Towards Temporal Knowledge Graph Alignment in the Wild
Technical Report

Abstract—**This technical report contains the dataset details and the full experimental setup of the paper "Towards Temporal Knowledge Graph Alignment in the Wild".**

I. Dataset

In this section, we introduce several key metrics used to quantify the characteristics and differences between temporal knowledge graph alignment (TKGA) datasets for assessing the difficulty and realism.

Overlapping Rate (Overlapping.%). To better reflect the real-world scenario where cross-KG alignment is rarely strictly 1-to-1, we use the overlapping rate to measure the degree of shared temporal entities between two TKGs [1]. The overlapping rate is calculated as follows:

$$\textbf{Overlapping.\%}(G^s) = \frac{|\phi_{seed}|}{|E^s|} \times 100\%, \tag{1}$$

$$\textbf{Overlapping.\%}(G^t) = \frac{|\phi_{seed}|}{|E^t|} \times 100\%, \tag{2}$$

where ϕ_{seed} denotes the set of aligned entity pairs, and E^s and E^t are the set of all temporal entities in G^s and G^t.

Temporal Interval Consistency (Inter. Consis.%). To effectively assess the differences in temporal intervals between aligned entities in TKGA, we introduce a new metric called *temporal interval consistency*, which is the proportion of aligned entities with consistent temporal intervals among all aligned entities, defined as follows:

$$\begin{aligned}&\textbf{Inter.Consis.\%}(G^s, G^t) = \\ &= \frac{|(e_i^s, e_j^t) \in \phi_{seed} \mid \mathcal{T}(e_i^s) \cap \mathcal{T}(e_j^t) \neq \emptyset|}{|\phi_{seed}|} \times 100\%,\end{aligned} \tag{3}$$

where ϕ_{seed} is the set of aligned entity pairs, and $\mathcal{T}(e)$ denotes the set of time intervals associated with entity e.

Multi-source Valid Temporal Fact Ratio (MTF.%). To assess the extent of missing valid temporal facts in multi-source TKGs, we define the multi-source valid temporal fact ratio as follows:

$$\textbf{MTF.\%}(G^s, G^t) = \frac{|\mathcal{V}G^s + \mathcal{V}G^t|}{|G^s + G^t|} \times 100\%, \tag{4}$$

where $|\mathcal{V}G^s + \mathcal{V}G^t|$ represents the total number of valid temporal facts in the source and target TKGs, and $|G^s + G^t|$ denotes the total number of all facts in both TKGs.

Differences in Valid Temporal Facts and Valid Temporal Density (ΔT.F.% and ΔT.D.%). To better capture the imbalance of valid temporal facts and temporal structures between two TKGs in real-world TKGA scenarios, we define the relative differences in valid temporal facts and valid temporal fact density, respectively, as follows:

$$\Delta\textbf{T.F.\%}(G^s, G^t) = \frac{|\mathcal{V}G^s - \mathcal{V}G^t|}{\min(|\mathcal{V}G^s|, |\mathcal{V}G^t|)} \times 100\%, \tag{5}$$

$$\Delta\textbf{T.D.\%}(G^s, G^t) = \frac{|\rho^s - \rho^t|}{\min(\rho^s, \rho^t)} \times 100\%, \tag{6}$$

where $|\mathcal{V}G^s|$ and $|\mathcal{V}G^t|$ represent the number of valid temporal facts in source TKG and targe TKG, and $\rho = \frac{|\mathcal{V}G|}{|G|}$ denotes the density of valid temporal facts in a TKG.

II. Experiments

In this section, we first introduce the experimental setting[1] in Section II-A.

A. Experimental Setting

Datasets. In our experiments, we conducted comprehensive evaluations on **eight datasets**, including `BETA`, `WildBETA`, and 6 current datasets (as detailed in Table I).

