage **P**assing (TeMP) framework to address these challenges by combining graph neural networks, temporal dynamics models, data imputation and frequency-based gating techniques. Experiments[1] on standard TKG tasks show that our approach provides substantial gains compared to the previous state of the art, achieving a 10.7% average relative improvement in Hits@10 across three standard benchmarks. Our analysis also reveals important sources of variability both within and across TKG datasets, and we introduce several simple but strong baselines that outperform the prior state of the art in certain settings.

## 1 Introduction

The ability to infer missing facts in temporal knowledge graphs is essential for applications such as event prediction (Leblay and Chekol, 2018; De Winter et al., 2018), question answering (Jia et al., 2018), social network analysis (Zhou et al., 2018; Trivedi et al., 2019) and recommendation systems (Kumar et al., 2018).

Whereas static knowledge graphs (KGs) represent facts as triples (e.g., *(Obama, visit, China)*), temporal knowledge graphs (TKGs) additionally associate each triple with a timestamp (e.g., *(Obama, visit, China, 2014)*). Figure 1 shows a subgraph of such TKG. Usually, TKGs are assumed to consist of discrete timestamps (Jiang et al., 2016), meaning that they can be represented as a sequence of static KG snapshots, and the task of inferring missing facts across these snapshots is referred to as temporal knowledge graph completetion (TKGC).

Recent works on TKGC have largely focused on developing time-dependent scoring functions, which score the likelihood of missing facts and build closely upon popular representation learning methods for static KGs (Dasgupta et al., 2018; Jiang et al., 2016; Goel et al., 2019; Xu et al., 2019; Lacroix et al., 2020). However, while powerful, these existing methods do not properly account for multi-hop structural information in TKGs, and they lack the ability to explicitly leverage temporal facts in nearby KG snapshots to answer queries. Knowing facts like *(Obama, make agreement with, China, 2013)* or *(Obama, visit, China, 2012)* is useful for answering the query *(Obama, visit, ?, 2014)*.

Moreover—and perhaps more importantly—there are also serious challenges regarding *temporal variability* and *temporal sparsity*, which previous works fail to address. In real-world TKGs, models have access to *variable* amounts of reference temporal information in near KG snapshots when answering different queries (Figure 2 and Figure 6 in the Appendix). For example, in a political event dataset, there are likely to be more quadruples with subject-relation pair *(Obama, visit)* than *(Trump, visit)* from 2008 to 2013.[2] Hence the model could access more reference information to answer *where Obama visited in 2014*.

The *temporal sparsity* problem reveals that only a small fraction of entities are *active*[3] at each time

[1] Code and data are published at https://github.com/JiapengWu/TeMP

[2] Obama was the president of US during the period.

[3] An entity is *active* at a time step if it has at least one neighboring entity in the same KG snapshot.



*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5730–5746,
November 16–20, 2020. 

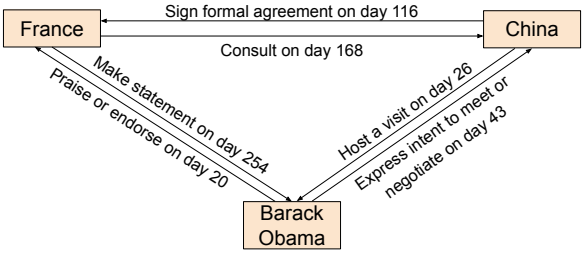



Figure 1: A sample temporal knowledge subgraph involving France, China and Barack Obama.

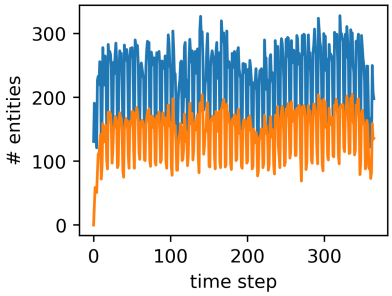



Figure 2: Dataset statistics of the ICEWS14 dataset. The blue (top) curve shows the number of active entities at each time step, while the orange (bottom) curve represents the number of active entities at each time step that are also active at least once in the past 15 time steps. While the total number of entities is 7,128, only 2% – 4% of these entities are active at each time step. (See Appendix A.5 for further examples and discussion).

step (Figure 7 in the Appendix). Previous methods usually assign the same embedding for *inactive* entities at different time steps, which is not fully representative of the time-sensitive features.

**Present work**. To address these issues, we introduce the **T**emporal **M**essage **P**assing (TeMP) framework, which combines neural message passing and temporal dynamic models. We then propose frequency-based gating and data imputation techniques to counter the temporal sparsity and variability issues.

We achieve state-of-the-art performance on standard TKGC benchmarks. In particular, on the standard ICEWS14, ICEWS05-15, and GDELT datasets, TeMP is able to provide an 7.3% average relative improvement in Hits@10 compared to the next-best model. Fine-grained error analysis on these three datasets demonstrates the unique contributions made by each of the different components of TeMP. Our analysis also highlights important sources of variability, in particular variations in temporal sparsity both within and across TKG datasets, and how effects of different components are affected by such variability.

## 2 Related Work

**Static KG representation learning** Much research exists on representation learning methods for static KGs, in which entities and relations are represented as low-dimensional embeddings (Nickel et al., 2011; Yang et al., 2014; Trouillon et al., 2016; Nickel et al., 2016). Generally, these methods involve a *decoding* method, which scores candidate facts based on entity and relation embeddings, and the models are optimized so that valid triples receive higher scores than random *negative examples*. While these methods typically rely on shallow encoders to generate the embeddings—i.e., single embedding-lookup layers (Hamilton et al., 2017)—message passing (or graph neural network; GNN) approaches have also been proposed (Schlichtkrull et al., 2018; Vashishth et al., 2019; Busbridge et al., 2019) to leverage multi-hop information around entities.

**Temporal KG representation learning** Recent works endeavor to extend static KGC models to the temporal domain. Typically, such approaches employ embedding methods with a shallow encoder and design time-sensitive quadruple decoding functions (Dasgupta et al., 2018; Jiang et al., 2016; Goel et al., 2019; Xu et al., 2019; Lacroix et al., 2020). While time-specific information is considered by these methods, entity-level temporal patterns such as event periodicity are not explicitly captured.

Another line of work on temporal (knowledge) graph reasoning uses message passing networks to capture intra-graph neighborhood information, which is sometimes combined with temporal recurrence or attention mechanisms (Manessi et al., 2020; Kumar et al., 2018; Pareja et al., 2019; Chen et al., 2018; Jin et al., 2019; Sankar et al., 2020; Hajiramezanali et al., 2019). Orthogonal to our work, Trivedi et al. (2017, 2019); Han et al. (2020) explore using temporal point processes. However, their focus is on continuous TKGC. The prior works that most resemble our framework are Recurrent Event Networks (RE-NET) (Jin et al., 2019) and DySAT (Sankar et al., 2020). RE-NET uses multi-level RNNs to model entity interactions, while DySAT uses self-attention to learn latent node representations on dynamic graphs. However, both these works were proposed for the task

of graph extrapolation (i.e., inferring the next time-step in a sequence), so they are not directly compatible with the TKGC setting.

## 3 Proposed Approach

We first define our key notation and provide an overview of our TeMP framework, before describing the individual components in detail in the following sections.

**Notation and task definition**. Our goal is to predict missing facts in a temporal knowledge graph (TKG) $\mathcal{G} = \{G^{(1)}, G^{(2)}, ..., G^{(T)}\}$, where $G^{(t)} = (E, R, D^{(t)})$. Here, $E$ and $R$ stand for the union of sets of entities and relations across all time steps and are known in advance. $D^{(t)}$ denotes the set of all *observed* triples $(s, r, o)$ at time $t$, with subjects $s \in E$ , objects $o \in E$ and relations $r \in R$. Let $\overline{D}^{(t)}$ denote the set of *true* triples at time $t$ such that $D^{(t)} \subseteq \overline{D}^{(t)}, \forall t$, the temporal knowledge graph completion (TKGC) problem is defined as ranking the subject and object entities given object queries $(s, r, ?, t)$ and subject queries $(?, r, o, t)$ where $(s, r, o) \in \overline{D}^{(t)}$ but $(s, r, o) \notin D^{(t)}, t \in \{0, ..., T\}$.

**Overview of TeMP**.

Following common practice, we structure our TeMP framework around the notion of an *encoder* and *decoder*. The encoder maps each entity $e_i \in E$ to time-dependent low-dimensional embedding $\boldsymbol{z}_{i,t}$ at each time-step $t$, while the decoder uses these entities' embeddings to score the likelihood of a temporal fact.

Figure 3 depicts the architecture of our model. A key insight in TeMP is that we use an encoder that combines a *structural* entity representation and *temporal* representations. The structural encoder (SE) based on a multi-relational message passing network produces entity representation $\boldsymbol{x}_{i,t} = \text{SE}(e_i, D^{(t)})$ while the temporal encoder (TE) integrates the output of SE at previous time steps to induce $\boldsymbol{z}_{i,t} = \text{TE}(\boldsymbol{x}_{i,t-\tau}, ..., \boldsymbol{x}_{i,t})$. Here $\tau$ stands for the number of temporal input KG snapshots to the model.

In addition, in Section 3.3, we propose a series of augmentations to TeMP that are designed to address the temporal sparsity and variability issues of real-world TKGs. Finally, in Section 3.4, we discuss how TeMP can leverage existing decoders from the static KG setting in order to train a model.

