Supplementary Material: Comparing Apples and Oranges? On the Evaluation of Methods for Temporal Knowledge Graph Forecasting

Julia Gastinger[1,2][0000−0003−1914−6723], Timo Sztyler[1][0000−0001−8132−5920], Lokesh Sharma[1][0009−0009−2522−1209], Anett Schuelke[1], and Heiner Stuckenschmidt[2][0000−0002−0209−3859]

[1] NEC Laboratories Europe, Kurfuersten-Anlage 36, 69115 Heidelberg
`{firstname.lastname}@neclab.eu`
[2] University of Mannheim, Chair of Artificial Intelligence B6, 26, 68131 Mannheim
`heiner.stuckenschmidt@uni-mannheim.de`

A Supplementary Material

A.1 Additional Information on Experimental Settings

In the following, we describe experimental settings in general, and subsequently special settings for each model, if they are different from the default settings or otherwise noteworthy.

For each model, we log the predicted scores for each quadruple from the test set D_{test}, in both directions (subject and object prediction) in a python dictionary. The dictionary contains as keys the query $(s, r, ?, t)$ or $(?, r, o, t)$, and as values the scores predicted by the model for each entity $v \in \mathcal{V}$ to belong to that query. After running all experiments on all datasets, settings, and methods, we compute the ranks (MRR, Hits@k, (with $k = 1, 3, 10$)) as described in Section 3.1 of the main paper based on these dictionaries.

If not stated otherwise, we use the hyperparameter settings (including the number of training epochs) reported in the respective papers. For each model, if the published source code provides the option to manually set a random seed, we did set the same seed. If the option was not explicitly provided in the published source code or could be reached with small ($\leq$ 5 lines) modifications, during training we did not validate the models on the same filter setting as we did test them on. Instead, we validated on the default filter setting (please see details for each model below). The reason is that, in a preliminary experiment on selected models, we found that the filter setting has only a small influence on the best validation epoch (i.e., different settings for validation lead more often than not to the same best epoch). Thus, we are confident that this did not significantly influence the final results. Due to issues regarding memory consumption and very high computing time, we were not able to conduct the experiments for the dataset ICEWS05-15 for the better part of models and thus excluded this dataset from our experiments.

RE-Net [3] We run RE-Net in multi-step setting only, because the published source code does not provide the option to set the single-step option in the arguments. Also, when we asked via e-mail and GitHub issue about how to conduct the implementation of the single-step option, unfortunately, we did not receive a concrete reply. We train the models on the static filtered MRR, following the training procedure provided in the source code. Due to GPU memory issues with the dataset GDELT, we run the model for this specific dataset on CPU, which leads to a very long runtime ($>$ 50 days of training). For the source code to be able to run on CPU, we have to conduct modifications to the source code. We run all other experiments for RE-Net on GPU.

RE-GCN [5] We train the models on the raw MRR, following the training procedure provided in the source code. While RE-GCN is originally also evaluated on relation prediction, we exclude this setting in our study, as the other models do not support relation prediction. We run RE-GCN in both settings, single-step and multi-step.

CyGNet [8] We run CyGNet on multi-step setting only, because non-trivial modifications in the source code would be necessary to run in single-step setting. Unfortunately, the authors did not reply to our question on the concrete implementation of multi-step setting. We train one model for each setting, raw, static, and time-aware filter. For testing, instead of allowing the model to only use the information from triples in the train set, we allow to also take into account the triples in the validation set. For this, we slightly adjust the original source code: In our version, the historical vocabulary (copy-sequences) now includes all timesteps from train and validation set, instead of only the timesteps from the train set. Please see Figure 1 for an overview of the change in testing scores for time-aware filter setting, when using the validation set (our modification) versus not using the validation set (original) during testing.

