Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR meets Machine Learning and Explanation

A Simple Baseline for Inductive Knowledge Base Completion

Christian Meilicke , Rainer Gemulla , Julie Naegelen , Heiner Stuckenschmidt
University of Mannheim, Germany
{christian.meilicke, rgemulla, julie.naegelen, heiner.stuckenschmidt}@uni-mannheim.de

Abstract

Inductive knowledge graph completion models aim to perform link prediction on knowledge graphs with completely new entities and/or relations. The motivation and key goal of these methods is to transfer relational knowledge and patterns across disjoint knowledge graphs. Although such methods have shown promising empirical results across a variety of datasets, we argue here that these results do not provide sufficient evidence for the successful transfer of relational knowledge. We identify biases in the construction principles of available benchmark datasets, provide a simple baseline that does not rely on multi-step relational paths or more sophisticated relational patterns, and show empirically that this baseline performs on par with the state-of-the-art zero-shot model ULTRA. To provide evidence that relational knowledge transfer really takes place in inductive knowledge graph completion models, improved benchmarks and stronger empirical results are therefore required.

1 Introduction

The goal of knowledge graph (KG) completion (KGC)—also called KG link prediction—is to predict missing triples in an incomplete knowledge graph (Pezeshkpour, Tian, and Singh 2020). In this paper, we focus on inductive KGC, in which KGC is performed on a previously unseen target KG with entirely new entities and/or entirely new relations. The key goal of inductive KGC is to transfer relational knowledge from other KGs to the target KG and to avoid potentially expensive training or fine-tuning on the target KG. Inductive KGC models can thus be seen as foundation models for knowledge graphs (Galkin et al. 2024).

A key motivation for inductive KGC is that KGs evolve over time, e.g., new entities and/or relations may be added or removed (Teru, Denis, and Hamilton 2020). Examples include new users in a social network, new products on an e-commerce platform, new proteins in a biomedical KG, or new relation types in an industrial KG. In all these examples, it is plausible that new triples involve new entities as well as existing entities and relations; e.g., a new product may belong to an existing category or a new protein may interact with an existing one. The assumption of inductive KGC that the target KG consists of "entirely new" entities, which we refer to as *disjointness assumption*, then does not hold.

To evaluate inductive KGC methods, benchmark datasets were constructed by introducing disjointness artificially into publicly available KGs as follows: (i) two disjoint subsets of entities and/or relations of the original KG were selected, (ii) subgraphs induced by these subsets were used as training KG and target KG, respectively, (iii) a set of held-out triples from the target KG was used as a test set.

In this paper, we analyze and discuss undesirable consequences of this approach. We first argue that the target graphs of the benchmark datasets are characterized by a frequency bias w.r.t. relations involving common entities (such as types, categories, or superspreaders). This makes the prediction tasks simpler than corresponding ones on the whole original KGs. We then construct a simple, cheap link prediction baseline which does not rely on relational paths or complex relational patterns, yet is able to exploit frequency biases. Finally, we show empirically that this baseline performs competitively with the state-of-the-art zero-shot inductive KGC model ULTRA (Galkin et al. 2024) as well as prior models.

Our results hint at systematic problems in the construction of the benchmark datasets for inductive KBC. They also imply that the current performance results of recent methods do not allow to conclude that relational knowledge transfer or deeper relational reasoning took place. To provide solid evidence for such transfer, improved benchmarks and stronger empirical results are therefore required. Simple baselines, such as the one we provide here, should also be considered.

2 Inductive Knowledge Graph Completion

2.1 Preliminaries

A knowledge graph $G = (E_G, R_G, T_G)$ consists of a set E_G of entities, a set R_G of relations, and a set $T_G = \{r(h,t) \mid h,t \in E_G, r \in R_G\}$ of atomic formulas called triples, where each entity in E_G and each relation in R_G occurs in at least one triple in T_G. Given a target graph G (also called test graph or inference graph), the goal of KGC is to predict new triples for queries of form $r(q,?)$ or $r(?,q)$, where $q \in E_G$ and $r \in R_G$ denote the query entity and query relation, respectively. An answer is a triple of form $r(q,e)$ or $r(e,q)$, where answer entity $e \in E_G$ as well.