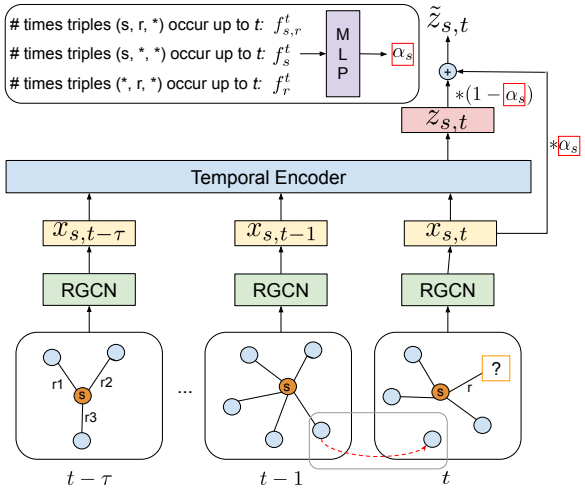



Figure 3: Architecture of TeMP Framework. TeMP combines structural graph encoder and temporal encoder to induce entity representations. Given query $(s, r, ?, t)$ at time $t$, TeMP takes graphs from time step $t-\tau$ to $t$ as input to compute structural embedding $\boldsymbol{x}_{s,t}$ and temporal embedding $\boldsymbol{z}_{s,t}$ for the centering entity $s$. The final representation $\tilde{\boldsymbol{z}}_{s,t}$ is obtained by further applying frequency-based gating, as illustrated in the upper rectangle. The red dotted arrow at the bottom indicates the imputation process for an inactive entity at time step $t$.

### 3.1 Structural Encoder

The first key component of TeMP is the structural encoder, which generates entity embeddings based on the graph $G^{(t)}$ within each time-step. We build our structural encoder by adapting existing techniques for message passing on static knowledge graphs (Schlichtkrull et al., 2018).

$$\boldsymbol{h}^{(0)}_{i,t} = \boldsymbol{W}_0 \boldsymbol{u}_i, \forall t \in 0, ..., T,$$
$$\boldsymbol{h}^{(l+1)}_{i,t} = \sigma\Big(\sum_{r \in R} \sum_{j \in N^r_i} \frac{1}{|N^r_i|} \boldsymbol{W}^{(l)}_r \boldsymbol{h}^{(l)}_{j,t} + \boldsymbol{W}^{(l)}_s \boldsymbol{h}^{(l)}_{i,t}\Big)$$

Here, $\boldsymbol{u}_i$ denotes a one-hot embedding indicating entity $e_i$, $\boldsymbol{W}_0$ is an entity embedding matrix, and $\boldsymbol{W}^{(l)}_r$ and $\boldsymbol{W}^{(l)}_s$ are transformation matrices specific to each layer of the model. These matrices are shared across all discrete time stamps. We use $N^r_i$ to denote the set of neighboring entities of $e_i$ connected by relation $r$, whose size acts as a normalizing constant for averaging the neighborhood information. After running $L$ layers of this message-passing approach on a snapshot $G^{(t)}$, we use $\boldsymbol{x}_{i,t} = \boldsymbol{h}^{(L)}_{i,t}$ to denote the resulting structural embedding of entity $e_i$, which summarizes its $L$-hop neighborhood within $G^{(t)}$.

While we focus on RGCN as the structural encoder, our framework is not tied to any specific multi-relational message passing network. One can swap RGCN with any multi-relational graph encoder, e.g. CompGCN (Vashishth et al., 2019) and EdgeGAT (Busbridge et al., 2019).

### 3.2 Temporal Encoder

The second key component of TeMP is the temporal encoder, which seeks to integrate information *across* time in the entity representations. We investigate two approaches to compute entity representation $\boldsymbol{z}_{i,t}$ leveraging temporal information: a recurrent architecture (inspired by Jin et al. (2019)) and a self-attention approach (inspired by Sankar et al. (2020)).

**Temporal recurrence model (TeMP-GRU).** We propose to couple a traditional recurrence mechanism with weight decay, in order to account the diminishing effect of historical facts. Let $t^-$ denote the last time step at which entity $e_i$ was *active* before $t$, the down-weighted entity representation $\hat{\boldsymbol{z}}_{i,t^-}$ is defined as follows:

$$\hat{\boldsymbol{z}}_{i,t^-} = \gamma^z_{i,t^-} \boldsymbol{z}_{i,t^-} \tag{1}$$

$$\gamma^z_{i,t^-} = \exp\{-\max(0, \lambda_z |t - t^-| + b_z)\}, \tag{2}$$

where $\gamma^z$ denotes the decay rate with $\lambda_z$ and $b_z$ as learnable parameters. This design is inspired by Che et al. (2018) and ensures that $\gamma^z$ is monotonically decreasing with respect to the temporal difference and ranges from 0 to 1. We ensure that $\hat{\boldsymbol{z}}_{i,t^-}$ is only nonzero if $t^- \in \{t-\tau, .., t-1\}$, otherwise it will be assigned a zero vector. Finally, we use a gated recurrent unit (GRU) to obtain the entity embedding $\boldsymbol{z}_{i,t}$ based on $\hat{\boldsymbol{z}}_{i,t^-}$ and the static representation $\boldsymbol{x}_{i,t}$:

$$\boldsymbol{z}_{i,t} = \text{GRU}(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{z}}_{i,t^-}), \tag{3}$$

where GRU denotes the standard cell defined by Cho et al. (2014).

**Temporal self attention model (TeMP-SA).** Another way to incorporate historical information is to selectively attend to the sequence of active temporal entity representations. We use the following equations—inspired by the transformer architecture (Vaswani et al., 2017)—to perform attentive pooling over the entity embeddings $\boldsymbol{x}_{i,t'}$ at each time step $t' \in \{t-\tau, .., t\}$, in order to generate time-dependent embeddings $\boldsymbol{z}_{i,t}$:

$$\boldsymbol{q}_{ij} = \frac{(\boldsymbol{x}_{i,t}\boldsymbol{W}_q)(\boldsymbol{x}_{i,t-j}\boldsymbol{W}_k)^T}{\sqrt{d}} \tag{4}$$

$$\boldsymbol{e}_{ij} = \boldsymbol{q}_{ij} - \max(0, \lambda_z j + b_z) + \boldsymbol{M}_{ij} \tag{5}$$

$$\boldsymbol{\beta}_{ij} = \frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k=0}^{\tau} \exp(\boldsymbol{e}_{ik})} \tag{6}$$

$$\boldsymbol{z}_{i,t} = \sum_{j=0}^{\tau} \boldsymbol{\beta}_{ij} (\boldsymbol{x}_{i,t-j}\boldsymbol{W}_v), \tag{7}$$

where $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d\times d}$ denote linear projection matrices, as in a transformer layer (Vaswani et al., 2017), $\boldsymbol{\beta} \in \mathbb{R}^{|E|\times\tau}$ denotes the attention weight matrix obtained by multiplicative attention function and $\{\lambda_z, b_z\}$ denotes the learnable parameters of the down-weighting function. The $\boldsymbol{M} \in \mathbb{R}^{|E|\times\tau}$ matrix is a mask defined as

$$\boldsymbol{M}_{ij} = \begin{cases} 0, & \text{if } e_i \text{ is active at time } t-j, \\ -\infty, & \text{otherwise.} \end{cases} \tag{8}$$

As $\boldsymbol{M}_{ij} \to -\infty$, the attention weights $\boldsymbol{\beta}_{ij} \to 0$, which ensures that only active temporal entity representations are assigned non-zero weights. Finally, note that the full self-attention model can be generalized to use multiple attention heads, as in Vaswani et al. (2017).

**Incorporating future information**. Note that in the TKGC setting, we assume that the model has access to all the time steps during training. In particular, we assume there is missing data within each time step but that all the (incomplete) snapshots information $D^{(t)}$ are available during training. Thus, in both the attention and recurrence-based approaches, it is worthwhile to integrate temporal information from both the past and future. We do so by employing a bi-directional GRU in the recurrent approach, and by attending over both past and future time steps in the attention-based approach.

### 3.3 Tackling Temporal Heterogeneities

Although TeMP jointly models structural and temporal information, the encoder alone is insufficient to deal with the *temporal heterogeneity* in real-world TKGs, namely *sparsity* and *variability* of entity occurrences. We explore data imputation and frequency-based gating techniques to address these temporal heterogeneities. Because the degrees of temporal heterogeneities vary drastically across datasets (Appendix A.5), our proposed techniques are optional model variations that may im-

prove model performance depending on the dataset characteristics.

**Imputation of inactive entities**. Recall that structural encoder only encodes neighboring entities within the same KG snapshot. For entity $e_i$ that is inactive at time step $t$, the static representation $\boldsymbol{x_{i,t}}$ is hence not informed by any structural neighbors, resulting in stale representations shared across multiple time steps. We propose an imputation (IM) approach that integrates stale representations with temporal representations for inactive entities, i.e., $\hat{\boldsymbol{x}}_{i,t} = \text{IM}(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t-})$, where $\hat{\boldsymbol{x}}_{i,t}$ represents imputed structural representation.

Without loss of generality, we define the imputation for a uni-directional model and refer the bidirectional case to Appendix A.2. We defined IM to be the weighted sum function, with the similar exponential decay mechanism used in Equation (1):

$$\gamma^x_{i,t^-} = \exp\{-\max(0, \lambda_x|t - t^-| + b_x)\}. \quad (9)$$

The imputed representation is defined as follows:

$$\hat{\boldsymbol{x}}_{i,t} = \gamma^x_{i,t^-}\boldsymbol{x}_{i,t^-} + (1 - \gamma^x_{i,t^-})\boldsymbol{x}_{i,t}. \quad (10)$$

This model-agnostic approach is applicable by replacing $\boldsymbol{x}_{i,t}$ in the temporal models with $\hat{\boldsymbol{x}}_{i,t}$ .

