A Literature Review

Table 8 lists existing KGC datasets. We can roughly classify them into two groups: inferential and non-inferential datasets. The first group are usually manually curated to ensure each testing sample can be inferred from training data through reasoning paths. Families (Garcia-Duran et al., 2015) test family relationships including cousin, ancestor, marriage, parent, sibling, and uncle, among the members of 5 families along 6 generations. Such that there are obvious compositional relationships like uncle $\approx$ sibling + parent or parent $\approx$ married + parent. Kinship (Kemp et al., 2006) contains kinship relationships among members of the Alyawarra tribe from Central Australia, while Country (Bouchard et al., 2015) contains countries, regions, and subregions as entities and is carefully designed to explicitly test the location relationship (i.e., locatedIn and neighbor) among them. The above datasets are clearly limited in scale and inference patterns, thus become not challenging. HOLE (Nickel et al., 2016) even achieves 99.7% ACU-PR on dataset Country (Bouchard et al., 2015).

The second group of datasets are automatically derived from public KGs and randomly split positive triples into train/valid/test, leading to a risk of testing samples non-inferential from training data. FB13 (Socher et al., 2013) and FB15K (Bordes et al., 2013) are commonly used benchmark from FreeBase. FB15k401 (Yang et al., 2014) is a subset of FB15k containing only frequent relations (relations with at least 100 training examples). To remove test leakage, FB15k-237 (Toutanova and Chen, 2015) removes all equivalent or inverse relations. Similarly, FB5M (Wang et al., 2014) removes all the entity pairs that appear in the testing set. WN18RR (Dettmers et al., 2018) is the challenging version of WN18 (Bordes et al., 2013) extracted from WordNet. Textual information is also included for specific task, such as FB40K (Lin et al., 2015) targeting relation extraction dataset New York Times (Riedel et al., 2010). FB24K (Lin et al., 2016) introduce Attributes. FB15K+ (Xie et al., 2016) introduce types and make fb15k more sparse by only filterring out relation with a frequency lower than one. Another popular knowledge source is YAGO, and the corresponding datasets include YAGO3-10 (Dettmers et al., 2018) and YAGO37 (Guo et al., 2018). Except for open-domain KG, NELL (Wang et al., 2015) concentrates on location and sports, and UMLS (Kok and Domingos, 2007) targets medical knowledge. CoDEx (Safavi and Koutra, 2020) argues the quality of the above benchmarks, such as NELL995 (Xiong et al., 2017) are nonsensical or overly generic. Thus they propose a comprehensive dataset consisting of three knowledge graphs varying in size and structure, entity types, multilingual labels and descriptions, and hard negatives.

B Annotation Guideline

We provide the following annotation guidelines for annotators to label inferred triples in Section 3.4.

Task This is a two-step annotations. First, you must annotate each triple with the label $y \in \{1, -1\}$, where 1 denotes that the triple is correct and -1 denotes that the triple is incorrect. You can find the answer from anywhere you want, such as commonsense, Wikipedia, and professional websites. If you cannot find any evidence to support the statement, you shall choose label -1. Second, you must annotate each incorrect triple with the label $\hat{y} \in \{0, -1\}$, where 0 denotes that you do not know the answer. Now, you can find the answer from our provided triples. If you cannot find any evidence to support the statement, you shall choose label 0.

Examples Here are some examples judged using three types of knowledge sources.

- **Commonsense**: (Cypriot Fourth Division, hasPart, 2018–19 Cypriot Third Division) is clearly incorrect, since the fourth division cannot has a part of third division.
- **Professional websites**: To annotate the triple (Bahrain-Merida 2019, hasPart, Carlos Betancur), you may search the person in professional websites, such as `https://www.procyclingstats.com/team/bahrain-merida-2019`. Since there is no Carlos Betancur listed in that website, please choose false.
- **Wikipedia**: Given the triples (Tōkaidō Shinkansen, connectsWith, Osaka Higashi Line) and (Tōkaidō Shinkansen, connectsWith, San'yō Main Line), you can find related station information from the page of Tōkaidō Shinkansen. You can find that Osaka Higashi Line shares a transfer station with Tōkaidō Shinkansen, thus label it with 1.