Among these, `DICEWS` and `YAGO-WIKI50K` are the most frequently used datasets for Temporal Knowledge Graph Alignment (TKGA), derived from `ICEWS05-15`, `YAGO`, and `Wikidata`. Specifically, `ICEWS05-15` is constructed from the `ICEWS` dataset [5], which comprises political events annotated with specific dates, using a daily temporal resolution and covering the period from 2005 to 2015. Xu et al. [3] randomly partitioned the quadruples in `ICEWS05-15` into two equally sized subsets, yielding the datasets `DICEWS-200` (`D200`). The `YAGO-WIKI50K` datasets are similarly constructed by Xu et al. [3], who first selected the top 50,000 most frequent entities from a Wikidata subset (as extracted in [6]) and linked them to corresponding entities in YAGO. Temporal facts were then added to form two temporally enriched knowledge graphs. The resulting dataset `YAGO-WIKI50K-1K` contains 1000 seed entity pairs.

In contrast, `ICEWS-WIKI` and `ICEWS-YAGO` [1] represent new heterogeneous and temporal datasets, posing more realistic and challenging alignment scenarios. These datasets are characterized by significant discrepancies not only in the

[1] The source codes and datasets of the previous work are available at https://github.com/DexterZeng/BETA. The source codes and datasets for this extended work will be released upon acceptance

TABLE I: Dataset statistics [1]–[4]. "*#Ent*", "*#Rel.*", "*#Facts*", "*#T.Facts*": The number of entities, relations, quadruples and quadruples with valid time interval in KG1 (KG2), respectively. "*Temp.*", "*Multi-Granularity*" : Indicates whether the dataset includes temporal knowledge information and the dataset includes multi-granularity temporal knowledge information, respectively. "*#Overlapping*": Represents the proportion of overlapping temporal entities in KG1 and KG2. "*Inter. Consis.*" : Represents the proportion of aligned entities with consistent temporal intervals among all aligned entities. "*Multi-Source*", "*MTF.%*" : Refers to whether both TKGs in the dataset are temporal incompleteness, and the average proportion of valid temporal facts in the two TKGs, respectively. "*Δ T.F.%*", "*Δ T.D.%*" : Relative difference in valid temporal facts/density values between two KGs, using the KG with the smaller valid temporal facts/lower valid temporal density as the base.

Dataset		#Ent.	#Rel.	Temp.	Multi-Granularity	#Seed	#Overlapping	Inter. Consis. ⇓	#Facts	#T.Facts	Multi-Source	MTF.% ⇓	Δ T.F.% ⇑	#T.Density	Δ T.D.% ⇑
DBP15K(EN-FR)	EN	15,000	193	✗	✗	15,000	100%	✗	96,318	0	✗	✗	✗	✗	✗
	FR	15,000	166	✗	✗		100%		80,112	0				✗	
DBP-WIKI	DBP	100,000	413	✗	✗	100,000	100%	✗	293,990	0	✗	✗	✗	✗	✗
	WIKI	100,000	261	✗	✗		100%		251,708	0				✗	
ICEWS-WIKI	ICEWS	11,047	272	✓	✗	5,058	45.79%	55.63%	3,527,881	3,527,881	✗	96.05%	6,817.1%	319.352	9,853.4%
	WIKI	15,896	226	✓	✗		31.82%		198,257	51,002				3.208	
ICEWS-YAGO	ICEWS	26,863	272	✓	✗	18,824	70.07%	7.39%	4,192,555	4,192,555	✗	98.21%	13,764.3%	156.072	11,633.3%
	YAGO	22,734	41	✓	✗		82.80%		107,118	30,240				1.330	
DICEWS	ICEWS	9,517	247	✓	✗	8,566	90.01%	95.09%	307,552	307,552	✗	100%	0%	32.316	0.2%
	ICEWS	9,537	246	✓	✗		89.82%		307,553	307,553				32.248	
YAGO-WIKI50K	YAGO	49,629	11	✓	✗	49,172	99.08%	93.63%	221,050	221,050	✗	100%	43.8%	4.454	45.0%
	WIKI	49,222	30	✓	✗		99.90%		317,814	317,814				6.457	
BETA	WIKI	42,666	257	✓	✓	40,364	94.60%	55.12%	199,879	104,774	✓	48.22%	49.9%	2.456	48.6%
	YAGO	42,297	45	✓	✓		95.43%		162,320	69,896				1.653	
WildBETA	WIKI	27,519	301	✓	✓	17,124	62.23%	**5.27%**	527,977	142,145	✓	**25.37%**	**14,001.7%**	5.165	**13,722.5%**
	YAGO	26,975	40	✓	✓		63.48%		36,283	1,008				0.037	

