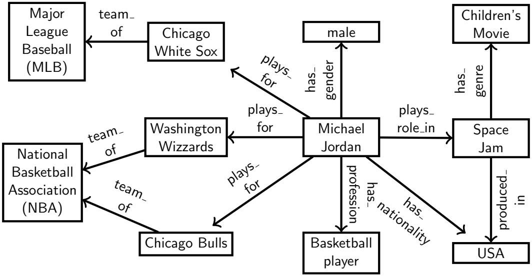
Debate Dynamics for Human-comprehensible Fact-checking on Knowledge Graphs

Marcel Hildebrandt, Jorge Quintero Serna, Yunpu Ma, Martin Ringsquandl, Mitchell Joblin, Volker Tresp

Siemens Corporate Technology, Germany
University of Munich (LMU), Germany

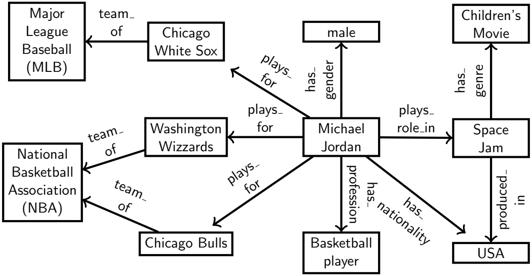
November 5, 2019

Knowledge Graphs



- Information about the real world can be expressed in terms of entities and their relations
- Knowledge graphs (KG) store facts about the world in terms of triples (s, p, o), where s (subject) and o (object) correspond to nodes in the graph and p (predicate) denotes the edge type connecting them.

Knowledge Graphs

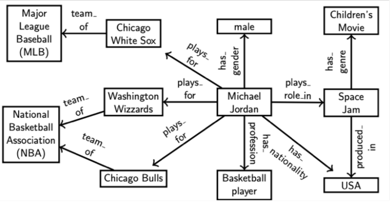


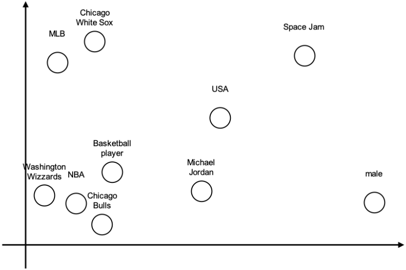
Definition: Knowledge Graphs (KG)

Let $\mathcal{E}$ denote the set of entities and $\mathcal{R}$ the set of binary relations. A knowledge graph $\mathcal{KG} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a collection of facts stored as triples as (s, p, o). To indicate whether a triple is true or false, we consider the characteristic function $\phi : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \{0, 1\}$. For all $(s, p, o) \in \mathcal{KG}$, we assume $\phi(s, p, o) = 1$ (i.e., KGs are collections of true facts).

KG Embeddings

Embed entities and relations into a continuous vector space while preserving structural similarities/dissimilarities.

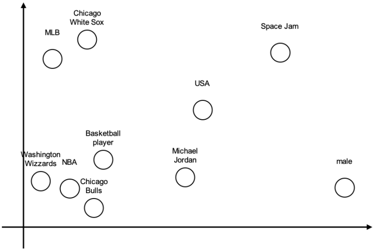




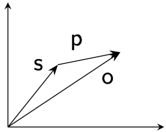
Canonical Machine Learning Tasks on KGs

Definition: Triple Classification

Given a triple $(s, p, o) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, triple classification is concerned with predicting the truth value $\phi(s, p, o)$.

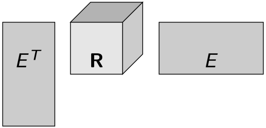


Categories of KG Embeddings

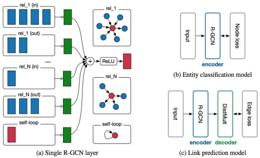


Embedding Space

Translational methods: TransE [1], TransR [5], RotatE [9], . . .



Factorization methods: RESCAL [6], ComplEx [11], DistMult [13], SimplE [4], . . .



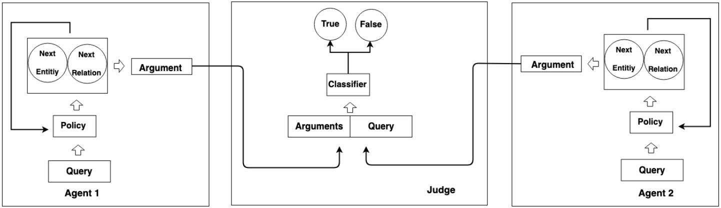
Neural Networks /graph convolutions: ConvE [3], R-GCN [7], NTN [8] . . .

→ All these methods have produce a scores for the plausiblity/likelihood of triples but it remains hidden what contributed to the scores.

