Supplementary Material for History repeats itself: A Baseline for Temporal Knowledge Graph Forecasting

Julia Gastinger[1,2] , **Christian Meilicke**[2] , **Federico Errica**[1] ,
Timo Sztyler[1] , **Anett Schuelke**[1] , **Heiner Stuckenschmidt**[2]
[1] NEC Laboratories Europe [2] University of Mannheim
{julia.gastinger, federico.errica, timo.sztyler, anett.schuelke}@neclab.eu
{christian.meilicke, heiner.stuckenschmidt}@uni-mannheim.de

1 Hyperparameters

We selected the hyperparameter λ from the following set: $\lambda \in L = \{0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.5, 0.9, 1.0001\}$. Figure 1 illustrates the time decay function $\Delta_\lambda(t,k) = 2^{\lambda(k-t)}$ for these values. Further, the hyperparameter α was chosen from the set: $\alpha \in A = \{0, 0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 0.9, 0.99, 0.999, 0.9999, 0.99999, 1\}$.

To assess the impact of hyperparameter values on prediction scores, we stipulate values for λ and α across all relations and show the prediction scores for different values of λ and α. Table 1 displays the values for α and λ, selected based on the validation MRR. Furthermore, Table 2 demonstrates how λ influences the test MRR for the *Strict Recurrency Baseline* (ψ_Δ) on ICEWS14. To analyse the influence of α on the *Combined Recurrency Baseline*, we set $\lambda = 0.04$, which corresponds to the value that yielded the best performance on ICEWS14 for ψ_Δ. Table 3 presents the test MRR on ICEWS14 for different values of α for the *Combined Recurrency Baseline* $\psi_{\Delta_{\lambda=0.04}}\xi$.

Table 4 presents the test MRR for the *Strict Recurrency Baseline* ψ for selected time decay settings on all datasets. We display results for $\lambda = \{0, 0.1, 1.0001\}$ stipulated for all relations, and values for λ_r(`per rel`), i.e., λ selected per relation based on validation MRR. From Proposition 3, we know that $\psi_{\Delta_{\lambda=0}}$ and ψ_1 are ranking equivalent, i.e. $\psi_{\Delta_{\lambda=0}} \Leftrightarrow \psi_1$. Further, we know from Proposition 2 that $\psi_{\Delta_{\lambda=1.0001}} \Leftrightarrow \phi_\Delta$.

Further, to understand the impact of the *Relaxed Recurrency Baseline* (ξ) on the combined baseline, Figure 2 presents a comparative analysis of the MRR of strict, relaxed, and combined baseline across individual relations within the YAGO dataset. For two relations hasWonPrize (tail) and diedIn (tail), the *Relaxed Recurrency Baseline* achieves significantly better results than the *Strict Recurrency Baseline*. Remarkably, the *Combined Recurrency Baseline* consistently demonstrates performance equal to or slightly better than the best-performing individual baseline across all relations examined. This observation underscores the effectiveness of the combined baseline approach and shows the benefit of selecting the value of α per relation.

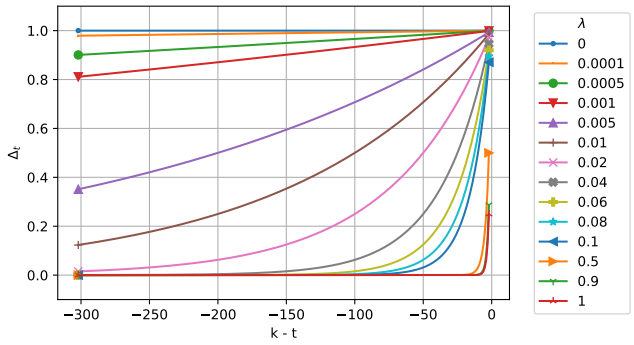

Figure 1: The time decay function $\Delta_\lambda(t,k) = 2^{\lambda(k-t)}$ for different values of λ. $k-t$ represents the distance between the query timestep t and the timestep of last occurence k.

2 Results for Multi-step Prediction

Multi-step prediction involves forecasting multiple future timesteps at once. In this setting, the model predicts all timesteps from the test set without having access to any ground truth information in between. Multi-step prediction presents a greater challenge because the model relies solely on its own forecasts, leading to an accumulation of uncertainty as the number of forecasted timesteps increases. [Gastinger *et al.*, 2023]

Table 5 shows the test results (MRR and Hits@10) for all datasets for the methods from related work that run in multi-step setting, as well as the *Relaxed Recurrency Baseline* (ξ), the *Strict Recurrency Baseline* (ψ_Δ), and the *Combined Recurrency Baseline* ($\psi_\Delta\xi$). We obtain the results for related work following the evaluation protocol in [Gastinger *et al.*, 2023]. The hyperparameter values for our baselines have been selected for each relation following the same procedure as described for the single-step setting.