number of entities, relations, and triples. Furthermore, the number of seeds is not directly proportional to the total entity count, adding complexity to the temporal alignment task.

In addition, two standard non-temporal KGA datasets, DBP15K(EN-FR) and DBP-WIKI [4], are also included. DBP15K(EN-FR) focuses on cross-lingual alignment, while DBP-WIKI offers a large-scale benchmark for aligning heterogeneous KGs. Both datasets exhibit similar structural properties and high overlap (100%) in aligned entities, relations, and facts.

Baselines. Currently, no specific solutions exist for TKGA-Wild. To establish a comprehensive baseline, we introduced 24 SOTA and classic baseline methods for extensive comparison:

- MTransE [7], which introduces translation vectors to align entity embeddings across languages; and
- AlignE [8], which employs neural relation extraction to identify key relationships; and
- BootEA [8], which is one of the most competitive translation-based EA methods; and
- GCN-Align [9], which trains GCNs to embed entities of each language into a unified vector space; and
- MRAEA [10], which applies attention over local neighborhoods and relation-level meta-information; and
- RREA [11], which implements relational reflection transformations to generate relation-aware embeddings; and
- RDGCN [12], which leverages GCNs for modeling structural information within knowledge graphs.
- Dual-AMN [13], which iointly captures intra-graph and cross-graph dependencies; and
- TEA-GNN [3], which treats timestamps as link attributes, using a time-aware attention mechanism to enrich entity and relation representations; and
- TREA [14], which enhances training using neighborhood aggregation and margin-based multi-class loss; and
- STEA [15], which utilizes a temporal dictionary to guide temporal alignment; and
- Dual-Match [16], which employs a temporal encoder for unsupervised layer-wise propagation; and
- MGTEA [17], which proposes a simple yet effective multi-granularity approach for temporal alignment; and
- LightTEA [18], which is a lightweight TKGA model, though its temporal component yields limited improvements on existing datasets; and
- BERT [19], utilized as a pretrained language model to initialize entity embeddings using name-based features; and
- FuAlign [20], which incorporates auxiliary information to address KG heterogeneity; and
- BERT-INT [21], which combines BERT-based augmentation with auxiliary cues for improved alignment; and
- PARIS [22], which is a probabilistic iterative method capable of aligning entities without prior alignments; and
- Simple-HHEA [1], which is a representation learning-based approach tailored for aligning heterogeneous and temporal KGs; and
- ChatEA [2], which applies large language models with fine-tuning to perform advanced KG alignment; and
- HTEA [23], which employs frequency-based temporal embeddings to enhance alignment performance; and
- Naive RAG [24], [25], a basic LLM-based RAG approach that first retrieves relevant information based on a user query and then generates answers using the retrieved content; and
- Self-Consistency [26], a chain-of-thought baseline that produces multiple reasoning paths and selects the most frequent answer as the final output. In our implementation, we further enhance it by using the top-1 most similar entity

from the similarity matrix produced by Simple-HHEA as a preprocessing step for the knowledge graph; and

- Self-RAG [27], a self-reflective RAG method aimed at improving the generation quality of LLMs.