R2D2: Triple Classification Based on Debates (I)

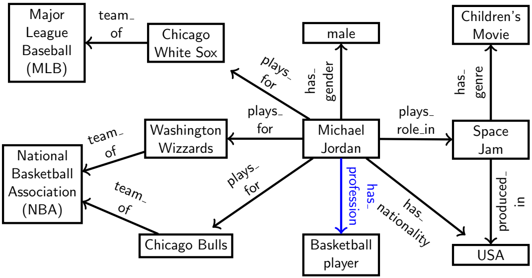
- Query statement: $(s, p, o) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$
- Two competing agents: Agent 1 argues that the query statement is true (thesis); Agent 2 argues that it is false (antithesis).
- Judge decides with agent to believe, i.e. whether the statement was truthful

R2D2: Triple Classification Based on Debates (II)



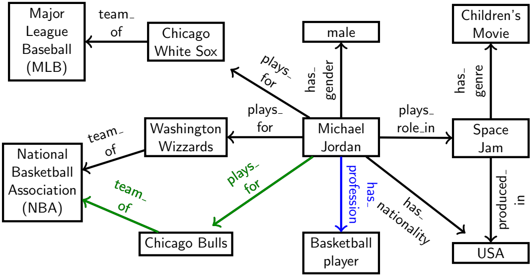
- Arguments are paths with fixed length in the KG.
- The agents are trained through reinforcement learning
- Judge is trained using supervised learning.

Example Debate



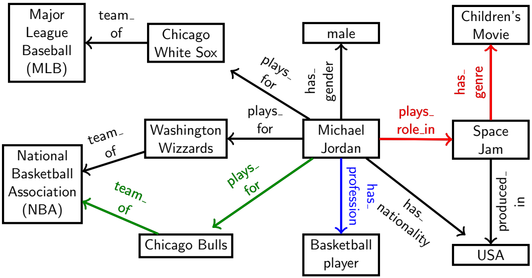
Query: (Michael Jordan, has_profession, basketball player)

Example Debate



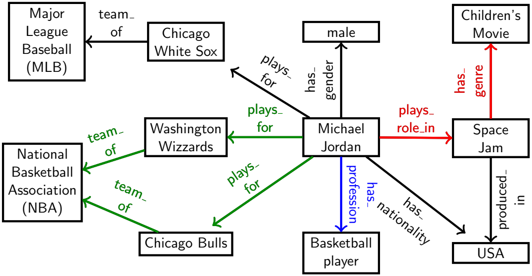
Query: (Michael Jordan, has_profession, basketball player)

Example Debate



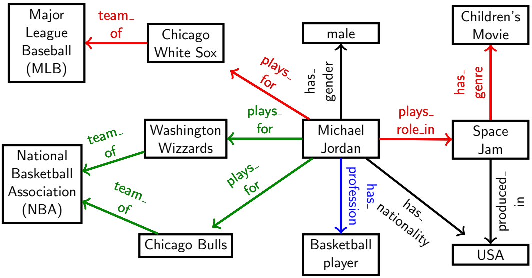
Query: (Michael Jordan, has_profession, basketball player)

Example Debate



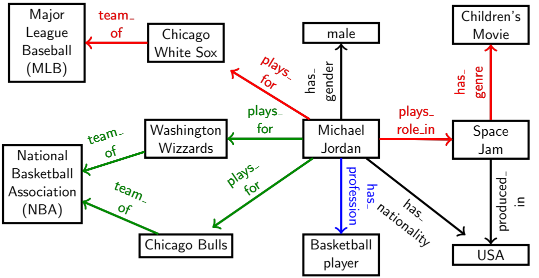
Query: (Michael Jordan, has_profession, basketball player)

Example Debate



Query: (Michael Jordan, has_profession, basketball player)

Example Debate



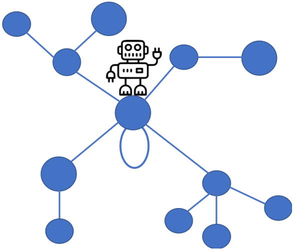
Query: (Michael Jordan, has_profession, basketball player)

Agent 1: (Michael Jordan, plays_for, Chicago Bulls)
$\wedge$ (Chicago Bulls, team_of, NBA)
Agent 2: (Michael Jordan, plays_role_in,Space Jam)
$\wedge$ (Space Jam, has_genre, Children's Movie)
Agent 1: (Michael Jordan, plays_for, Washington Wizzards)
$\wedge$ (Washington Wizzards, team_of, NBA)
Agent 2: (Michael Jordan, plays_for, Chicago White Sox)
$\wedge$(Chicago White Sox, team_of, MLB)

Judge: Query is true.