Datasets	source	#Entity	#Relation	#Triples (train/valid/test)
FB13 (Socher et al., 2013)	FreeBase	75,043	13	316,232/5,908/23,733
FB15k (Bordes et al., 2013)	FreeBase	14,951	1,345	483,142/50,000/59,071
FB15k237 (Toutanova and Chen, 2015)	FreeBase	14,541	237	272,115/17,535/20,466
FB15k+ (Xie et al., 2016)	FreeBase	14,951	1,855	486,446/50,000/62,374
FB15k401 (Yang et al., 2014)	FreeBase	14,541	401	560,209/-/-
FB24k (Lin et al., 2016)	FreeBase	23,634	987	402,493/-/21,067
FB40k (Lin et al., 2015)	FreeBase	39,528	1,336	370,648/67,946/96,678
FB5M (Wang et al., 2014)	FreeBase	5,385,322	1,192	19,193,556/50,000/59,071
WN11 (Socher et al., 2013)	WordNet	38,696	11	112,581/2,609/10,544
WN18 (Bordes et al., 2013)	WordNet	40,943	18	141,442/5,000/5,000
WN18RR (Dettmers et al., 2018)	WordNet	40,943	11	86,835/3,034/3,134
YAGO3-10 (Dettmers et al., 2018)	YAGO	123,182	37	1,079,040/5,000/5,000
YAGO37 (Guo et al., 2018)	YAGO	123,189	37	989,132/50,000/50,000
CoDEx (Safavi and Koutra, 2020)	Wikidata	77,951	69	551,193/30,622/30,622
NELL995 (Xiong et al., 2017)	NELL	75,492	200	154,213/-/-
NELL$_{loc}$ (Wang et al., 2015)	NELL	672	10	941/-/-
Family (Garcia-Duran et al., 2015)	Artificial	721	7	8,461/2,820/2,821
Kinship (Kemp et al., 2006)	Artificial	104	26	8,548/2,820/2,821
Countries (Bouchard et al., 2015)	Artificial	272	2	1,111/24/24
UMLS (Kok and Domingos, 2007)	UMLS	135	49	5,216/-/-

Table 8: An overview of Knowledge Graph Completion Datasets.

And, San'yō Main Line doesn't show up in the page, you may label it with -1.

C Relation Patterns

InferWiki is able to analyze relation patterns for each path, including symmetry, inversion, hierarchy, and composition, where detailed explanations and examples are listed in Table 9.

D Relation Types

We illustrate the most frequent relation types and their distribution of InferWiki64k and InferWiki16k in Figure 8.

E Comparison with Existing Datasets

Figure 9 shows the distribution of entities and their neighbors as compared to widely used datasets: FB15k237 and CoDEx-m.

F Experiment Setup

Our experiments are run on the server with the following configurations: OS of Ubuntu 16.04.6 LTS, CPU of Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, and GPU of GeForce RTX 2080 Ti. We use OpenKE[5] for re-implementing TransE, ComplEx, and RotatE. For the rest models, we use the original codes for ConvE[6], TuckER [7], Multi-hop[8], and AnyBURL[9]. Because we utilize various types of KGC models including embedding-based, multi-hop reasoning (reinforcement learning), and rule-based models, these models largely have their own hyperparameters. To avoid exhaustive parameter search in a large range, we conduct a series of preliminary experiments and find that the suggested parameters work well on Wikidata-based data. We then search the embedding size in the range of $\{256, 512\}$, number of negative samples in the range of $\{15, 25\}$ and margin in the range of $\{4, 8\}$. The optimal parameters of each model on all of three datasets are listed in Table 10. The thresholds in triples classification are listed in Table 11