Implementation details. All experiments were conducted on a server equipped with four NVIDIA GeForce RTX 4090 graphics cards, each with 24 GB of GDDR6X memory. The system features a 64-core processor and 480 GB of RAM. For storage, the server utilizes a 30 GB system disk alongside a 50 GB solid-state drive (SSD) for data storage. All implementations were carried out using the PyTorch framework.

The large language models (LLMs) reported in Table III and Table IV were evaluated under identical settings, employing GPT-4[2], except for ChatEA, which directly follows the results reported in its original paper. For subsequent experiments, unless otherwise specified, GPT-3.5[3] was adopted as the default LLM owing to its cost-effectiveness.

The multi-granular information encoders and the integrated training were configured with a learning rate of 0.01, a weight decay of 0.001, gamma set to 1.0, and were trained for 500 epochs. The training set proportions follow the settings used in prior work and are set as follows: `WildBETA` (2%), `YAGO-WIKI50K-1K` (2%), `DICEWS-200` (2.3%), `BETA` (10%) [17], `ICEWS-WIKI` (30%) [1], `ICEWS-YAGO` (30%) [1], `DBP15K(EN-FR)` (30%), and `DBP-WIKI` (30%).

Evaluation metrics. Consistent with prior benchmark studies [1], [4], we adopt two widely recognized evaluation metrics to assess the effectiveness of entity alignment models: Hits@k and Mean Reciprocal Rank (MRR).

1) Hits@k evaluates the proportion of correctly aligned entity pairs that appear among the top-k ranked candidates. Formally, let N denote the total number of reference (ground truth) alignments, and for each reference entity e_i, let rank_i denote the rank position of its correct counterpart in the candidate list. The Hits@k is defined as:

$$\text{Hits@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{rank}_i \leq k), \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise. In practice, Hits@1 reflects the strict accuracy of top-1 predictions, while Hits@10 provides insight into broader top-k retrieval performance.

2) Mean Reciprocal Rank (MRR) measures the average of the reciprocal ranks of the correct entities in the prediction lists. It is computed as:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}. \tag{8}$$

MRR captures both the presence and position of correct alignments, thereby emphasizing early correct retrieval.

Both metrics are positively oriented, meaning higher values indicate better alignment quality. Notably, in cases where models yield only the final alignment predictions (i.e., without ranked candidate lists), the Hits@1 score is substituted with standard precision.

[2] gpt-4-0125-preview from the OpenAI API, https://openai.com/api/

[3] gpt-3.5-turbo-1106 from the OpenAI API, https://openai.com/api/

References

[1] X. Jiang, C. Xu, Y. Shen, Y. Wang, F. Su, Z. Shi, F. Sun, Z. Li, J. Guo, and H. Shen, "Toward practical entity alignment method design: Insights from new highly heterogeneous knowledge graph datasets," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, T. Chua, C. Ngo, R. Kumar, H. W. Lauw, and R. K. Lee, Eds. ACM, 2024, pp. 2325–2336. [Online]. Available: https://doi.org/10.1145/3589334.3645720

[2] X. Jiang, Y. Shen, Z. Shi, C. Xu, W. Li, Z. Li, J. Guo, H. Shen, and Y. Wang, "Unlocking the power of large language models for entity alignment," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 7566–7583. [Online]. Available: https://aclanthology.org/2024.acl-long.408

[3] C. Xu, F. Su, and J. Lehmann, "Time-aware graph neural network for entity alignment between temporal knowledge graphs," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 2021, pp. 8999–9010.

[4] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami, and C. Li, "A benchmarking study of embedding-based entity alignment for knowledge graphs," *Proc. VLDB Endow.*, vol. 13, no. 11, pp. 2326–2340, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf

[5] J. Lautenschlager, S. Shellman, and M. Ward, "ICEWS Event Aggregations," 2015. [Online]. Available: https://doi.org/10.7910/DVN/28117

[6] T. Lacroix, G. Obozinski, and N. Usunier, "Tensor decompositions for temporal knowledge base completion," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[7] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment," in *IJCAI*, 2017, pp. 1511–1517.