TLogic [6] For the datasets ICEWS14 and ICEWS18 we use the hyperparameters as described in the paper. The hyperparameter that changes across datasets is the window size w. According to the authors, the higher the window size, the better the performance, but also the higher the memory need. This means, that generally, the smaller (and less dense) the graph, the higher the window size can be in regard to memory usage. The datasets YAGO, WIKI, and GDELT have not been evaluated in the original paper. For the small dataset YAGO we set the window size to $w = 0$, which includes all past timesteps. For WIKI and GDELT we experience memory issues when using the machines described in Section 5 (main paper), and even when using a machine with 2 TB Memory. Integrating instructions kindly provided by the authors, for the datasets WIKI and GDELT we can circumvent these memory issues by decreasing the rule length to $l = \{1, 2\}$, instead of $l = \{1, 2, 3\}$. In addition, for WIKI and GDELT we set the window size to $w = 200$, the value reported by [6] for the larger dataset ICEWS18.

For the multi-step setting, we modify the published source code. Instead of allowing the model to only apply the rules based on occurrences of quadruples in the train set, we allow to also take into account the quadruples in the validation set. We modify the highest timestep for the rule application to be the highest timestep from the validation set, instead of the highest timestep from the training set. In addition, for datasets ICEWS18, WIKI and GDELT, we implement the option to set the window size of $w = 200$ also for multi-step prediction (instead of using all quadruples from training and validation set). Please see Figure 1 for an overview of the impact of using the validation set (our modification) versus not using the validation set (original) during testing (rule application) on testing scores for time-aware filter setting.

TimeTraveler [7] We run TimeTraveler only in single-step setting, because, as kindly confirmed by the authors, non-trivial modifications in the source code would have been necessary to run in multi-step setting. For GDELT, no hyperparameters were specified in the original paper. We use the same hyperparameters as for WIKI, because this dataset is the most similar in size. TimeTraveler is capable of doing inductive link prediction for future timesteps, i.e., prediction of triples with previously unseen nodes. We do not specifically evaluate this capability, as it is not in the scope of our study.

xERTE [1] We run xERTE only in single-step setting, because, as kindly confirmed by the authors, non-trivial modifications in the source code would have been necessary to run in multi-step setting. In the paper, hyperparameters are not specified for the datasets WIKI and GDELT. We use the hyperparameters as specified for the ICEWS18 dataset because ICEWS18 is most similar in size. For each epoch during training, we log the validation results for raw, static, and time-aware filter settings. We run separate testing for the three filter settings, where we select the trained model from the best training epoch for the respective setting. Please note, that in most cases, the best epoch was the same across settings[3]. We experienced a very long training time ($>$ 30 days) for xERTE on the GDELT dataset.

TANGO [2] We only run TANGO in single-step setting, because non-trivial modifications in the source code would be necessary to run in multi-step setting. Unfortunately, the authors did not reply to our question on the possibility of implementing the multi-step setting, nor on the question of how to realize the long-horizontal Forecasting experiment they report in their paper. For GDELT, no hyperparameters were specified in the original paper. We use the same hyperparameters as for WIKI, because it is the most similar in size. We train one model for each setting, raw, static, and time-aware filter. TANGO is capable of doing inductive link prediction for future timesteps, i.e., prediction of triples with previously unseen nodes. We do not specifically evaluate this capability, as it is not in the scope of our study.

[3] The best epoch e_{best} was $e_{best} = 8$ for 9 out of 12 cases (3 settings across 4 experiment runs), with $variance_{bestepoch} = 0.52$.

CEN [4] We only run CEN in single-step setting, because non-trivial modifications in the source code would be necessary to run in multi-step setting. For GDELT and YAGO, no hyperparameters were specified in the original paper. We use the same hyperparameters for GDELT as for WIKI, and for YAGO as for ICEWS14 because they are the most similar in size. We train the models on the raw MRR, following the training procedure provided in the source code. We first run the pretraining step with the minimum length as specified in the paper, and the curriculum training step. After this, we extract the *test-history-len* k, for each model, i.e. the history length where the model received the best validation score. We find the values of k to be $k = \{10, 7, 6, 2, 3\}$ for {GDELT, ICEWS14, ICEWS18, WIKI, YAGO}. We provide these values during testing and during online learning, as described by the authors in their GitHub repository.