**Frequency-based gating**. In addition to imputation, we also implement an approach to perform frequency-based gating (FG). The encoded representation of an entity is modulated depending on how many recent temporal facts it participates in. In particular, we propose to learn a gating term in order to fuse the embeddings $\boldsymbol{x}_{i,t}$ from output of the structural encoder (Section 3.1) with the temporal embeddings $\boldsymbol{z}_{i,t}$ (Section 3.2) in a frequency-dependent way. We differentiate the weights by the query types (subject or object query) and entity position (whether $e_i$ is subject or object in the queried fact) in order to contextualize the entities into their role within a quadruple.

In what follows, we use the term *pattern* to denote a non-empty subset of the quadruple $(s, r, o, t)$ (not containing time $t$). The *temporal frequency* of a pattern is defined as the number of facts with such pattern in the defined time window. Consider the quadruple *(Obama, visit, China, 2014)*, the temporal frequency of the pattern *(Obama, visit)* is the number of quadruples *(Obama, visit, $*$, $t'$)* with $t'$ in the time window (e.g., from 2000 to 2014).

We define the following *temporal pattern frequencies* (TPFs) associated with the quadruple $(s, r, o, t)$: (1) subject frequency $f^t_s$, (2) object frequency $f^t_o$, (3) relation frequency $f^t_r$, (4) subject-relation frequency $f^t_{s,r}$, (5) relation-object frequency $f^t_{r,o}$.

Without loss of generality, we define our gating mechanism from the perspective of object queries $(s, r, ?, t)$, where the goal is to predict the missing object in a quadruple. The definition for subject queries is analogous and detailed in Appendix A.3.

When answering the object query $(s, r, ?, t)$ the model has only the access to frequencies $F_s = [f^t_s, f^t_r, f^t_{s,r}]$. Thus, we use the frequency vector $F_s$ to define a gating term over the embeddings in the query:

$$\tilde{\boldsymbol{z}}_{s,t} = \alpha_{os}\boldsymbol{x}_{s,t} + (1 - \alpha_{os})\boldsymbol{z}_{s,t} \quad (11)$$

$$\tilde{\boldsymbol{z}}_{o,t} = \alpha_{oo}\boldsymbol{x}_{o,t} + (1 - \alpha_{oo})\boldsymbol{z}_{o,t}, \quad (12)$$

where $\alpha_{os} = \text{MLP}_{os}(F_s), \alpha_{oo} = \text{MLP}_{oo}(F_s)$ are weights in the range $[0, 1]$ learned via a two-layer dense neural network. Here the calculation for object embedding $\tilde{\boldsymbol{z}}_{o,t}$ covers all entities.

### 3.4 Decoder and Training

Let $\phi(.)$ denote the score for a tuple and let DEC denote any proper decoding function for static KGs, e.g., the TransE decoder (Bordes et al., 2013). The score for the quadruple is defined as follows:

$$\phi(s, r, o, t) = \text{DEC}(\tilde{\boldsymbol{z}}_{s,t}, \boldsymbol{z}_r, \tilde{\boldsymbol{z}}_{o,t}). \quad (13)$$

Here, $\tilde{\boldsymbol{z}}_{s,t}$ and $\tilde{\boldsymbol{z}}_{o,t}$ are the subject and object embeddings (as defined in Sections 3.1-3.3) while $\boldsymbol{z}_r$ is a learned embedding of the relation $r$. To train a model using this score function, the model parameters are learned using gradient-based optimization in mini-batches. For each triple $\eta = (s, r, o) \in D^{(t)}$, we sample a negative set of entities $D^-_\eta = \{o' | (s, r, o') \notin D^{(t)}\}$ and define the cross-entropy loss as follows:

$$L = -\sum_{t=1}^{T}\sum_{\eta \in D^{(t)}} \frac{\exp(\phi(s, r, o, t))}{\sum_{o' \in D^-_\eta}\exp(\phi(s, r, o', t))}.$$

Note that without loss of generality, we defined the above loss over object queries (as in Section 3.3), with an analogous loss and negative sampling used for subject queries defined in Appendix A.3.

## 4 Experiments

We evaluate the performances of TeMP models on three standard TKGC benchmark datasets and analyze the strengths and shortcomings when answering queries with different characteristics. Code to

Table 1: Temporal KG completion evaluation results on ICEWS, ICEWS05-15 and GDELT. The Hit@1, Hit@3, and Hit@10 metrics are multiplied by 100. Best results are in bold.

| Model | ICEWS14 | | | | ICEWS05-15 | | | | GDELT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | Hit@1 | Hit@3 | Hit@10 | MRR | Hit@1 | Hit@3 | Hit@10 | MRR | Hit@1 | Hit@3 | Hit@10 |
| TransE | 0.326 | 15.4 | 43.0 | 64.4 | 0.330 | 15.2 | 44.0 | 66.0 | 0.155 | 6.0 | 17.8 | 33.5 |
| DistMult | 0.441 | 32.5 | 49.8 | 66.8 | 0.457 | 33.8 | 51.5 | 69.1 | 0.210 | 13.3 | 22.4 | 36.5 |
| ComplEx | 0.442 | 40.0 | 43.0 | 66.4 | 0.464 | 34.7 | 52.4 | 69.6 | 0.213 | 13.3 | 22.5 | 36.6 |
| SimplE | 0.458 | 34.1 | 51.6 | 68.7 | 0.478 | 35.9 | 53.9 | 70.8 | 0.206 | 12.4 | 22.0 | 36.6 |
| TTransE | 0.255 | 7.4 | - | 60.1 | 0.271 | 8.4 | - | 61.6 | 0.115 | 0.0 | 16.0 | 31.8 |
| HyTE | 0.297 | 10.8 | 41.6 | 65.5 | 0.316 | 11.6 | 44.5 | 68.1 | 0.118 | 0.0 | 16.5 | 32.6 |
| TA-DistMult | 0.477 | 36.3 | - | 68.6 | 0.474 | 34.6 | - | 72.8 | 0.206 | 12.4 | 21.9 | 36.5 |
| DE-TransE | 0.326 | 12.4 | 46.7 | 68.6 | 0.314 | 10.8 | 45.3 | 68.5 | 0.126 | 0.0 | 18.1 | 35.0 |
| DE-DistMult | 0.501 | 39.2 | 56.9 | 70.8 | 0.484 | 36.6 | 54.6 | 71.8 | 0.213 | 13.0 | 22.8 | 37.6 |
| DE-SimplE | 0.526 | 41.8 | 59.2 | 72.5 | 0.513 | 39.2 | 57.8 | 74.8 | 0.230 | 14.1 | 24.8 | 40.3 |
| AtiSEE | 0.569 | 46.3 | 63.9 | 76.3 | 0.520 | 39.7 | 59.5 | 77.3 | - | - | - | - |
| AtiSER | 0.571 | 46.5 | 64.3 | 75.5 | 0.484 | 35.0 | 55.8 | 74.9 | - | - | - | - |
| TNTComplEx | **0.620** | **52.0** | 66.0 | 76.0 | 0.670 | **59.0** | 71.0 | 81.0 | - | - | - | - |
| TED | 0.441 | 35.3 | 49.1 | 60.8 | 0.503 | 40.8 | 56.1 | 68.4 | 0.237 | 14.9 | 26.3 | 40.7 |
| SRGCN | 0.604 | 48.3 | 68.0 | 83.0 | 0.662 | 53.5 | 74.7 | 89.9 | 0.239 | 15.7 | 25.6 | 39.8 |
| TeMP-GRU | 0.601 | 47.8 | 68.1 | 82.8 | **0.691** | 56.6 | **78.2** | **91.7** | **0.275** | **19.1** | **29.7** | **43.7** |
| TeMP-SA | 0.607 | 48.4 | **68.4** | **84.0** | 0.680 | 55.3 | 76.9 | 91.3 | 0.232 | 15.2 | 24.5 | 37.7 |

reproduce all our experiments is included in the submission and will be made publicly available.

### 4.1 Datasets

We evaluate our model on Global Database of Events, Language and Tone (GDELT) (Leetaru and Schrodt, 2013) and Integrated Crisis Early Warning System (ICEWS) (Boschee et al., 2015) datasets. For ICEWS, we use the two subsets generated by García-Durán et al. (2018): ICEWS14, corresponding to the facts in 2014 and ICEWS 05-15, containing all facts from 2005 to 2015. For GDELT, we use the subset provided by Trivedi et al. (2017) corresponding to facts from April 1, 2015 to March 31, 2016. We utilize the same partitioning of train, validation and test set as specified by Goel et al. (2019). More dataset statistics are summarized in Appendix A.5.

### 4.2 Evaluation Metrics

For each quadruple $(s, r, o, t)$ in the test set, we evaluate two queries $(s, r, ?, t)$ and $(?, r, o, t)$. For the first query we calculate scores for $(s', r, o, t), \forall s' \in E$ using Equation (13). Similar procedure applies to the second query. We then calculate the metrics based on the rank of $(s, r, o, t)$ in each query. Evaluation is performed under filtered settings defined by Bordes et al. (2013). We report the Hits@1,@3, @10 scores and MRR (mean reciprocal rank). Please see Appendix A.6 for detailed definitions.

### 4.3 Baseline Methods

We compare TeMP against a broad spectrum of existing approaches, including a novel rule-based baseline, static embedding methods, and existing state-of-the-art approaches for TKGC.