[8] Z. Sun, W. Hu, Q. Zhang, and Y. Qu, "Bootstrapping entity alignment with knowledge graph embedding," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. ijcai.org, 2018, pp. 4396–4402.

[9] Z. Wang, Q. Lv, X. Lan, and Y. Zhang, "Cross-lingual knowledge graph alignment via graph convolutional networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Association for Computational Linguistics, 2018, pp. 349–357.

[10] X. Mao, W. Wang, H. Xu, M. Lan, and Y. Wu, "MRAEA: an efficient and robust entity alignment approach for cross-lingual knowledge graph," in *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*. ACM, 2020, pp. 420–428.

[11] X. Mao, W. Wang, H. Xu, Y. Wu, and M. Lan, "Relational reflection entity alignment," in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 2020, pp. 1095–1104.

[12] Z. Chen, Y. Wu, Y. Feng, and D. Zhao, "Integrating manifold knowledge for global entity linking with heterogeneous graphs," *Data Intelligence*, vol. 4, no. 1, pp. 20–40, 2022.

[13] X. Mao, W. Wang, Y. Wu, and M. Lan, "Boosting the speed of entity alignment 10 ×: Dual attention matching network with normalized hard sample mining," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 2021, pp. 821–832.

[14] C. Xu, F. Su, B. Xiong, and J. Lehmann, "Time-aware entity alignment using temporal relational attention," in *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 2022, pp. 788–797.

[15] L. Cai, X. Mao, M. Ma, H. Yuan, J. Zhu, and M. Lan, "A simple temporal information matching mechanism for entity alignment between temporal knowledge graphs," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju,*

Republic of Korea, October 12-17, 2022. International Committee on Computational Linguistics, 2022, pp. 2075–2086.

[16] X. Liu, J. Wu, T. Li, L. Chen, and Y. Gao, "Unsupervised entity alignment for temporal knowledge graphs," in *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo, and G. Houben, Eds. ACM, 2023, pp. 2528–2538. [Online]. Available: https://doi.org/10.1145/3543507.3583381

[17] W. Zeng, J. Zhou, and X. Zhao, "Benchmarking challenges for temporal knowledge graph alignment," *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273501043

[18] L. Cai, X. Mao, Y. Xiao, C. Wu, and M. Lan, "An effective and efficient time-aware entity alignment framework via two-aspect three-view label propagation," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 2023, pp. 5021–5029.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[20] C. Wang, Z. Huang, Y. Wan, J. Wei, J. Zhao, and P. Wang, "FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs," *Inform. Fusion*, vol. 89, pp. 41–52, Jan. 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2022.08.002

[21] X. Tang, J. Zhang, B. Chen, Y. Yang, H. Chen, and C. Li, "BERT}-{INT: A {BERT}-based interaction model for knowledge graph alignment," *interactions*, vol. 100, p. e1, 2020.

[22] F. M. Suchanek, S. Abiteboul, and P. Senellart, "Paris: Probabilistic alignment of relations, instances, and schema," *Proceedings of the VLDB Endowment*, vol. 5, no. 3, 2011.

[23] J. Li, W. Hua, F. Jin, and X. Li, "HTEA: heterogeneity-aware embedding learning for temporal entity alignment," in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10-14, 2025*, W. Nejdl, S. Auer, M. Cha, M. Moens, and M. Najork, Eds. ACM, 2025, pp. 982–990. [Online]. Available: https://doi.org/10.1145/3701551.3703588

[24] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, and X. Huang, "A survey of graph retrieval-augmented generation for customized large language models," *CoRR*, vol. abs/2501.13958, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2501.13958

[25] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting in retrieval-augmented large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 5303–5315. [Online]. Available: https://doi.org/10.18653/v1/2023.emnlp-main.322

[26] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=1PL1NIMMrw

[27] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=hSyW5go0v8