A.2 Additional Experiment Results

Usage of the Validation Set Figure 1 shows the performance of the methods TLogic and CyGNet on all datasets in multi-step setting, when not leveraging the information from the validation set D_{valid} (option a), versus leveraging D_{valid} (option b) (see Section 3 in main paper) during testing. Please note that the drop in scores for the sixth timestep on the YAGO dataset is due to the dataset only having two samples in this snapshot, and all models performing bad on these two samples.

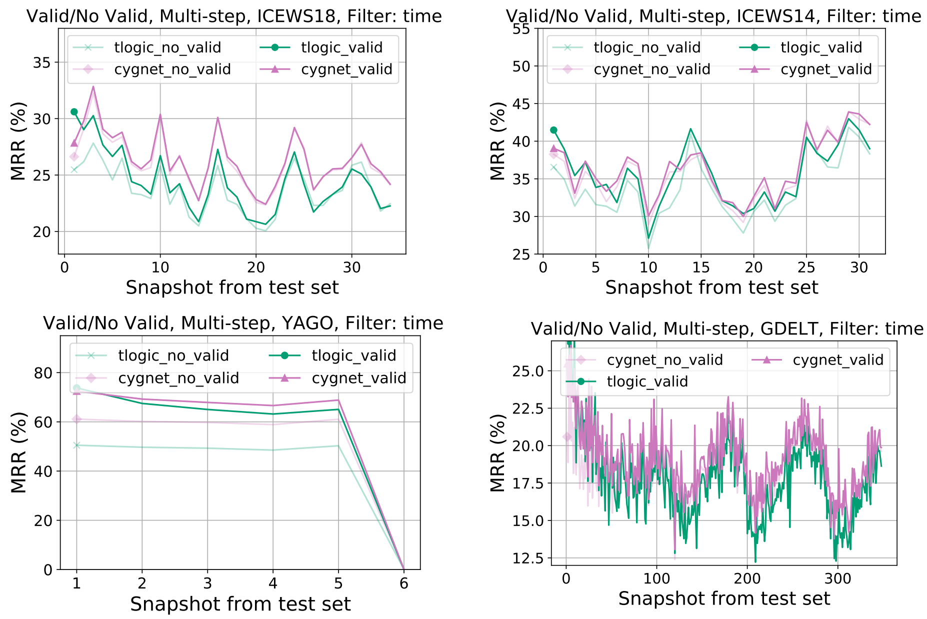

Fig. 1. MRR (in %) over snapshots from test set per method for datasets ICEWS18 (top left), ICEWS14 (top right), YAGO (bottom left), and GDELT (bottom right) for methods CyGNet and TLogic for multi-step prediction in time-aware filter setting. Each Subfigure shows the MRR when leveraging the information from the validation set during testing, vs. when not using it. Figures for static and raw setting are available upon request.

Static and Raw Filters Tables 1 and 2 report results on static filter setting and raw filter setting, for the five datasets GDELT, YAGO, WIKI, ICEWS14, and ICEWS18. Although we do not encourage evaluation on these settings, we have added the results for reasons of completeness and comparability. Figure 2 shows the MRR over test timestamps (snapshots), for multi-step and single-step prediction for the three remaining datasets (ICEWS14, YAGO, and GDELT) that have not been shown in the main paper. Please note that the drop in scores for the sixth timestep on the YAGO dataset is due to the dataset only having two samples in this snapshot, and all models performing badly on these two samples.