Overall, the scores for multi-step prediction are lower than those for single-step prediction, aligning with expectations. Furthermore, the results reveal a consistent pattern with the results obtained in single-step prediction. In comparison to related work, the *Combined Recurrency Baseline* demonstrates superior performance in two out of the five datasets (YAGO and WIKI) and ranks third in performance for a third dataset (GDELT). For ICEWS14 and ICEWS18, the major-

Table 1: Values for parameters λ and α that have been selected per dataset when stipulating values for λ and α across all relations.

	GDELT	YAGO	WIKI	ICEWS14	ICEWS18
λ	0.01	1.0001	1.0001	0.02	0.02
α	0.99	0.99	0.99999	0.999	0.99

Table 2: Test MRR for the *Strict Recurrency Baseline* (ψ_Δ) for different values of λ. Example for dataset ICEWS14.

λ	0	0.0001	0.0005	0.001	0.005	0.01	0.02	0.04	0.06	0.08	0.1	0.5	0.9	1.0001
MRR	34.4	35.3	35.3	35.3	35.6	35.7	35.9	36.0	36.0	35.9	35.5	30.3	27.9	27.5

Table 3: Test MRR for the *Combined Recurrency Baseline* ($\psi_{\Delta_{\lambda=0.04}}\xi$) for different values of α. Example for dataset ICEWS14.

α	0	1.00E-05	0.0001	0.001	0.01	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	1
MRR	14.4	14.6	14.6	14.6	14.8	17.5	27.0	34.7	35.9	36.6	37.1	37.2	37.2

Table 4: Experimental results for single-step prediction on the *Strict Recurrency Baseline* ψ for different time decay settings, i.e. for different values of λ.

	GDELT		YAGO		WIKI		ICEWS14		ICEWS18	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
$\psi_{\Delta_{\lambda=0}} \Leftrightarrow \psi_1$	20.8	35.3	81.6	90.9	68.1	83.5	34.4	47.2	26.2	40.3
$\psi_{\Delta_{\lambda=0.1}}$	19.5	31.3	85.7	92.3	73.7	85.6	35.5	47.8	26.8	40.4
$\psi_{\Delta_{\lambda=1.0001}} \Leftrightarrow \phi_\Delta$	12.1	18.3	90.7	92.8	81.6	87.0	27.5	33.8	21.0	29.9
$\psi_{\Delta_{\lambda_{r\texttt{(per rel)}}}}$	23.7	38.3	90.7	92.8	81.6	87.0	36.3	48.4	27.8	41.4

Table 5: Experimental results for multi-step prediction.

	GDELT		YAGO		WIKI		ICEWS14		ICEWS18	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
RE-GCN	19.6	33.6	75.4	81.7	62.7	67.9	**37.8**	**57.5**	**29.0**	**47.5**
RE-Net	**19.7**	**33.9**	58.2	66.3	49.5	53.0	37.0	54.9	27.9	46.2
CyGNet	19.1	33.1	69.0	83.4	58.3	67.6	36.1	54.5	26.0	44.4
TLogic	17.7	30.3	66.9	71.6	64.0	68.2	35.5	53.1	24.0	41.2
ξ	14.1	23.4	5.1	10.5	14.0	24.7	14.1	27.8	11.6	21.8
ψ_Δ	18.1	30.4	81.4	84.2	63.7	68.2	30.1	40.8	23.4	35.8
$\psi_\Delta\xi$	19.2	32.5	**81.7**	**84.6**	**64.3**	**69.0**	31.7	45.4	24.6	38.6

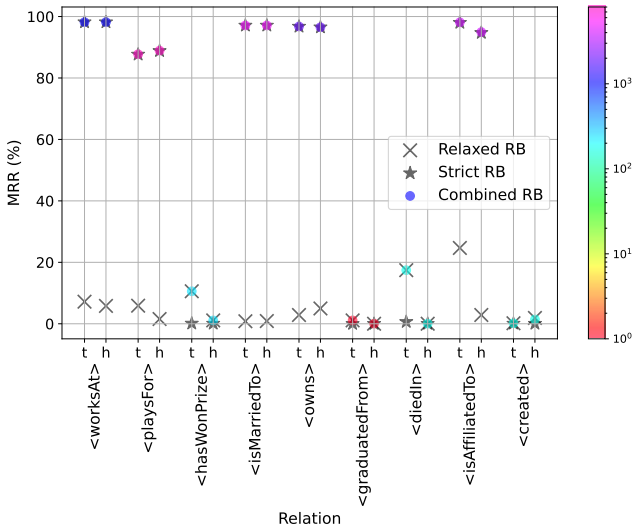

Figure 2: Test MRRs for the *Relaxed Recurrency Baseline* (cross), *Strict Recurrency Baseline* (star), and *Combined Recurrency Baseline* (dot) for each relation and direction ("t" means tail and "h" head, respectively) for YAGO. Colors indicate the number of queries for relation and its direction in the test set.

ICEWS14	ICEWS18	GDELT	YAGO	WIKI
728 s	6758 s	22812 s	778 s	9215 s

Table 6: Runtime including hyperparameter selection and testing for the Combined Recurrency Baseline $\psi_\Delta\xi$.

ity of methods from related work outperform the *Combined Recurrency Baseline*, yielding better results.