[5] https://github.com/thunlp/OpenKE

[6] https://github.com/TimDettmers/ConvE

[7] https://github.com/ibalazevic/TuckER

[8] https://github.com/salesforce/MultiHopKG

[9] http://web.informatik.uni-mannheim.de/AnyBURL/

Pattern	Notation	Example
symmetry	$r_1(x, y) \Rightarrow r_1(y, x)$	(Prince Christopher$_{Q44775}$, partner, Friederike$_{Q93614}$) $\Rightarrow$ (Friederike, partner, Prince Christopher)
inversion	$r_1(x, y) \Leftrightarrow r_2(y, x)$	(Amravati district$_{Q1771774}$, capital, Amravati$_{Q269899}$) $\Rightarrow$ (Amravati, capitalOf, Amravati district)
hierarchy	$r_1(x, y) \Rightarrow r_2(y, x)$	(Superman$_{Q79015}$, derivativeWork, Superman Returns$_{Q328695}$) $\Rightarrow$ (Superman, presentInWork, Superman Returns)
composition	$r_1(x, y) \wedge \cdots \wedge r_p(y, z) \Rightarrow r_{p+1}(x, z)$	(Eleanor$_{Q156045}$, mother, Joanna$_{Q171136}$) $\wedge$ (Ferdinand I$_{Q150611}$, mother, Joanna) $\wedge$ (Isabella$_{Q157884}$, sibling, Ferdinand I) $\Rightarrow$ (Eleanor, sibling, Isabella)

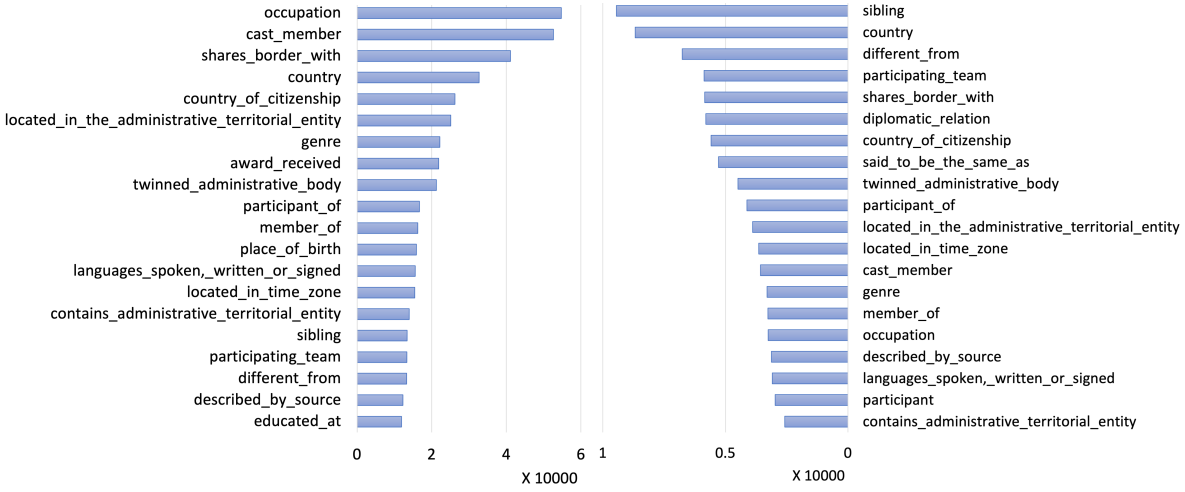
Table 9: Explanations and examples for various relation patterns.

Hyperparameter	TransE	ComplEx	RotatE	ConvE	TuckER	Multihop
InferWiki16k						
Embedding Size	256	512	512	512	512	256
# Negatives	15	25	25	-	-	-
Margin	4	4	8	-	-	-
Learning Rate	1.0	0.5	2e-5	1e-4	1e-4	1e-3
Optimizer	SGD	adagrad	adam	adam	adam	-
Batch Size	1,625	1,625	2,000	256	256	128
InferWiki64k						
Embedding Size	256	512	512	256	512	256
# Negatives	15	15	25	-	-	-
Margin	4	4	8	-	-	-
Learning Rate	1.0	0.5	2e-5	1e-4	1e-4	1e-3
Optimizer	SGD	adagrad	adam	adam	adam	-
Batch Size	7,823	7,823	2,000	256	256	128
CoDEx-m-infer						
Embedding Size	512	256	512	256	512	256
# Negatives	25	25	25	-	-	-
Margin	8	4	4	-	-	-
Learning Rate	1.0	0.5	2e-5	1e-4	1e-4	1e-3
Optimizer	SGD	adagrad	adam	adam	adam	-
Batch Size	1,856	1,856	2000	256	256	128

Table 10: Best hyperparameter configurations.