Table 1. Experiment results for multi-step and single-step prediction with datasets GDELT, YAGO, WIKI, ICEWS14, and ICEWS18. Results for single-step prediction should not be compared to results for multi-step prediction. We report mean reciprocal rank (MRR), and Hits@k (H@k), with $k = 1, 2, 3$ in static filter setting (static filter).

multi-step setting (static filter)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	39.90	32.38	43.12	53.27	77.81	75.36	78.95	82.30	66.16	64.73	66.80	68.49
RE-Net	41.45	34.68	43.84	54.04	64.99	63.44	65.31	67.73	52.18	51.27	52.31	53.83
CyGNet	**53.01**	**46.52**	**56.62**	**64.14**	**84.57**	**83.93**	**84.76**	**85.52**	**69.00**	68.38	**69.26**	**70.02**
TLogic	35.77	30.00	37.80	46.68	71.39	71.10	71.27	71.87	68.54	**68.52**	68.54	68.55
single-step setting (static filter)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	40.26	32.84	43.35	53.55	83.81	81.05	85.18	88.99	81.83	79.96	82.92	85.05
xERTE	29.38	24.62	32.05	39.00	**90.44**	**89.90**	90.82	**91.28**	78.73	77.65	79.58	80.42
TLogic	37.62	30.47	41.11	51.78	79.10	78.97	79.06	79.28	**87.18**	**87.16**	**87.19**	**87.20**
TANGO	41.03	**35.12**	42.88	52.25	67.88	66.95	67.85	69.47	52.46	52.12	52.58	53.06
Timetraveler	28.62	23.90	29.29	37.33	90.26	89.37	**90.99**	91.24	82.60	82.19	82.73	83.27
CEN	**42.19**	34.77	**45.39**	**55.35**	85.24	82.72	86.56	89.88	82.26	80.47	83.37	85.21
online setting (single-step with model update) (time filter)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
CEN	41.26	33.24	44.68	56.20	86.27	83.74	87.70	90.78	83.22	81.66	84.15	85.78

multi-step setting (static filter)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	48.14	40.12	51.78	**63.51**	41.23	33.58	44.28	55.81
RE-Net	48.21	41.52	50.81	61.13	42.88	36.19	45.36	55.97
CyGNet	**53.10**	**47.83**	**55.47**	62.85	**47.97**	**42.70**	**50.05**	**57.81**
TLogic	51.15	46.37	53.28	60.66	43.10	38.37	45.22	52.20
single-step setting (static filter)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	52.91	44.97	56.91	**67.91**	45.57	37.49	49.04	**61.23**
xERTE	46.22	40.12	50.09	58.61	37.30	31.14	40.65	49.91
TLogic	**57.76**	**52.86**	**60.55**	66.86	**47.66**	**42.07**	**50.62**	58.27
TANGO	50.71	45.20	52.90	61.58	41.88	34.68	44.90	55.32
Timetraveler	48.09	41.62	51.00	60.17	36.09	30.10	38.37	47.25
CEN	51.30	43.87	54.82	65.66	44.13	36.23	47.64	59.10
online setting (single-step with model update) (time filter)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
CEN	51.67	44.00	55.14	66.61	43.89	35.66	47.66	59.52

Table 2. Experiment results for multi-step and single-step prediction with datasets GDELT, YAGO, WIKI, ICEWS14, and ICEWS18. Results for single-step prediction should not be compared to results for multi-step prediction. We report mean reciprocal rank (MRR), and Hits@k (H@k), with $k = 1, 2, 3$ in raw setting (raw).