**TED model**. We propose a rule-based baseline by directly copying facts from quadruples in the recent past and future, denoted as temporal exponential decay (TED) model. The basic idea in this approach is that we predict missing facts by simply copying facts from nearby time steps. The probability of copying each fact is dependent on (1) number of elements overlapping with the queried quadruple and (2) temporal distance to the current time step. For a detailed description of this baseline, please refer to Appendix A.4.

**Static KGC methods**. We include TransE (Nguyen et al., 2016), DistMult (Yang et al., 2014), ComplEx (Trouillon et al., 2016) and SimplE (Kazemi and Poole, 2018) in the realm of static KG embedding methods. We also include a Static RGCN baseline (denoted as SRGCN), which implements the RGCN message-passing approach proposed by Schlichtkrull et al. (2018). Note that all

these static baseline methods are employed without considering the time information in the input.

**Temporal KGC methods**. We also compare with state-of-the arts models designed for TKGC including TTransE (Leblay and Chekol, 2018), TADistMult (García-Durán et al., 2018), HyTe (Dasgupta et al., 2018), Diachronic Embedding (DE) (Goel et al., 2019), AtisEE, AtisER (Xu et al., 2019) and TNTComplEx (Lacroix et al., 2020). We don't compare with RE-NET, GHN (Han et al., 2020), DartNet (Garg et al., 2020) and Know-Evolve since these work focus on graph extrapolation task.

### 4.4 Implementation and Hyperparameters

All the models except TED are implemented in PyTorch, making use of the PyTorch lightning module and the Deep Graph Library (Wang et al., 2019). We set the negative sampling ratio to 500, i.e. 500 negative samples per positive triple. Because we corrupt subjects and objects separately, there are in total 1000 negative samples collected to estimate the probability of a factual triple. For full details on all the model hyperparameters for TeMP and the baselines, refer to Appendix A.7.

### 4.5 Results and Analysis

#### 4.5.1 Comparative Study

We compare the baseline models with two instantiations of the TeMP framework: *TeMP-GRU*, *TeMP-SA*, corresponding to the GRU and self-attention variants discussed in Section 3.2. Incorporating imputation or frequency-based gating is treated optional and we explore different model variants in Section 4.5.2. Results on each dataset are given by the model variant that achieves the best validation set performance. The core experimental results are summarized in Table 1.

**TeMP achieves a new state of the art**. We find that TeMP-SA and TeMP-GRU achieve state-of-the-art results on all three datasets in terms of Hits@10. Compared to the most recent work (Lacroix et al., 2020)—which achieves the best performance to-date on the ICEWS datasets—our results are 8.0% and 10.7% higher on the Hits@10 evaluation, though they are slightly worse on Hits@1. Additionally, our model achieves a 3.7% improvement on GDELT compared with DE, the prior state-of-the-art on that dataset. The results of the AtiSEE and TNTComplEx methods on the GDELT dataset are not available.

**Strong baseline performance**. Interestingly, we find that two of our proposed baseline models also achieve surprisingly strong performance, even outperforming the prior state of the art in some settings. For example, our rule-based TED baseline achieves relatively strong performance on all three datasets, in particular on GDELT, where it is better than all existing neural models by all measures. This highlights the power of simply copying temporal facts with the same patterns as the queried quadruples. Similarly, our static RGCN baseline (*SRGCN*) also achieves very strong performance, with the next-best Hits@10 results behind the TeMP framework. We hypothesize that the message-passing procedure in SRGCN allows the model to leverage multi-hop structural information that is specific to each time-step, enabling strong performance.

#### 4.5.2 Exploration of Model Variations

We study the effect of the imputation and frequency-based gating approaches proposed in Section 3.3 by running model variants on three datasets. We highlight the performance comparison as well as the implication of dataset characteristics on the performance variations.

Our results are reported on the corresponding validation sets of these benchmarks. The results regarding the incorporation of imputation (IM) and frequency-based gating (FG) are shown in Table 2. We use a ✓ to indicate a certain component being used in the experiment, and blank for the absence of the corresponding component. [4]

**ICEWS14**. On the ICEWS14 dataset, we find that combining both TeMP-GRU and TeMP-SA models with both imputation and gating achieves the best results on validation set (3.3% improvement). Additionally, each individual component helps improve the overall model performance by about 1%.

**ICEWS05-15**. On ICEWS05-15, models with gating improved the performance by more than 1% compared to those without gating. However, the additional incorporation of imputation does not result in improvement in the results.

**GDELT**. As for GDELT dataset, we find neither imputation nor gating is significant for model performance. However, it is evident from dataset characteristics that GDELT does not exhibit the same temporal variability and sparsity as the ICEWS datasets. Discussion in Appendix A.5 shows that all entities are active at every time step in GDELT

[4] Imputation is an intrinsic part of TeMP-SA thus it is used in all experiments. See Appendix A.2 for details.

Table 2: MRR results for different model variations on ICEWS14, ICEWS05-15 and GDELT

| Model | IM | FG | ICEWS14 | ICEWS05-15 | GDELT |
|---|---|---|---|---|---|
| TeMP-GRU | ✓ | ✓ | **0.610** | 0.680 | 0.269 |
| TeMP-GRU | | ✓ | 0.599 | **0.689** | 0.270 |
| TeMP-GRU | ✓ | | 0.593 | 0.670 | **0.275** |
| TeMP-GRU | | | 0.577 | 0.673 | 0.274 |
| TeMP-SA | ✓ | ✓ | **0.623** | **0.676** | 0.233 |
| TeMP-SA | ✓ | | 0.619 | 0.670 | **0.235** |

(unlike the ICEWS datasets). Additionally, on average each active entity has roughly 150 reference temporal facts in the last 15 time steps, suggesting that each entity involved in TKGC queries are sufficiently informed by the nearby KG snapshots. Data imputation and gating methods are thus unnecessary complexities in GDELT.

#### 4.5.3 Fine-grained Error Analysis

To assess how models perform on TKGC queries with different temporal pattern frequencies (TPFs; see Section 3.3), we group queried quadruples based on different TPFs and calculate the Hits@10 metrics in each group.

We plot the temporal subject-relation frequency $f^t_{s,r}$ (defined in Section 3.3) versus the model performances on subject and object queries to study the *replication* and *reference* effects of temporal facts, respectively. Here, we use the term *replication effect* to denote the situation where the model can make predictions by copying the exact correct answer to a query from temporal facts. For example, copying *China* from *(Biden, visit, China, 2013)* to answer the query *(Obama, visit, ?, 2014)*. We use the term *reference effect* to denote the effect of having facts that are related (but do not not contain answer entity) to the query fact in the temporal context. For example, selecting *China* from a set countries where Obama visited in the year 2013.

We compare the performances of static models (DE and SRGCN) and temporal models (TeMP-GRU models) on different TPFs. *TeMP-GRU-Vanilla* represents the vanilla version of the model and *TeMP-GRU-Gating* refers to TeMP-GRU model combined with gating technique. Detailed analysis regarding TKGC performance versus other TPFs are discussed in Appendix A.8.

**Replication effect analysis** Here, we examine how the subject-relation TPF correlates with model performance on subject queries. Figure 4 illustrates that temporal models exhibit positive correlation between subject-relation TPF and subject query performance, while static models show relatively negative correlation between the two quantities. This suggests that the replication effect is stronger in TeMP, indicating that the TeMP model is better at utilizing temporal information for TKGC queries. Additionally, gating helps improve over the vanilla version by a slight margin on all subject-relation frequency values. On the other hand, SRGCN achieves better performance on low-TPF queries than temporal models. However, coupling the TeMP model with gating helps close the gap, sometimes surpassing SRGCN on such queries.

**Reference effect analysis**. Here, we examine how the occurrence of related facts (not containing the answer) in the temporal context impacts performance. We find that the temporal models exhibit non-linear correlations between object query performance and subject-relation TPF (Figure 5). In particular, on the ICEWS datasets the performance increases as the log-frequencies grows from $-\infty$ to 2 and drops at higher frequency values. We hypothesize that it is harder for temporal model to select the answer from a very large set of object candidates, e.g,. choosing *China* from more than 100 countries that Obama visited from 2008 to 2013. In terms of model comparisons, we find that gating helps TeMP-GRU to surpass its vanilla version and SRGCN on most TPF values. The margin of improvement is especially significant on queries of high TPF in ICEWS05-15.

The null effect of frequency-based gating on GDELT can be attributed to the same reason as discussed in Section 4.5.2.

## 5 Conclusion

In this work, we present a novel framework named TeMP for temporal knowledge graph completion (TKGC). TeMP computes entity representation by jointly modelling multi-hop structural information and temporal facts from nearby time-steps.

Additionally, we introduce novel frequency-based gating and data imputation techniques to address the temporal variability and sparsity problems in TKGC. We show that our model is able to achieve superior performance (10.7% relative improvement) over the state-of-the-arts on three benchmark datasets. Our work is potentially beneficial to other tasks such as temporal information extraction and temporal question answering, by providing beliefs about the likelihood of facts at particular points in time.