KGC methods are evaluated using a test set T_G^{test} of true but held-out triples for G, i.e., $T_G^{\text{test}} \cap T_G = \emptyset$. Each test triple has form $r(h,t)$, where $h,t \in E_G$, $r \in R_G$ and $r(h,t) \notin T_G$, and is used to generate queries $r(h,?)$

and $r(?, t)$. KGC methods usually produce a ranked list of possible answers to each of these queries. The reciprocal rank of the original test triple $r(h, t)$ is then determined, and the mean reciprocal rank (MRR; Voorhees and Tice (1999)) across all queries is used as a performance metric. As is common in the literature, we use the filtered MRR in which all other valid triples (i.e., the ones in T_G) are removed from the answer list (Kazemi and Poole 2018). The MRR ranges from $0 < 1/|E_G|$ to 1, where the minimum means the correct answer was always ranked last and 1 means the correct answer was always ranked first. Note that some earlier prior work used a sampling-based protocol, which is problematic (Kochsiek and Gemulla 2021; Ott, Meilicke, and Stuckenschmidt 2024) and not used here.

2.2 Construction of Benchmark Datasets

In the classical transductive KGC setting, the target graph G is also used as a training graph $G_{\text{train}} = (E_{\text{train}}, R_{\text{train}}, T_{\text{train}})$ for the KGC model. For this reason, knowledge and patterns involving the entities and relations in G can be learned during training. For inductive KGC, which is the focus of this paper, the training graph and target graph are disjoint.[1] In entity-inductive KGC, this means that $E_G \cap E_{\text{train}} = \emptyset$. In entity-and-relation-inductive KGC, it additionally holds that $R_G \cap R_{\text{train}} = \emptyset$. Direct learning of entity knowledge and relation knowledge (in the latter setting) is thus impossible.

To evaluate entity-inductive KGC methods, the available and commonly used benchmark datasets—i.e., FB, WN, NELL (Teru, Denis, and Hamilton 2020) as well as ILPC (Galkin, Berrendorf, and Hoyt 2022)—construct both G and G_{train} from a common source KG $G_{\text{full}} = (E_{\text{full}}, R_{\text{full}}, T_{\text{full}})$. This is done by first selecting a subset $E_S \subseteq E_{\text{full}}$ of the entities—e.g., a randomly sampled set of seed entities and their k-hop neighborhoods—and then using the corresponding induced subgraph $G_S = (E_S, R_{\text{full}}, T_S)$, where $T_S \subseteq T_{\text{full}}$ is given by the set of triples (or a subset thereof) in T_{full} that involve only entities in E_S. Both G and G_{train} are constructed in this fashion, using disjoint subsets. Entity-and-relation-inductive benchmark datasets are constructed similarly.

We argue below that this construction pipeline results in evaluation datasets exhibiting statistical regularities which can be leveraged by simple link prediction approaches.

2.3 Frequency Bias in Benchmark Datasets

We examplarily illustrate our argument using the FB15k-237 dataset (G_{full}; Toutanova and Chen (2015)) and its inductive counterpart (Teru, Denis, and Hamilton 2020). The inductive version consists of four splits (v1–v4), each with a training graph and a target graph with disjoint entity sets.

Consider the *profession* relation of FB15k-237. Queries of form *profession(q,?)* ask for the profession of query entity q. Table 1 shows the frequency of answer entities in the full graph as well as target graphs of each inductive split. As can be seen, there are significant differences between the full KG and the inductive splits. In the full dataset, the most frequent profession appears in 20.7% of all *profession*-triples, whereas all other professions appear in less than 10% of the triples. In the inductive splits, however, the most frequent profession has a frequency between 35% and 66.2%, and the second most frequent profession still appears with a frequency between 18.1% and 35.9%. Moreover, individuals that may have a profession in the original dataset may not have a profession in the inductive splits. This phenomenon arises due to the disjointness constraint, i.e., the professions of these individuals may not occur in the entity set of the inductive split.

profession$(q, ?)$	full	inductive v1	v2	v3	v4
Actor	20.7	-	-	66.2	-
Screenwriter	7.9	-	-	-	-
Film Producer	7.8	-	-	-	36.9
Film Director	5.5	-	-	18.1	27.4
Musician	5.2	-	35.0	-	-
Writer	3.9	-	27.1	-	-
Record Producer	2.7	-	20.3	-	-
Singer-Songwriter	2.6	38.4	-	-	-
Comedian	2.5	35.9	-	7.7	12.6
Journalist	0.5	8.9	-	-	-
Pop. ranking MRR	43.2	58.3	48.2	84.8	66.9

Table 1: Common professions in the full FB15k-237 dataset and its inductive splits (in %). The bottom row shows the MRR (in %) obtained by ranking answer triples solely by the frequency of the answer entity. For brevity, we only show professions that are among the top-3 most frequent professions in any target graph.