multi-step setting (raw)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	19.21	11.96	20.47	33.34	**58.20**	**47.94**	**65.47**	75.73	39.96	31.74	44.57	54.01
RE-Net	**19.27**	**11.97**	**20.51**	**33.63**	46.49	37.74	52.13	61.55	31.00	25.12	33.76	41.29
CyGNet	18.68	11.41	19.90	32.81	54.88	43.52	61.54	**77.77**	37.58	28.37	42.72	54.08
TLogic	17.35	10.88	18.57	30.05	52.36	42.22	60.46	69.90	**40.58**	**32.49**	**45.67**	**54.72**
single-step setting (raw)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	19.31	11.99	20.62	33.56	63.07	51.96	71.00	82.24	51.51	40.90	58.21	69.49
xERTE	18.51	12.26	20.76	31.75	64.70	52.07	74.48	**87.31**	53.15	42.07	61.15	71.93
TLogic	19.30	11.69	21.23	**35.31**	57.88	46.56	67.10	77.52	53.38	42.18	**61.41**	**72.09**
TANGO	18.80	11.69	20.04	32.52	49.02	40.61	55.04	63.01	30.72	25.07	33.74	40.42
Timetraveler	19.77	**13.52**	**21.84**	31.02	**64.83**	**52.23**	**74.57**	**87.31**	**53.41**	**42.27**	61.35	71.98
CEN	**19.95**	12.40	21.37	34.73	63.23	51.82	71.55	83.02	51.81	41.10	58.82	69.75
online setting (single-step with model update) (time filter)												
	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
CEN	21.18	13.13	23.02	36.95	63.75	52.04	72.48	83.89	52.09	41.11	59.47	70.35

multi-step setting (raw)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	**37.23**	**27.14**	**41.59**	**57.31**	**27.77**	**17.94**	**31.46**	**46.99**
RE-Net	36.27	26.75	40.31	54.68	26.59	16.87	30.29	45.67
CyGNet	35.41	25.86	39.72	54.42	25.05	15.53	28.66	43.83
TLogic	34.85	25.70	39.05	52.92	23.09	14.51	26.30	40.76
single-step setting (raw)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN	41.29	30.22	46.66	**62.45**	**31.02**	**20.38**	**35.41**	**51.94**
xERTE	40.06	31.74	45.01	56.91	27.95	19.23	32.46	45.84
TLogic	**41.56**	**31.81**	**46.95**	60.08	28.16	18.59	32.35	47.50
TANGO	36.04	26.29	40.34	54.88	27.03	17.42	30.79	45.74
Timetraveler	40.08	30.88	44.89	57.44	28.04	19.85	31.63	43.59
CEN	40.94	30.71	46.01	60.68	29.92	19.65	34.11	50.11
online setting (single-step with model update) (time filter)								
	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
CEN	42.17	31.83	47.28	62.23	30.21	19.81	34.46	50.70