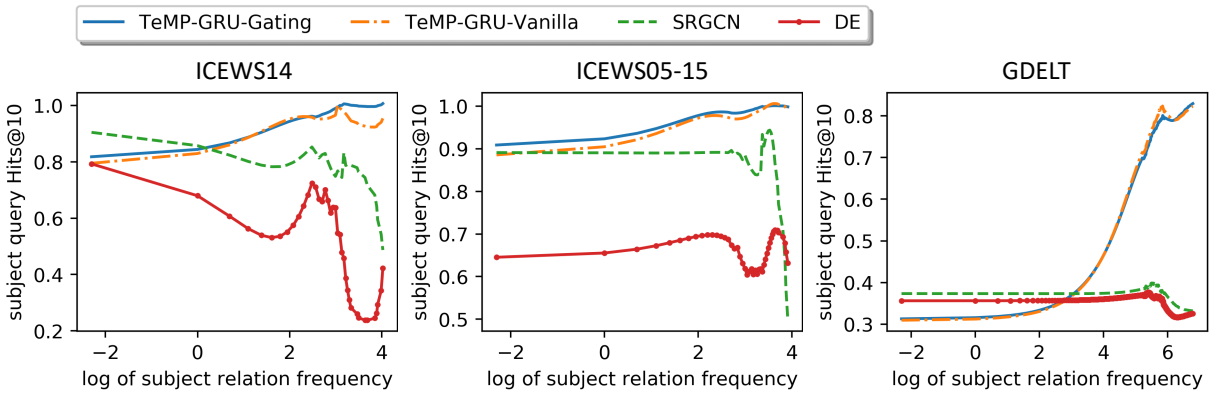



Figure 4: Subject query hit@10 performance comparison of TeMP with different variations and baseline methods.

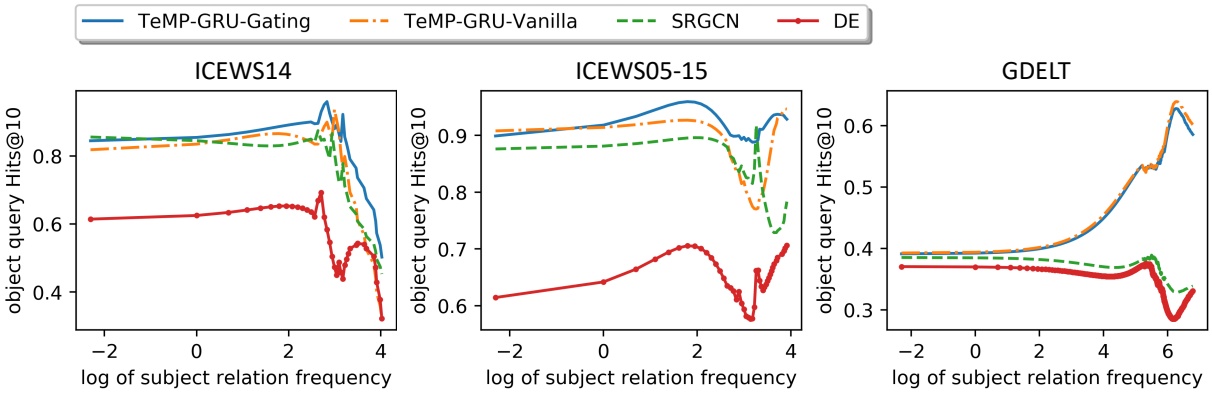



Figure 5: Object query hit@10 performance comparison of TeMP with different variations and baseline methods.

Future work involves exploring the generalization of TeMP to continuous TKGC and better imputation techniques to induce representations for infrequent and inactive entities.

## Acknowledgement

This research is supported by CIFAR Canada AI Chair program and Samsung Electronics. The authors would like to thank Compute Canada for providing the computational resources.

## References

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*, pages 2787–2795.

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

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. 2019. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12.

Jinyin Chen, Xuanheng Xu, Yangyang Wu, and Haibin Zheng. 2018. Gc-lstm: Graph convolution embedded lstm for dynamic link prediction. *arXiv preprint arXiv:1812.04206*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

Sam De Winter, Tim Decuypere, Sandra Mitrović, Bart Baesens, and Jochen De Weerdt. 2018. Combining temporal aspects of dynamic networks with node2vec for a more efficient dynamic link prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1234–1241. IEEE.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. 2018. Learning sequence encoders

for temporal knowledge graph completion. *arXiv preprint arXiv:1809.03202*.

Sankalp Garg, Navodita Sharma, Woojeong Jin, and Xiang Ren. 2020. Temporal attribute prediction via joint modeling of multi-relational structure evolution. *arXiv preprint arXiv:2003.03919*.

Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. 2019. Diachronic embedding for temporal knowledge graph completion. *arXiv preprint arXiv:1907.03143*.

Ehsan Hajiramezanali, Arman Hasanzadeh, Nick Duffield, Krishna R Narayanan, Mingyuan Zhou, and Xiaoning Qian. 2019. Variational graph recurrent neural networks. *arXiv preprint arXiv:1908.09710*.

W. Hamilton, R. Ying, and J. Leskovec. 2017. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*.

Zhen Han, Yuyi Wang, Yunpu Ma, Stephan Guünnemann, and Volker Tresp. 2020. The graph hawkes network for reasoning on temporal knowledge graphs. *arXiv preprint arXiv:2003.13432*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1807–1810. ACM.

Tingsong Jiang, Tianyu Liu, Tao Ge, Lei Sha, Baobao Chang, Sujian Li, and Zhifang Sui. 2016. Towards time-aware knowledge graph completion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1715–1724.

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

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in neural information processing systems*, pages 4284–4295.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2018. Learning dynamic embeddings from temporal interactions. *arXiv preprint arXiv:1812.02289*.

Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. 2020. Tensor decompositions for temporal knowledge base completion. *arXiv preprint arXiv:2004.04926*.

Julien Leblay and Melisachew Wudage Chekol. 2018. Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, pages 1771–1776. International World Wide Web Conferences Steering Committee.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone. In *ISA Annual Convention*. Citeseer.

Franco Manessi, Alessandro Rozza, and Mario Manzo. 2020. Dynamic graph convolutional networks. *Pattern Recognition*, 97:107000.

Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. 2016. Stranse: a novel embedding model of entities and relationships in knowledge bases. *arXiv preprint arXiv:1606.08140*.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Thirtieth Aaai conference on artificial intelligence*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816.

Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, and Charles E Leisersen. 2019. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. *arXiv preprint arXiv:1902.10191*.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2019. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*.

Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. 2020. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 519–527.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. 2017. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3462–3471. JMLR. org.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2019. Dyrep: Learning representations over dynamic graphs. In *International Conference on Learning Representations*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. 2019. Deep graph library: Towards efficient and scalable deep learning on graphs. *arXiv preprint arXiv:1909.01315*.

Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Jens Lehmann, and Hamed Shariat Yazdi. 2019. Temporal knowledge graph embedding model based on additive time series decomposition. *arXiv preprint arXiv:1911.07893*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Lekui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic network embedding by modeling triadic closure process. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

# A Appendix

## A.1 Architecture Details

**Temporal Edge Dropout**.

The replication effect illustrated in Figure 4 and 8 shows that TeMP is increasingly better capable at copying from temporal facts when TPFs also increase. We refer to this as "overfitting" to the temporal facts. In order to alleviate such problem, we propose temporal edge dropout: randomly dropping facts occurred in the defined time window used to induce the entity representation.

Rong et al. (2019) propose dropping a proportion in the local graph context to combat over-fitting and over-smoothing. We extend this technique to TKG by either (1) randomly dropping a certain percentage of quadruples in each temporal snapshot and (2) drop quadruples with different probabilities based on certain quadruple characteristics. Details of the second method is omitted since we find the two methods working equally well. We use 0.2 as temporal edge dropout rate in all experiments.

**Positional Embedding**. We capture the time-sensitive information in the TKG by combining the entity representation with positional embedding. The positional embedding is denoted as $\{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_T\}$, which embeds absolute positional information of each time step. The set of representations for entity $e_i$ at all time steps is $\{\boldsymbol{p}_1 + \boldsymbol{z}_{i,1}, \boldsymbol{p}_2 + \boldsymbol{z}_{i,2}, ..., \boldsymbol{p}_T + \boldsymbol{z}_{i,T}\}$, which are used as input entity representation to the decoding function.

## A.2 Extended Imputation Formulation

For bidirectional temporal recurrent model, we defined the imputed representation analogous to Equation (9) and 10. We use $t^+$ to denote the very next time step at which entity $e_i$ is active after $t$. The decay rate for imputing from future representations as follows:

$$\gamma^x_{i,t^+} = \exp\{-\max(0, \lambda_x|t - t^+| + b_x)\}.$$

To calculate the imputed representation of the $e_i$ at time $t$, we divide both exponential decay rates by two and renormalize:

$$\gamma^x_{i,t} = 1 - \frac{\gamma^x_{i,t^-}}{2} - \frac{\gamma^x_{i,t^+}}{2}$$
$$\hat{\boldsymbol{x}}_{i,t} = \frac{\gamma^x_{i,t^-}}{2}\boldsymbol{x}_{i,t^-} + \frac{\gamma^x_{i,t^+}}{2}\boldsymbol{x}_{i,t^+} + \gamma^x_{i,t}\boldsymbol{x}_{i,t}\ .$$

**Intrinsic imputation for TeMP-SA**. We use Equation (4) - (7) to derive entity representations for

both active and inactive entities and view it as an intrinsic way of imputation. Hence imputation is tagged with all TeMP-SA results in Table 2.