This frequency bias facilitates high predictive quality for any model that is aware of it. At the bottom of Table 1, we list the MRR obtained by popularity ranking on the subset of the *profession(q,?)* queries from the test set. This naive approach simply ranks answer entities by how often they co-occur with the query relation, completely ignoring anything known about the query entity q.[2] On the full dataset, popularity ranking already achieves a decent MRR of 43.2%. On the inductive splits, where frequency skew is higher, it is possible to achieve a substantially higher and sometimes even very good MRR.

This example illustrates that although the frequency distribution is already helpful on the full dataset, it is substantially more informative on the inductive splits. We will show below that a simple baseline which does not use any complex relational patterns beyond target graph frequencies and co-ocurrence statistics is competitive with the state of the art (SOTA) on the commonly used benchmarks.

3 A Simple Baseline

Our baseline computes various summary statistics of the target graph and encodes them as a set of simple rules, each

[1] Pretrained KGC models use additional pretraining graphs, which are also disjoint from the target graph.

[2] E.g., for queries of form *profession(q,?)* on split v3 of Tab. 1, pop. ranking predicts: 1. Actor, 2. Film Director, 3. Comedian, ...

	WN_e				FB_e				$NELL_e$				ILPC	
	v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4	small	large
Simple baseline	64.6	62.3	37.7	58.9	46.7	51.6	49.0	47.4	81.0	61.0	57.7	54.9	27.9	29.2
ULTRA-0	64.8	66.3	37.6	61.1	49.8	51.2	49.1	48.6	78.5	52.6	51.5	47.9	30.2	29.0
ULTRA-F	68.5	67.9	41.1	61.4	50.9	52.4	50.4	49.6	75.7	57.5	56.3	46.9	30.3	30.8
Prior SOTA	74.1	70.4	45.2	66.1	45.7	51.0	47.6	46.6	63.7	41.9	43.6	36.3	13.0	7.0

Table 2: MRR (in %) for entity-inductive KGC, where relations (but not entities) are shared across training and target graph

	$FB_{e,r}$				$WK_{e,r}$				$NL_{e,r}$				
	25	50	75	100	25	50	75	100	0	25	50	75	100
Simple baseline	35.9	30.7	37.5	40.6	28.0	15.9	37.9	18.6	33.1	34.9	37.6	30.0	47.4
ULTRA-0	38.8	33.8	40.3	44.9	31.6	16.6	36.5	16.4	34.2	39.5	40.7	36.8	47.1
ULTRA-F	38.3	33.4	40.0	44.4	32.1	14.0	38.0	16.8	32.9	40.7	41.8	37.4	45.8
Prior SOTA	13.3	11.7	18.9	22.3	18.6	06.8	24.7	10.7	26.9	33.4	28.1	26.1	30.9

Table 3: MRR (in %) for entity-and-relation-inductive KGC, where a fraction of the relations (but not entities) are shared across training and target graph

associated with a confidence. In particular, we consider the following rule patterns as well as their variants obtained by reordering terms in subject/object position:

$$r(X, c) \leftarrow \quad \text{(P1)}$$
$$r(X, c) \leftarrow \exists Y.\, s(X, Y) \wedge r \neq s \quad \text{(P2)}$$
$$r(X, c) \leftarrow s(X, d) \quad \text{(P3)}$$
$$r(X, Y) \leftarrow s(X, Y) \quad \text{(P4)}$$

Given a target graph G, we instantiate these patterns by substituting the constants c and d as well as the relations r and s with all entities from E_G and relations from R_G, respectively. The following example rules (one for each pattern) have been obtained from the target graph of the v3 split of FB15k-237 (cf. Tab. 1). Constant *visual* refers to an award for best special visual effects.

$$66.2\%\ \ \textit{profession}(X, \textit{actor}) \leftarrow \quad \text{(R1)}$$
$$33.3\%\ \ \textit{profession}(X, \textit{filmdirector}) \leftarrow \exists Y\ \textit{film\_by}(Y, X) \quad \text{(R2)}$$
$$26.3\%\ \ \textit{film\_role}(X, \textit{makeup\_artist}) \leftarrow \textit{award}(X, \textit{visual}) \quad \text{(R3)}$$
$$66.6\%\ \ \textit{friends}(X, Y) \leftarrow \textit{friends}(Y, X) \quad \text{(R4)}$$

For example, R3 states that any person who won a visual-effects award is a makeup artist. As this is clearly not always the case, each rule is associated with a confidence which states "how often" the head is true (person is a makeup artist) given that the body is true (person won visual-effects award).