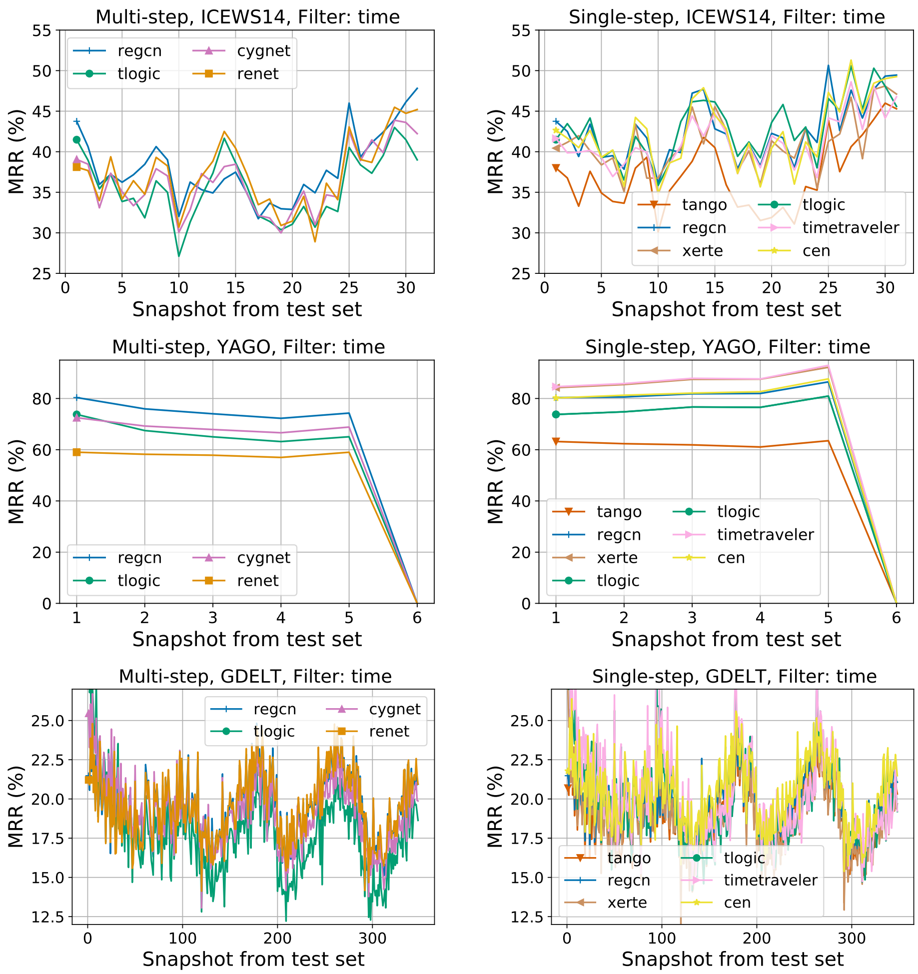

Fig. 2. MRR (in %) over snapshots from test set (one snapshot is one timestamp) per method for datasets ICEWS14 (top), YAGO (middle) and GDELT (bottom) for multi-step prediction (left) and single-step prediction (right). Figures for static and raw setting are available upon request.

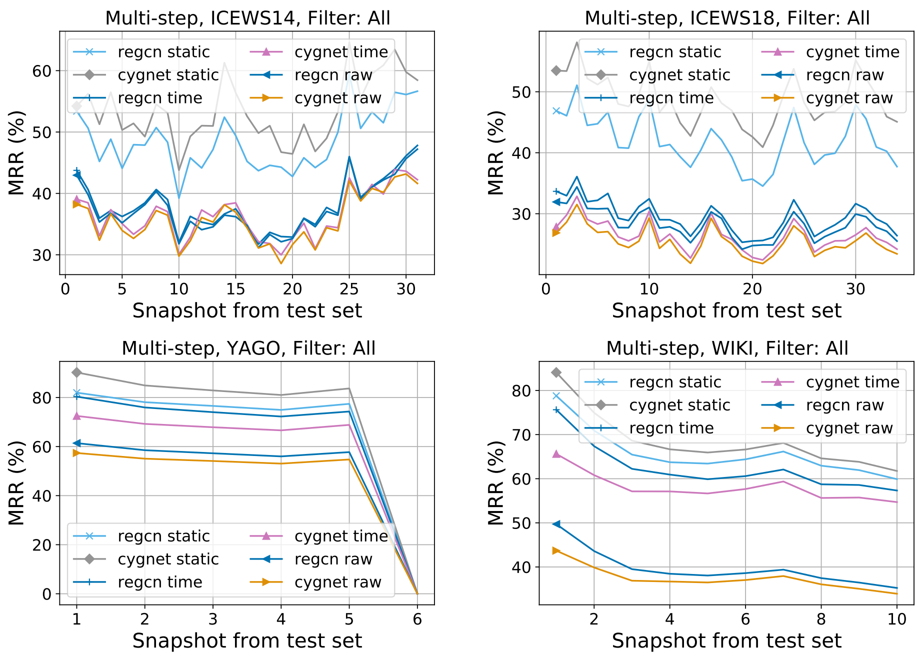

Fig. 3. MRR (in %) over snapshots from test set (one snapshot is one timestamp) for methods CyGNet and RE-GCN on all filter settings (raw, static, time-aware filter) for datasets ICEWS14 (top left), ICEWS18 (top right), YAGO (bottom left) and WIKI (bottom right) for multi-step prediction. Figures for other methods and single-step prediction are available upon request.