### A.3 Analogous Definition of Frequency Based Gating and Training Loss

We define the process for deriving entity representation for subject queries $(?, r, o, t)$ analogous to Equation (11) and (12). The model is only allowed the access to frequencies $F_o = [f_o^t, f_r^t, f_{o,r}^t]$, we use it to define a similar gating over static and temporal entity representations:

$$\boldsymbol{z}_{s,t} \coloneqq \alpha_{ss}\boldsymbol{x}_{s,t} + (1-\alpha_{ss})\boldsymbol{z}_{s,t}$$
$$\boldsymbol{z}_{o,t} \coloneqq \alpha_{so}\boldsymbol{x}_{o,t} + (1-\alpha_{so})\boldsymbol{z}_{o,t},$$

where $\alpha_{ss} = \text{MLP}_{ss}(F_o)$, $\alpha_{so} = \text{MLP}_{ss}(F_o)$, $\alpha_{ss}, \alpha_{so} \in [0,1]$. With the negative subject entity set being $D^-_{\eta,s} = \{s' | (s', r, o) \notin D^{(t)}\}$, the training loss for subject queries is defined as follows:

$$L_{sub} = -\sum_{t=1}^{T} \sum_{\eta \in D^{(t)}} \frac{\exp(\phi(s,r,o,t))}{\sum_{s' \in D^-_{\eta,s}} \exp(\phi(s',r,o,t))}.$$

The final training loss is the sum of losses for two types of queries: $L = L_{sub} + L_{obj}$.

### A.4 Detailed TED Formulation and Analysis

**TED Model Definition**. We hypothesize that certain quadruples with *more frequent* occurrence in *more recent* time steps are informative for the current-step KGC. For each query, we construct a set of reference entities from training data. Similar to the down-weighting mechanism of temporal encoder (Section 3.2), we score each entity based on exponential decaying mechanism with respect to the temporal distance to the current time step. We then rank the entities in the reference set according to such scores.

For each queried quadruple $(s, r, o, t)$, we collect reference entity sets consisting of tuples $\{(e, t'), t' \neq t\}$ where $e$ is the subject or object entity and $t'$ is the corresponding time of occurrence. The tuples are extracted from the temporal facts sharing at least one element with $(s, r, o, t)$. We divide them into subject and object reference sets two types of queries. The subject reference set consists of:

(1) subjects with shared relation-object pair, i.e., $\{(s', t') | \exists t' \neq t, (s', r, o) \in D^{(t')}_{train}\}$,

(2) subjects with shared object, i.e., $\{(s', t') | \exists t' \neq t \wedge r' \in R, (s', r', o) \in D^{(t')}_{train}\}$,

(3) subjects with shared relation, i.e., $\{(s', t') | \exists t' \neq t \wedge o' \in E, (s', r, o') \in D^{(t')}_{train}\}$.

Symmetrically, object reference set consists of:

(1) objects with shared subject-relation pair, i.e., $\{(o', t') | \exists t' \neq t, (s, r, o') \in D^{(t')}_{train}\}$,

(2) objects with shared subject, i.e., $\{(o', t') | \exists t' \neq t \wedge r' \in R, (s, r', o') \in D^{(t')}_{train}\}$,

(3) objects with shared relation, i.e., $\{(o', t') | \exists t' \neq t \wedge s' \in E, (s', r, o') \in D^{(t')}_{train}\}$.

We don't collect triples in the current time step $t$ as we assume $D^{(t)}_{train} \cap D^{(t)}_{test} = \emptyset, \forall t$.

Note that (1) is a subset of (2) and (3), also (2) and (3) contain overlapping tuples. We define the priority to be (1) > (2) > (3), such that if some tuple is present in (1), then it will be removed from both (2) and (3). This is based on the assumption that objects with the same subject-relation pair as the current triple are the most ideal candidates. For example, because of the characteristics of *police*, the fact *(police, arrest, citizen)* occurred multiple times across in the dataset. Objects with same shared subject and different relation comes second, e.g. *(Obama, visit, China, 2013)*, *(Obama, visit, Russia, 2014)* are important information for predicting *(Obama, make_announcement_to, ?, 2015)*.

Let $S$ be some set of tuple defined above. The score for $e$ is the sum over all tuples containing $e$,

$$\sum_{t', (e,t') \in S} \exp(-\sigma|t - t'|), \sigma > 0. \quad (14)$$

**TED Results and Analysis**. Table 3 shows the sensitivity analysis for parameter $\sigma$ on validation set. We notice that the performances are low when $\sigma$ is either extremely large or small, while peaks when $\sigma = 0.1$ on ICEWS datasets and $\sigma = 1$ on GDELT dataset. This suggests an existing trade-off between *recency* and *frequency* heuristics.

TED model results also expose the bias of recurring events in political event datasets, particularly in GDELT. However, TED should be considered by future work as an important baseline to gauge the

relative model performance. Additionally, the results suggests the potential for pointer-style TKGC – deciding between coping an entity from historical facts and selecting an entity in the current snapshot to answer a query.

### A.5 Dataset Statistics and Characteristics

The dataset statistics are summarized in Table 4. The numbers of entities are 7,128, 10,488 and 500 respectively in three datasets, indicating that temporal sparsity issue is severe on ICEWS datasets but trivial on GDELT dataset. The temporal variability of three datasets is demonstrated in Figure 7. The average number of associated temporal facts for each entity is much lower in ICEWS datasets compared to GDELT. The difference can be attributed to the fact that GDELT dataset is constructed by extracting facts among the most frequent 500 entities in the entire dataset. This intrinsically eliminates the sparsity and variability bias in the original datasets.

### A.6 Definitions for Evaluation Metrics

We use MRR, Hits@1, Hits@3 and Hits@10 to evaluate the model performance. MRR is defined as:

$$\frac{1}{2 * |D_{test}|} \sum_{t=1}^{T} \sum_{\eta=(s,r,o)\in D_{test}^{(t)}} ( \frac{1}{\text{rank}(o|s,r,t)} + \frac{1}{\text{rank}(s|r,o,t)}) \quad (15)$$

The Hit@1, Hit@3, Hit@10 are the percentages of test facts for which the $k$ highest ranked predictions contain the correct prediction, $k = 1, 3, 10$. That is,

$$\frac{1}{2 * |D_{test}|} \sum_{t=1}^{T} \sum_{\eta=(s,r,o)\in D_{test}^{(t)}} (I(\text{rank}(o|s,r,t) \leq k) + I(\text{rank}(s|r,o,t) \leq k)) \quad (16)$$

where $k = 1, 3, 10$, $I$ is the indicator function.

### A.7 Detailed Implementation and Hyperparameters

We use the Adam optimizer and set the learning rate to 0.001. The batch size is set to 8 for ICEWS14 and ICEWS05-15, i.e. each batch contains facts in 8 snapshots. We additionally sample 3,000 quadruples in each snapshot to avoid out-of-memory issue. Embedding size and hidden sizes for both recurrent and self-attentive models are both set to 128. We use 8 attention heads in TeMP-SA to model the multi-faced evolution of TKG. As required by reproducibility checklist, the complete hyperparameter setting and run-time information for TeMP-GRU model on all benchmark datasets are summarized in Table 5.

Suggested by ablation study in (Jin et al., 2019) we set the number of relational convolution layers to 2 to encode two-hop neighbors. We apply temporal edge dropout technique to TKG, in each training epoch we randomly drop 50% of the quadruples in current KG and 20% triples in each temporal reference KG to combat over-fitting and over-smoothing. We experimented with TransE, DistMult and ComplEx on validation set and found that ComplEx (Trouillon et al., 2016) yields the best performance overall. Hence ComplEx is used as decoding function to score head or tail entities given queries. During inference on $D_{valid}^{(t)}$ and $D_{test}^{(t)}$, our models take $D_{train}^{(t-\tau)}, ..., D_{train}^{(t)}$ as input and compute the scores to compute the entity representations.

The parameter $\tau$ stands for the number of KG snapshots available for answering query. This is applied to temporal models as a *budget*. Single-direction models take temporal entity embedding from the past $\tau$ graphs while bidirectional models focus on $\frac{\tau}{2}$ historical and future snapshots.

We use early stopping with patience 10 with respect to the average MRR on the validation set. All ablation studies are conducted on the validation set. For the best model variants, we use the model checkpoint that achieves the best MRR score on validation set to perform final evaluation on test set.

### A.8 Detailed Analysis of Performances versus TPFs

We studied the correlation between subject-relation TPF and query answering performances in Section 4.5.3. Here, we first define a complete set of TPFs that covers all possible subsets of a quadruple. In Section 3.3 we defined (1) subject frequency $f_s^t$, (2) object frequency $f_o^t$, (3) relation frequency $f_r^t$, (4) subject-relation frequency $f_{s,r}^t$, (5) relation-object frequency $f_{r,o}^t$ related to quadruple $(s, r, o, t)$. We additionally define (6) subject-object frequency $f_{s,o}^t$ and (7) triple frequency $f_{s,r,o}^t$. We use the following combinations of TPFs and query types to study replication and reference effects respectively.

| $\sigma$ | ICEWS14 | | | | ICEWS05-15 | | | | GDELT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | Hit@1 | Hit@3 | Hit@10 | MRR | Hit@1 | Hit@3 | Hit@10 | MRR | Hit@1 | Hit@3 | Hit@10 |
| $10^{-5}$ | 0.434 | 36.2 | 48.9 | 60.5 | 0.466 | 36.2 | 52.4 | 66.6 | 0.179 | 10.1 | 18.7 | 33.1 |
| $10^{-2}$ | 0.445 | 35.5 | 49.9 | 61.2 | 0.498 | 39.7 | 55.9 | **68.8** | 0.192 | 11.0 | 20.3 | 35.6 |
| $10^{-1}$ | **0.455** | **36.7** | **50.7** | **61.6** | **0.505** | **40.8** | **56.3** | 68.7 | 0.226 | 13.7 | 24.7 | 40.4 |
| 1 | 0.449 | 35.9 | 50.3 | 61.4 | 0.500 | 40.1 | 55.9 | 68.5 | **0.238** | **15.0** | **26.3** | **40.8** |
| $10^{1}$ | 0.449 | 35.9 | 50.3 | 61.5 | 0.496 | 39.8 | 55.4 | 68.1 | 0.237 | 14.9 | 26.2 | 40.7 |
| $10^{2}$ | 0.446 | 35.5 | 50.0 | 61.2 | 0.482 | 38.3 | 53.9 | 66.8 | 0.232 | 14.4 | 25.8 | 40.2 |
| $10^{5}$ | 0.359 | 24.9 | 41.7 | 57.2 | 0.362 | 23.8 | 42.8 | 60.6 | 0.091 | 3.0 | 8.0 | 20.2 |