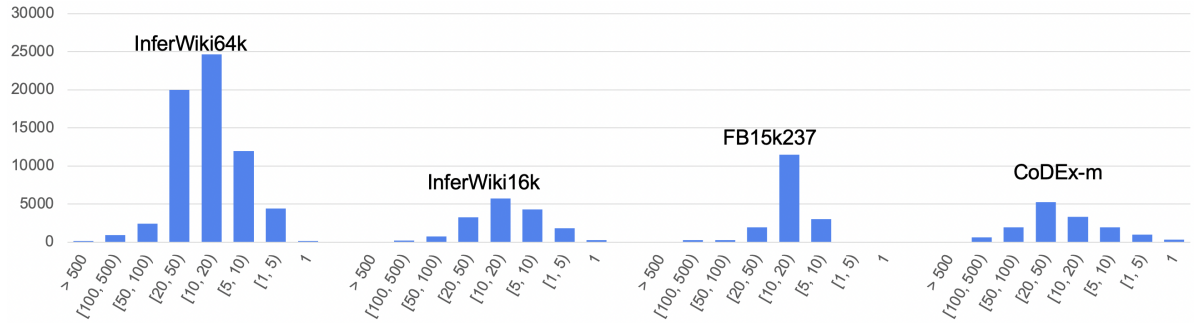
	InferWiki	TransE	ComplEx	RotatE	ConvE	TuckER
Closed World	64k	[-24.4663, -9.0235] -16.7449	[-43.0342, 30.6942] -0.2717	[-15.7235, 7.8291] -0.6498	[0.0, 0.9999] 0.1	[0.0, 0.9982] 0.01
	16k	[-24.0588, -4.333] -13.4069	[-21.5906, 24.7742] 2.5191	[-21.2362, 7.8282] -0.6005	[0.0, 1.0] 0.19	[0.0, 0.9734] 0.0097
Open World	16k	[-24.0588, -4.333] -16.1685, -11.8288	[-21.5906, 24.7742] -3.5084, 3.4464	[-21.2362, 7.8282] -2.3444, 0.8527	[0.0, 1.0] 0.01, 0.37	[0.0, 0.9734] 0.0097, 0.0389

Table 11: Best thresholds in triple classification, where the upper side is the search range and the lower side is the best values. They are searched on validation.

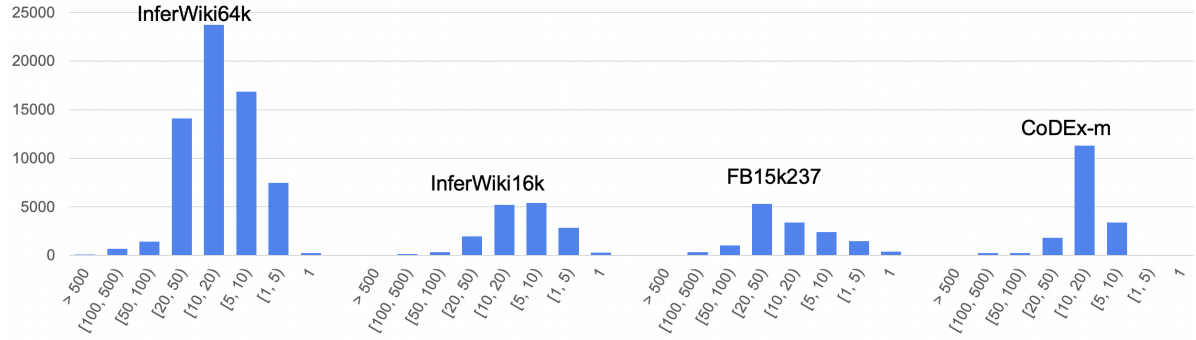


(a) InferWiki64k. (b) InferWiki16k.

Figure 8: Distribution of most frequent relation types.



(a) Distribution of entities, where x-axis denotes different ranges regarding entity frequency in the train set.



(b) Distribution of entity neighbors, where x-axis denotes ranges regarding average number of neighbors in the train set.

Figure 9: Distribution of entities and their neighbors.