Result Consistency Tables 3 and 4 show the difference Δ of scores (MRR and Hits) reported by the authors of the original papers to the results from our experiments (as reported in Table 2, 3, and 4), if computable. As we show in Table 1, various differences in evaluation settings exist, and not all papers report results on all datasets, thus it is not possible to compute the differences for all datasets and settings for each method. Note: The results we find for RE-Net on the Wiki dataset in static filter setting are not consistent with the results reported by the authors of the original paper. However, our results are consistent with the results that the authors of CyGNet [8] report for RE-Net in this setting on this dataset, where we have $\Delta_{\text{MRR}} = \text{MRR}_{\text{CygNet reports for ReNet}} - \text{MRR}_{\text{This Work}} = -0.21$, $\Delta_{\text{H@3}} = -0.24$, and $\Delta_{\text{H@10}} = 0.08$.

Table 3. Experiment result Consistency: Difference Δ in reported scores (MRR and Hits) on multi-step and single-step setting, for time-aware, static, and raw setting on the datasets ICEWS14, ICEWS18, with $\Delta_{\text{Score}} = \text{Score}_{\text{Original Paper}} - \text{Score}_{\text{This Work}}$. Entry **n.r.**: result was **n**ot **r**eported by the original paper in this setting. Entry **d.v.**: a different **d**ataset **v**ersion was used in the original paper.

	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN								
multi-step								
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	0.55	0.03	0.94	1.53	-0.26	-0.12	-0.29	-0.44
single-step								
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	0.21	0.64	-0.06	0.02	-0.47	-0.38	-0.68	-0.48
RE-Net								
multi-step								
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	d.v.	d.v.	d.v.	d.v.	0.05	n.r.	0.11	-0.17
raw	d.v.	d.v.	d.v.	d.v.	0.03	n.r.	-0.02	-0.1
single-step	not computed							
CyGNet	Different usage of validation set: option (a) as described in section 3.4 in main paper. Thus, results are not comparable.							
Tlogic	Different usage of validation set: option (a) as described in section 3.4 in main paper. Thus, results are not comparable.							
xERTE								
multi-step	not reported							
single-step								
time-aware	d.v.	d.v.	d.v.	d.v.	0.08	0.11	0.01	0.22
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
TANGO								
multi-step	not reported							
single-step								
time-aware	d.v.	d.v.	d.v.	d.v.	0.62	0.41	0.73	1.24
static	d.v.	d.v.	d.v.	d.v.	2.68	3.19	2.56	1.74
raw	d.v.	d.v.	d.v.	d.v.	0.56	0.35	0.61	1.18
TimeTraveler								
multi-step	not reported							
single-step								
time-aware	d.v.	d.v.	d.v.	d.v.	0.85	0.76	0.92	0.91
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
CEN								
multi-step	not reported							
single-step								
time-aware	0.40	0.23	0.87	0.44	0.00	0.01	0.04	-0.10
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
single-step online time-aware	0.17	-0.02	0.46	0.15	0.88	0.73	1.02	1.23
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.

Table 4. Experiment result Consistency: Difference Δ in reported scores (MRR and Hits) on multi-step and single-step setting, for time-aware, static, and raw setting on the datasets GDELT, YAGO, and WIKI, with $\Delta_{\text{Score}} = \text{Score}_{\text{Original Paper}} - \text{Score}_{\text{This Work}}$. Entry **n.r.**: result was **n**ot **r**eported by the original paper in this setting. Entry **d.v.**: a different **d**ataset **v**ersion was used in the original paper.