Table 3: TKGC evaluation results(filtered setting) using TED model under various $\sigma$ values. The Hit@1, Hit@3, and Hit@10 metrics are multiplied by 100.

| Dataset | # entities | # relations | # time steps | N˙train | N˙valid | N˙test | N˙total |
|---|---|---|---|---|---|---|---|
| ICEWS14 | 7,128 | 230 | 365 | 72,826 | 8,941 | 8,963 | 90,730 |
| ICEWS05-15 | 10,488 | 251 | 4017 | 386,962 | 46,275 | 46,092 | 479,329 |
| GDELT | 500 | 20 | 366 | 2,735,685 | 341,961 | 341,961 | 3,419,607 |

Table 4: Statistics of ICEWS14, ICEWS05-15 and GDELT datasets.

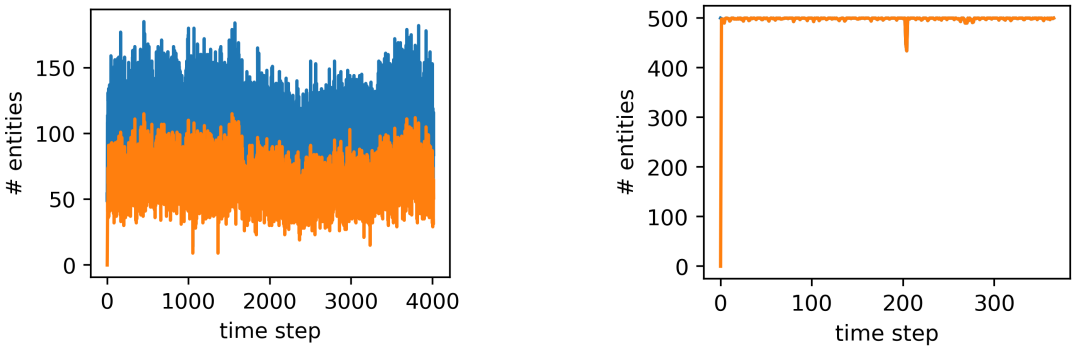



Figure 6: Dataset statistics of ICEWS05-15 (left) and GDELT (right) as a supplement of Figure 2.

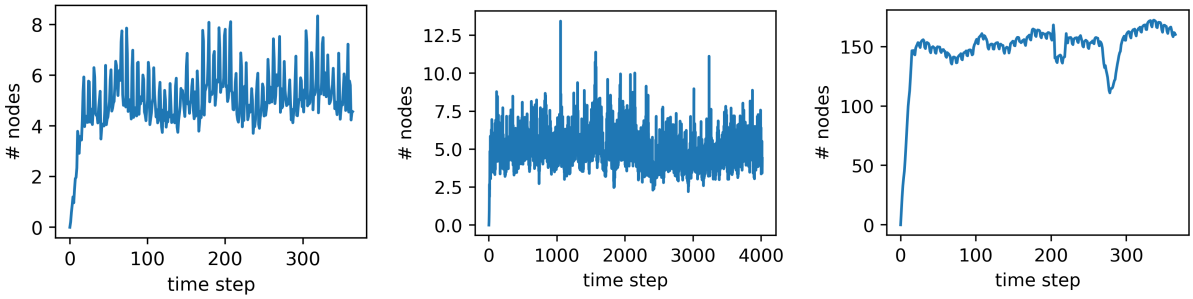
Table 5: Hyperparameters setting for TeMP-GRU model on three benchmark datasets

| Dataset | batch size | # temporal snapshots | GPU type | # GPU | Time limit | runtime per epoch | # parameters |
|---|---|---|---|---|---|---|---|
| ICEWS14 | 8 | 15 | GeForce GTX TiTan | 1 | 24h | 8m | 885K |
| ICEWS05-15 | 8 | 10 | Nvidia V100 | 1 | 60h | 70m | 2856K |
| GDELT | 4 | 15 | Nvidia V100 | 2 | 60h | 13m | 878K |

For replication effect, we compare subject query results against (1), then compare object query results against (2) and (5). Values of (6) and (7) are compared with the results of both subject and object queries. For reference effect, we compare object query results against (1), and subject query results against (2) and (5). Results are summarized in Figure 8 and Figure 9 respectively.

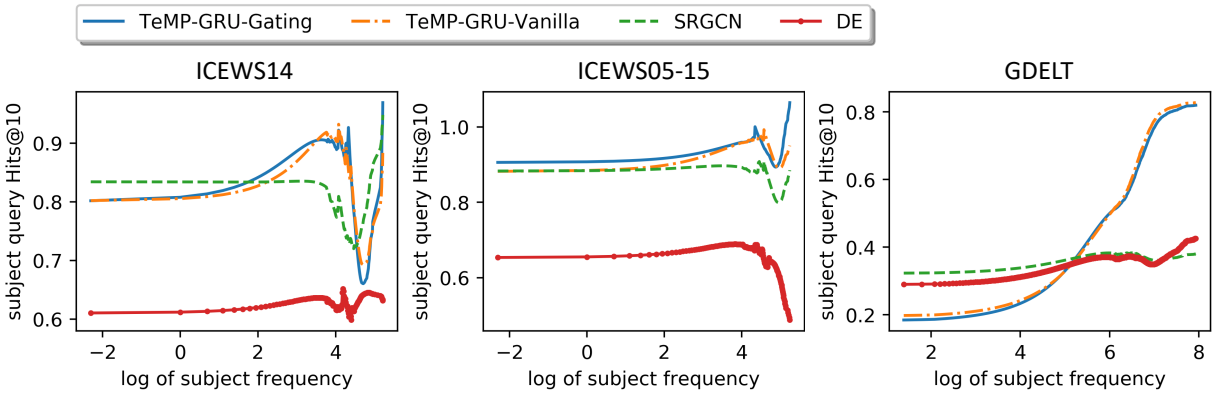
The general observation is similar to the discussion in Section 4.5.3. In the replication analysis , TeMP-GRU models show significantly more positive trends than the static models (SRGCN and DE). However, we witness drops in performances when TPFs become large in the reference effect analysis. Performance of TeMP-GRU-Vanilla model improves with the help of gating on ICEWS datasets on TPFs. The benefit is less obvious on GDELT dataset due to the observation that GDELT is less affected by temporal sparsity and variability problem (Appendix A.5).

We conclude that TeMP models are significant more advantageous in utilizing temporal facts for TKGC task. In addition, frequency-based gating improves the overall performance with respect to all different TFPs.

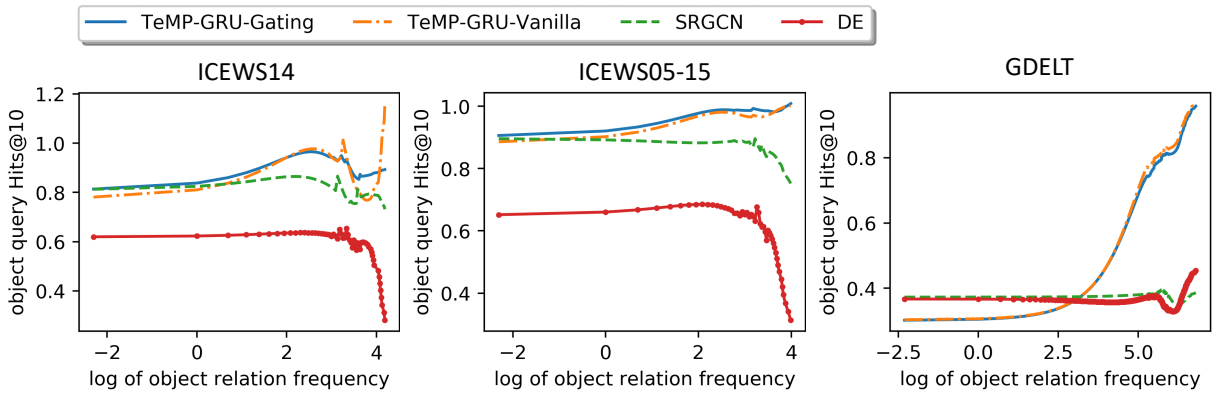



(a) (b) (c)