More precisely, the confidence of a rule is given by the number of its correct predictions divided by the number of all its predictions w.r.t. T_G. Pattern P1 corresponds to the naive popularity ranking approach described in Section 2.3. The confidence of a corresponding rule (e.g., R1 with $r = \textit{profession}$ and $c = \textit{actor}$) is given by the fraction of r-triples in T_G that have tail entity c (e.g., 66.2% of all professions are actors, cf. Tab. 1). For Patterns P2 and P3, we count over X-values (taking values in E_G). For example, the confidence of R3 states that 26.3% of the persons who won a visual-effects award (X-values) were makeup artists. For Pattern P4, we count over pairs of (X, Y)-values (taking values in $E_G \times E_G$). For example, the confidence of R4 states that when Y is a friend of X, the friendship is mutual in 66.6% of the cases. Note that all counts can be obtained by iterating once over the adjacency list of the target graph; they constitute simple frequency statistics.

Let $r(q, ?)$ be a query. Our baseline then considers all rules in which the substitution of X (likewise Y for P4) by an entity $e \in E_G$ would result in a rule head of form $r(q, e)$. We score answer $r(q, e)$ by the highest confidence of such a rule that fires, i.e., where the body is also true. Note that the score of an answer ultimately depends on the presence of only a single triple in the target graph (the one in the body) and is computed by solely considering the 1-hop neighborhood of the query entity q. The baseline thus does not use any relational knowledge beyond one-hop neighbourhood statistics.

Note that we do not consider this baseline a competing or even a suitable method for inductive KGC.[3] Instead, our goal is to understand to what extent existing results provide evidence for the transfer of relational knowledge in inductive KGC. The intended role of the target graph in inductive KGC benchmarks is to provide descriptions of the subject and objects involved in a test triple in order to support relational inference. However, the problematic construction pipeline produces target graphs that themselves contain a rich set of simple statistical dependencies *not present* in the full dataset G_{full}.

4 Experiments

In this section, we revisit the state-of-the-art results for inductive KGC reported by Galkin et al. (2024) and compare them with the results obtained by our simple baseline. As our baseline does not perform any relational reasoning, good

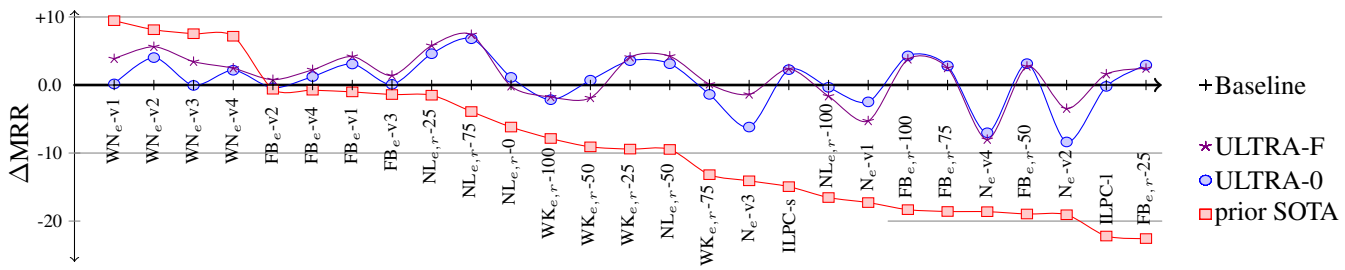

Figure 1: Absolute differences in percentage points w.r.t. baseline MRR of ULTRA-0, ULTRA-F, and the prior SOTA. N is used as abbreviation for NELL.

inductive KGC methods should clearly outperform it. Our results show that this is not the case on commonly used benchmark datasets, however.