3 Runtime

All baselines run on CPU; we use an Intel Xeon Silver (2.10GHz) 4208 CPU with 16 cores (32 threads) and 512 GB RAM. The overall runtime is linear in the size of the dataset times the number of different hyperparameter settings, no matter the number of relations. For each relation r, we tune the hyper-parameter values of λ_r, α_r on the subset of quadruples with relation r, which is a strict subset of the total quadruples. This is possible because relations are independent from each other. Thus, for each hyper-parameter λ_r, α_r to be tuned, the complexity is linear in the number of quadruples of that specific relation times the number of values to try.

In Table 6 we report total runtimes required for our baseline (to perform all steps, i.e. data loading, hyperparameter selection, and testing). In Table 7, as an example, we compare this number for ICEWS14 to time spent to perform all steps except hyperparameter tuning for the other methods considered in Section 5.2 (thus ignoring hyperparameter tuning altogether). Despite performing hyperparameter selection, we need 20x less time than the fastest related method to get the final scores.

$\psi_\Delta\xi$	RE-GCN	CEN	xERTE	TLogic
728 s	31052 s	14535 s	51212 s	15212 s

Table 7: Total Runtime on ICEWS14 for $\psi_\Delta\xi$ and the methods considered in Section 5.2 of the main paper. Total runtime means time to perform all steps, i.e. data loading, hyperparameter selection, and testing for $\psi_\Delta\xi$ and all steps except hyperparameter tuning for the other methods (thus ignoring hyperparameter tuning altogether). TLogic and $\psi_\Delta\xi$ run on CPU, the others on GPU.

4 Additional Information on Related Work

TRKG Note on results for TRKG: We found that the authors of TRKG [Kiran *et al.*, 2023] compute their results (MRR and Hits@1,3,10) on the computation protocol "best". If there are multiple triples in the candidate set with the same score from the model, this protocol assigns the lowest rank, i.e. the best score, to the ground truth triple. As reported in [Sun *et al.*, 2020], this is an unfair evaluation protocol. Thus, we rerun the experiments for TRKG following the evaluation protocol introduced by [Gastinger *et al.*, 2023], who compute the ranks based on the "random" computation protocol. This explains the drop in result scores from the original TRKG paper as compared to our table.

5 Additional Information on Datasets

In the following we provide additional information on the datasets that we use for our experiments. For each dataset we use the version provided by [Li *et al.*, 2021] and [Gastinger *et al.*, 2023].

ICEWS Datasets: ICEWS14 [García-Durán *et al.*, 2018] and ICEWS18 [Jin *et al.*, 2019] are derived from the Integrated Crisis Early Warning System (ICEWS) [Boschee *et al.*, 2015], covering different time spans (for the years 2014 and 2018). ICEWS provides event data on various global events such as conflicts, protests, and diplomatic relations.

GDELT: The Global Database of Events, Language, and Tone (GDELT) [Leetaru and Schrodt, 2013] provides a comprehensive collection of global events derived from news articles and other sources. It covers a broad range of events, including political events, conflicts, and societal movements, spanning various geographic regions and time periods.

YAGO and WIKI: YAGO [Mahdisoltani *et al.*, 2015] offers structured data regarding entities, facts, and their relationships. The WIKI dataset is extracted from Wikidata [Vrandečić and Krötzsch, 2014] and has first been preprocessed to contain temporal information by [Leblay and Chekol, 2018]. We use versions of YAGO and WIKI that have further been preprocessed according to [Jin *et al.*, 2019], who formulate temporal information in quadruples and remove noisy events of early years (before 1786 for WIKI and 1830 for YAGO).

References

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

[García-Durán *et al.*, 2018] Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4816–4821, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Gastinger *et al.*, 2023] Julia Gastinger, Timo Sztyler, Lokesh Sharma, Anett Schuelke, and Heiner Stuckenschmidt. Comparing apples and oranges? On the evaluation of methods for temporal knowledge graph forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 533–549, 2023.

[Jin *et al.*, 2019] Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. *arXiv preprint arXiv:1904.05530*, 2019. preprint version.

[Kiran *et al.*, 2023] Rage Uday Kiran, Abinash Maharana, and Krishna Reddy Polepalli. A novel explainable link forecasting framework for temporal knowledge graphs using time-relaxed cyclic and acyclic rules. In *Proceedings of the 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Part I*, pages 264–275, 2023.

[Leblay and Chekol, 2018] Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1771–1776. ACM, 2018.

[Leetaru and Schrodt, 2013] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, pages 1–49. Citeseer, 2013.

[Li *et al.*, 2021] Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. Temporal knowledge graph reasoning based on evolutional representation learning. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2021.

[Mahdisoltani *et al.*, 2015] Farzaneh Mahdisoltani, Joanna Asia Biega, and Fabian M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*, 2015.

[Sun *et al.*, 2020] Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha P. Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5516–5522, 2020.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.