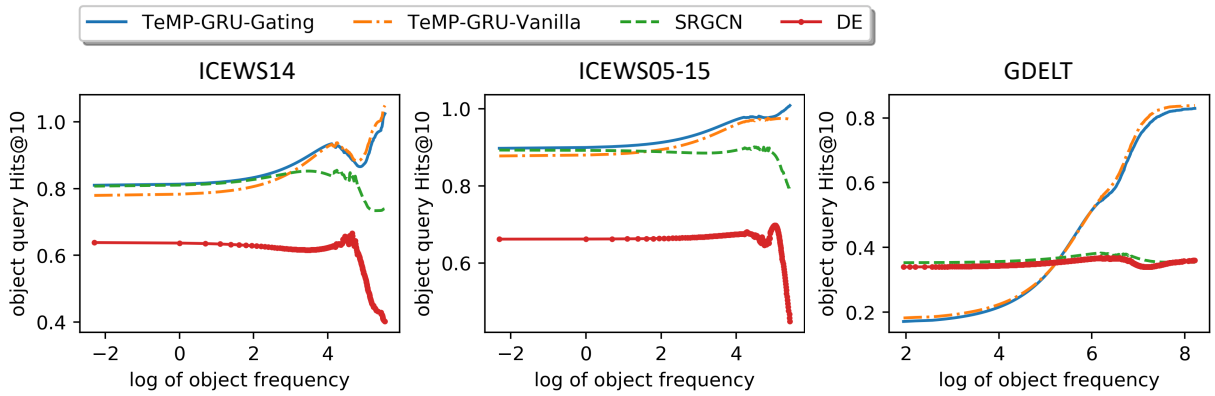
Figure 7: At each time step, for every active entity we calculate how many times each active entity occurred in that last 15 time steps and take average. We show the distirbution of such quantities on (a)ICEWS14, (b) ICEWS05-15 and (c) GDELT



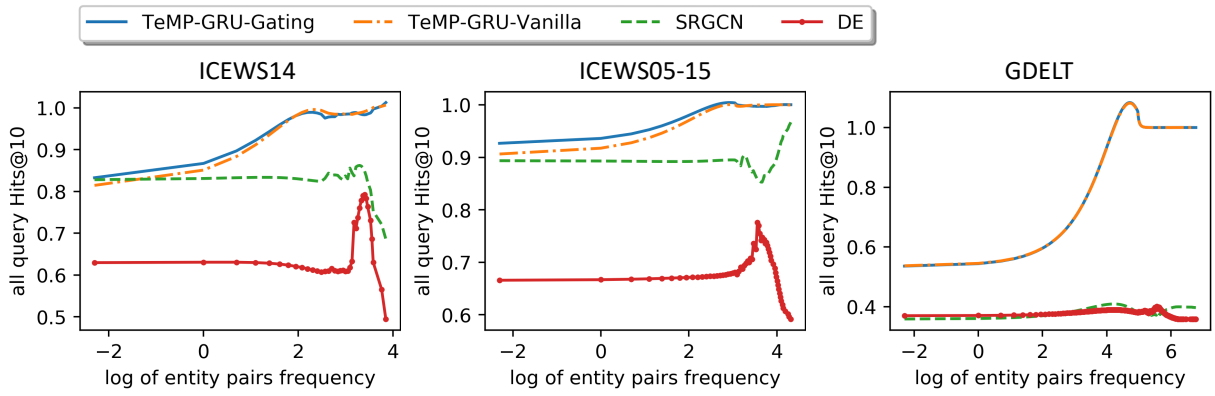

(a) Subject query Hits@10 performances versus temporal subject frequencies



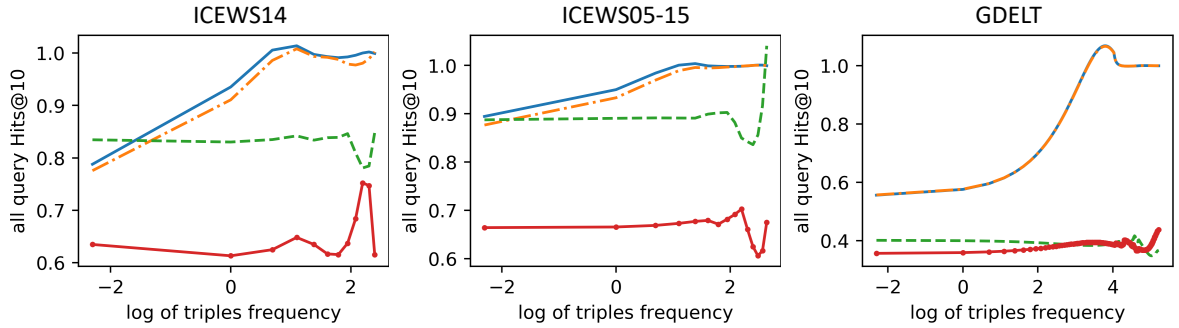

(b) Object query Hits@10 performances versus temporal object-relation frequencies




(c) Object query Hits@10 performances versus temporal object frequencies

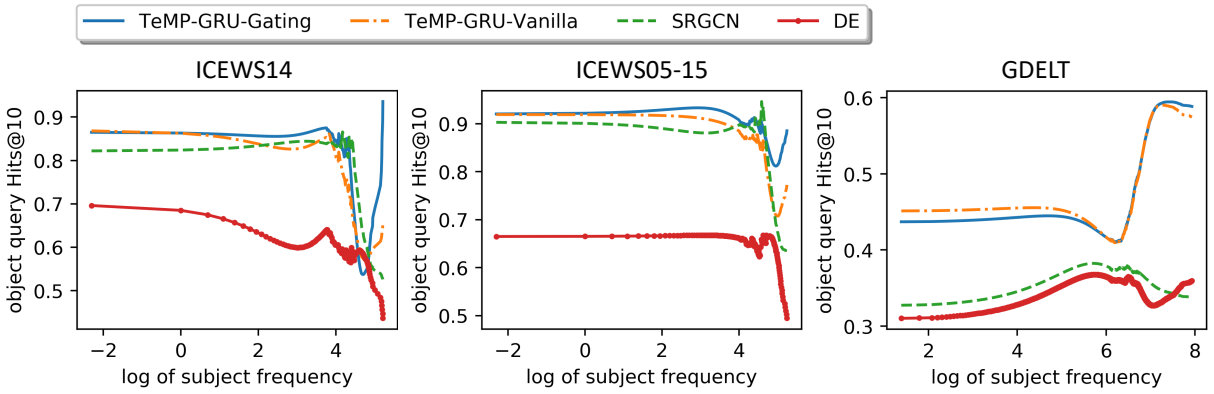



(d) All query Hits@10 performances versus temporal entity pair frequencies

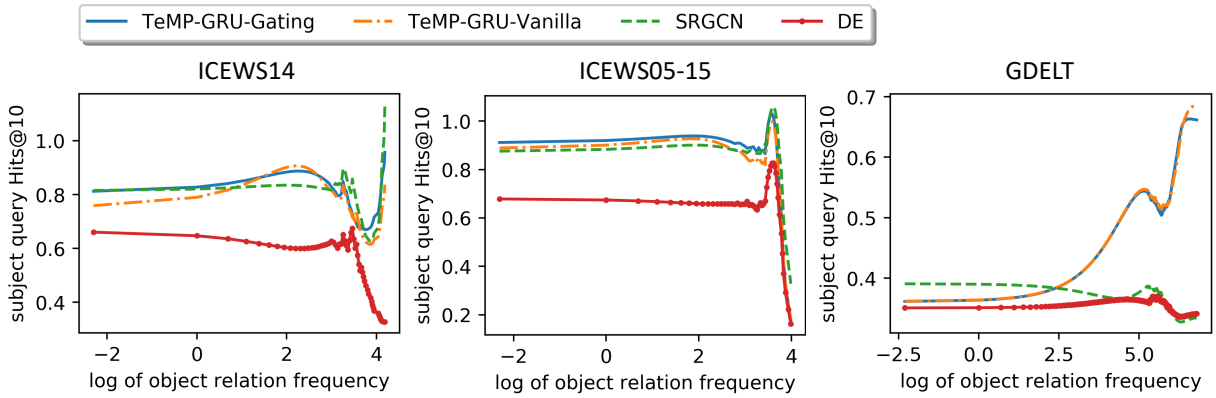



(e) All query Hits@10 performances versus temporal triple frequencies

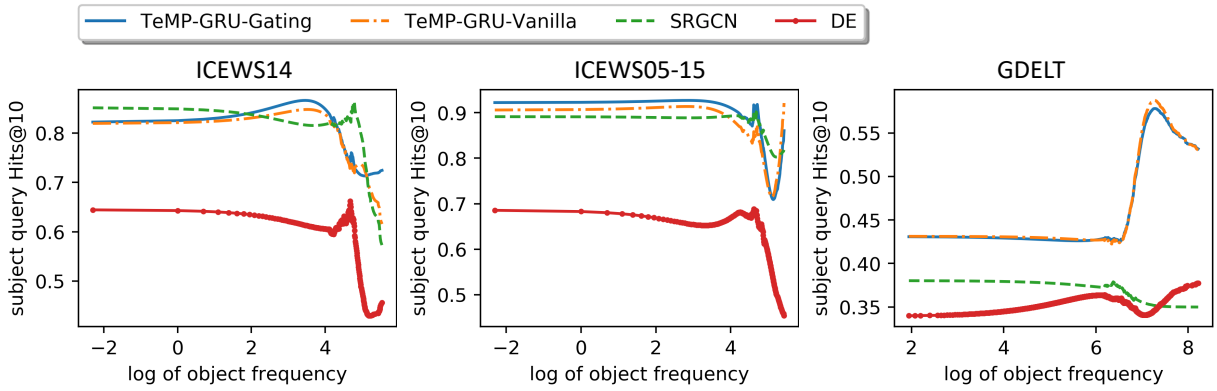
Figure 8: Plots of replication effect group




(a) Object query Hits@10 performances versus temporal subject frequencies




(b) Subject query Hits@10 performances versus temporal relation-object frequencies




(c) Subject query Hits@10 performances versus temporal object frequencies

Figure 9: Plots of reference effect group