For entity-inductive KGC, we considered the WN_e, FB_e, and NELL_e benchmarks proposed by Teru, Denis, and Hamilton (2020), which were constructed from the full datasets WN18RR, FB15k-237, and NELL-995 (Dettmers et al. 2018; Toutanova and Chen 2015; Xiong, Hoang, and Wang 2017), respectively. Each benchmark has four different splits v1–v4. We also included the newer and larger ILPC benchmark of Galkin, Berrendorf, and Hoyt (2022). For entity-and-relation-inductive KGC, we used the $\text{FB}_{e,r}$, $\text{NL}_{e,r}$ and $\text{WK}_{e,r}$ benchmarks of Lee, Chung, and Whang (2023), which were obtained from the FB15k-237, NELL-995, and Wikidata68K (Gesese, Sack, and Alam 2022) full datasets and have five splits denoted 0, 25, 50, 75 and 100. The pure entity-and-relation-inductive setting is referred to as 0; other settings (e.g., 25) allow for partial overlap of relations in the training and target graph (e.g., 25% of the triples in the target graph have relations that also occur in the training graph).

We computed the results for our baseline using AnyBURL (Meilicke et al. 2024), which we ran in a restricted setting on the target graph. [4] This setting ensures that only the rule patterns described in Section 3 are instantiated. We also report results for ULTRA (Galkin et al. 2024), a state-of-the-art inductive KGC model. ULTRA first extracts a relation graph from the target graph, which encodes the presence of length-2 relational paths. Then it applies a pretrained NBFNet model (Zhu et al. 2021) on the relation graph, and finally runs another pretrained NBFNet model on the target graph. We give results for both the zero-shot version (ULTRA-0) and the fine-tuned version (ULTRA-F, fined-tuned on training graph) as well as from prior work; all numbers are taken from Galkin et al. (2024). The prior SOTA results were obtained with the models A*Net (Zhu et al. 2023), NBFNet (Zhu et al. 2021), RED-GNN (Zhang and Yao 2022), Nodepiece (Galkin et al. 2022), and InGram (Lee, Chung, and Whang 2023).

Our results are shown in Tables 2 and 3. With some exceptions, the MRRs of the baseline and ULTRA-0 are surprisingly similar, while prior SOTA shows a significant deviation. This can be more easily seen in Figure 1, where we show the absolute difference in MRR (in pp.) of ULTRA-0, ULTRA-F, and prior SOTA over the simple baseline (shown as the "zero line"). Positive values indicate better-than-baseline results, negative values indicate worse.

Overall, we conducted experiments on 27 different splits. For eight of the splits, the simple baseline obtained an MRR higher than 0.5. In fact, in five of these splits, the simple baseline achieved the best overall result, outperforming even the prior SOTA and ULTRA-F (which was additionally fine-tuned on the target graph's training split). In 23 splits, the simple baseline outperformed the SOTA prior to ULTRA. The good performance of this simple baseline, which does not conduct any complex relational reasoning, raises doubts about the suitability of current benchmarks for evaluating knowledge transfer in inductive KGC.

The perhaps simplest potential explanation for these observations is that both the simple baseline and ULTRA-0 use the same information source for their predictions: the frequency skew in the target graph. Nevertheless, our baseline performs sometimes better and sometimes worse. This might be caused by the fact that the simple baseline only considers the maximum-confidence rule, whereas ULTRA-0 learned how to aggregate messages within the GNN architecture during pretraining. This may sometimes be helpful, sometimes detrimental. Another important difference has been described in Betz et al. (2024), where the authors argue that NBFNet, which is used by ULTRA, can learn negative patterns. These cannot be captured by rule patterns such as the ones used in our baseline.

For most of the splits, the simple baseline performs clearly better than what has been reported as (prior) state of the art models in the ULTRA paper. This perhaps surprising result may have multiple reasons. First, the prior SOTA models are apparently incapable to leverage key information from the target graph, including frequency skew or simple patterns. Second, most of these models have been

[3] As our baseline accesses the target graph to perform counting, it may appear debatable whether the approach is in line with the inductive evaluation protocol. All inductive KGC methods access the target graph during prediction, however, typically by running a GNN (Zhu et al. 2021). Moreover, methods as ULTRA (Galkin et al. 2024) also obtain summary statistics from the target graph.