	GDELT				YAGO				WIKI			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN												
multi-step												
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	-0.06	-0.04	-0.07	-0.15	0.07	n.r.	0.15	0.21	-0.12	n.r.	-0.14	-0.13
single-step												
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	0	0	-0.01	0.03	0	n.r.	0.17	-0.17	0.02	n.r.	0.08	0.04
RE-Net												
multi-step												
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
static	-1.03	n.r.	-0.44	-0.34	0.17	n.r.	0.32	0.35	12.98	n.r.	13.32	14.25
raw	0.33	n.r.	0.05	0.26	0.32	n.r.	0.58	0.38	-0.13	n.r.	1.79	-0.02
single-step	not computed											
CyGNet	Different usage of validation set: option (a) as described in section 3.4 in main paper. Thus, results are not comparable.											
Tlogic	Different usage of validation set: option (a) as described in section 3.4 in main paper. Thus, results are not comparable.											
xERTE												
multi-step	not reported											
single-step												
time-aware	n.r.	n.r.	n.r.	n.r.	d.v.	d.v.	d.v.	d.v.	n.r.	n.r.	n.r.	n.r.
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
TANGO												
multi-step	not reported											
single-step												
time-aware	n.r.	n.r.	n.r.	n.r.	0.95	1	0.5	1.04	2.96	3.22	2.43	2.7
static	n.r.	n.r.	n.r.	n.r.	0.46	0.1	0.54	1.23	1.59	-0.6	1.26	2.4
raw	n.r.	n.r.	n.r.	n.r.	0.47	0.29	0.38	0.73	1.81	1.26	2.01	2.75
TimeTraveler												
multi-step	not reported											
single-step												
time-aware	n.r.	n.r.	n.r.	n.r.	-0.26	0.34	-0.91	-0.93	-3.15	-2.19	-4.54	-4.03
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
CEN												
multi-step	not reported											
single-step												
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	-0.36	-0.46	-0.47	-0.01
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
single-step online												
time-aware	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	-0.15	-0.25	-0.14	0.11
static	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
raw	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.

A.3 Checklist for Benchmark Experiments on TKG Forecasting

In the following, we provide a checklist for benchmark experiments on TKG Forecasting.

- Are the datasets used the same version for all models? Check e.g., number of triples in train, validation, test set.
- Are the hyperparameters set as reported in the papers?
- Is the single-step/ multi-step setting consistent across models?
- Is the validation set used during testing?
- Are you sure that the test set is not leaked during training?
- Does the model predict in both directions, $(s, r, ?, t)$ and $(?, r, o, t)$?
- Are evaluation scores computed based on time-aware filtered setting? Is the implementation to compute the evaluation scores consistent across all models?

Bibliography

[1] Han, Z., Chen, P., Ma, Y., Tresp, V.: Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=pGIHq1m7PU

[2] Han, Z., Ding, Z., Ma, Y., Gu, Y., Tresp, V.: Learning neural ordinary equations for forecasting future links on temporal knowledge graphs. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 8352–8364. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.658, https://doi.org/10.18653/v1/2021.emnlp-main.658

[3] Jin, W., Qu, M., Jin, X., Ren, X.: Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. pp. 6669–6683. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.541, https://doi.org/10.18653/v1/2020.emnlp-main.541

[4] Li, Z., Guan, S., Jin, X., Peng, W., Lyu, Y., Zhu, Y., Bai, L., Li, W., Guo, J., Cheng, X.: Complex evolutional pattern learning for temporal knowledge graph reasoning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 290–296. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-short.32, https://aclanthology.org/2022.acl-short.32

[5] Li, Z., Jin, X., Li, W., Guan, S., Guo, J., Shen, H., Wang, Y., Cheng, X.: Temporal knowledge graph reasoning based on evolutional representation learning. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 408–417. ACM (2021). https://doi.org/10.1145/3404835.3462963, https://doi.org/10.1145/3404835.3462963

[6] Liu, Y., Ma, Y., Hildebrandt, M., Joblin, M., Tresp, V.: Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 4120–4127. AAAI Press (2022), https://ojs.aaai.org/index.php/AAAI/article/view/20330

[7] Sun, H., Zhong, J., Ma, Y., Han, Z., He, K.: Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 8306–8319. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.655, https://doi.org/10.18653/v1/2021.emnlp-main.655

[8] Zhu, C., Chen, M., Fan, C., Cheng, G., Zhang, Y.: Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 4732–4740. AAAI Press (2021), https://ojs.aaai.org/index.php/AAAI/article/view/16604