[4] Confidence computation and all predictions jointly took less than 6 minutes per split on a standard office laptop (CPU only); for WN_e, FB_e, and NELL_e less than 1 minute.

developed for or appear more suited for a transductive or semi-inductive setting, where the entities in the training and target graph overlap and direct knowledge transfer is possible. Finally, in the fully inductive setting, only "entity-independent" relational knowledge (such as rule R4, but unlike rule R2) can potentially be transferred from the training graph to the target graph. This appears to be rather restrictive.

5 Conclusion

Inductive KGC methods aim to exploit relational knowledge obtained from pretraining or training datasets on new KGs with entirely new sets of entities or relations. Key motivations are the evolving nature of knowledge graphs—i.e., new, previously unseen entities or relations may be added—and the relational knowledge transfer across related domains (e.g., from a geographical dataset about the US to a dataset about Europe).

We argued that there is currently insufficient evidence that relational knowledge transfer indeed takes place. First, the strict disjointness principle used to construct current benchmark datasets leads to frequency biases. Second, a simple baseline approach based on one-hop neighborhood statistics but without any more sophisticated relational reasoning obtained competitive results in our experimental study. Future work should include simple baseline approaches and consider improved benchmarks. In particular, the strict disjointness principle should be rethought.

AI Declaration

The authors have not employed any Generative AI tools.

References

Betz, P.; Stelzner, N.; Meilicke, C.; Stuckenschmidt, H.; and Bartelt, C. 2024. A*Net and NBFNet learn negative patterns on knowledge graphs. *arXiv preprint arXiv:2412.05114*.

Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *AAAI Conference on Artificial Intelligence*.

Galkin, M.; Berrendorf, M.; and Hoyt, C. T. 2022. An open challenge for inductive link prediction on knowledge graphs. In *Workshop on Graph Learning Benchmarks at the International World Wide Web Conference*.

Galkin, M.; Denis, E.; Wu, J.; and Hamilton, W. L. 2022. Nodepiece: Compositional and parameter-efficient representations of large knowledge graphs. In *International Conference on Learning Representations*.

Galkin, M.; Yuan, X.; Mostafa, H.; Tang, J.; and Zhu, Z. 2024. Towards foundation models for knowledge graph reasoning. In *International Conference on Learning Representations*.

Gesese, G. A.; Sack, H.; and Alam, M. 2022. Raild: Towards leveraging relation features for inductive link prediction in knowledge graphs. In *International Joint Conference on Knowledge Graphs*.

Kazemi, S. M., and Poole, D. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems*.

Kochsiek, A., and Gemulla, R. 2021. Parallel training of knowledge graph embedding models: a comparison of techniques. *Proceedings of the VLDB Endowment* 15(3):633–645.

Lee, J.; Chung, C.; and Whang, J. J. 2023. InGram: Inductive knowledge graph embedding via relation graphs. In *International Conference on Machine Learning*.

Meilicke, C.; Chekol, M. W.; Betz, P.; Fink, M.; and Stuckenschmidt, H. 2024. Anytime bottom-up rule learning for large-scale knowledge graph completion. *VLDB Journal* 33(1):131–161.

Ott, S.; Meilicke, C.; and Stuckenschmidt, H. 2024. Reevaluation of inductive link prediction. In *8th International Joint Conference on Rules and Reasoning*.

Pezeshkpour, P.; Tian, Y.; and Singh, S. 2020. Revisiting evaluation of knowledge base completion models. In *Automated Knowledge Base Construction*.

Teru, K.; Denis, E.; and Hamilton, W. 2020. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning*.

Toutanova, K., and Chen, D. 2015. Observed versus latent features for knowledge base and text inference. In *Third Workshop on Continuous Vector Space Models and their Compositionality*.

Voorhees, E., and Tice, D. 1999. The trec-8 question answering track evaluation. In *The Eighth Text Retrieval Conference*, 82.

Xiong, W.; Hoang, T.; and Wang, W. Y. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y., and Yao, Q. 2022. Knowledge graph reasoning with relational digraph. In *ACM Web Conference*.

Zhu, Z.; Zhang, Z.; Xhonneux, L.-P.; and Tang, J. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems*.

Zhu, Z.; Yuan, X.; Galkin, M.; Xhonneux, L.-P.; Zhang, M.; Gazeau, M.; and Tang, J. 2023. A*Net: A scalable path-based reasoning approach for knowledge graphs. In *Advances in Neural Information Processing Systems*.