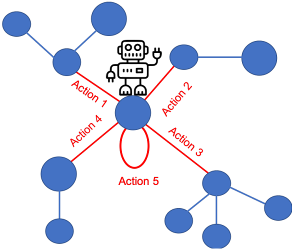
States

- The fully observable state space $\mathcal{S}$ for each agent is given by $\mathcal{E}^2 \times \mathcal{R} \times \mathcal{E}$. Intuitively, we want the state to encode the query triple $q = (s_q, p_q, o_q)$ and the location of exploration $e_t^{(i)}$ (i.e., the current location) of agent $i \in \{1, 2\}$ at time t. Thus, a state $S_t^{(i)} \in \mathcal{S}$ for time $t \in \mathbb{N}$ is represented by $S_t^{(i)} = \left(e_t^{(i)}, q\right)$.



Actions

- The set of possible actions for agent i from a state $S_t^{(i)} = \left(e_t^{(i)}, q\right)$ is denoted by $\mathcal{A}_{S_t^{(i)}}$. It consists of all outgoing edges from the vertex $e_t^{(i)}$ and the corresponding target nodes. More formally, $\mathcal{A}_{S_t^{(i)}} = \left\{(r, e) \in \mathcal{R} \times \mathcal{E} : S_t^{(i)} = \left(e_t^{(i)}, q\right) \wedge \left(e_t^{(i)}, r, e\right) \in \mathcal{KG}\right\}$. Moreover, we denote with $A_t^{(i)} \in \mathcal{A}_{S_t^{(i)}}$ the action that agent i performed at time t.

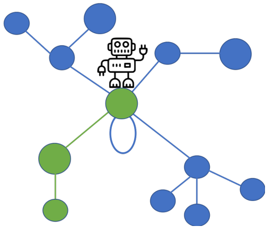


History

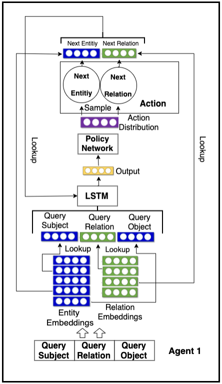
- We denote the history of agent i up to time t with the tuple $H_t^{(i)} = \left(H_{t-1}^{(i)}, A_{t-1}^{(i)}\right)$ for $t \geq 1$ and $H_0^{(i)} = (s_q, p_q, o_q)$.
- The agents encode their histories via an LSTM

$$\boldsymbol{h}_t^{(i)} = \mathrm{LSTM}^{(i)}\left(\left[\boldsymbol{a}_{t-1}^{(i)}, \boldsymbol{q}^{(i)}\right]\right) \tag{1}$$

where $\boldsymbol{a}_{t-1}^{(i)} = \left[\boldsymbol{r}_{t-1}^{(i)}, \boldsymbol{e}_{t-1}^{(i)}\right] \in \mathbb{R}^{2d}$ corresponds to embedding of the action at $t-1$ with $\boldsymbol{r}_{t-1}^{(i)}$ and $\boldsymbol{e}_{t-1}^{(i)}$ denoting the embeddings of the relation and the target entity into $\mathbb{R}^d$.



Policies



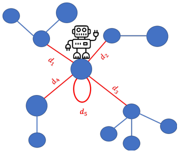
- The action distributions of each agent are given by

$$\boldsymbol{d}_t^{(i)} = \text{softmax}\left(\boldsymbol{A}_t^{(i)}\left(\boldsymbol{W}_2^{(i)}\text{ReLU}\left(\boldsymbol{W}_1^{(i)}\boldsymbol{h}_t^{(i)}\right)\right)\right), \quad (2)$$

where the rows of $\boldsymbol{A}_t^{(i)} \in \mathbb{R}^{|\mathcal{A}_{S_t^{(i)}}| \times d}$ contain embeddings of all actions.

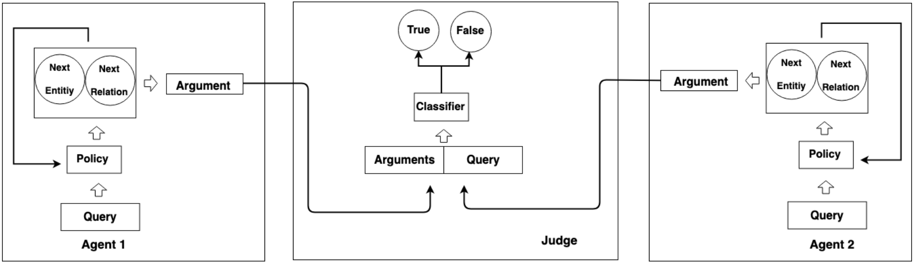
- The action $A_t^{(i)} = (r, e) \in \mathcal{A}_{S_t^{(i)}}$ is drawn according to

$$A_t^{(i)} \sim \text{Categorical}\left(\boldsymbol{d}_t^{(i)}\right). \quad (3)$$

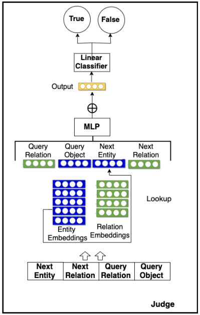


Judge (I)

- The role of the judge in R2D2 is twofold:
  1. The judge is a binary classifier that tries to distinguish between true and false facts.
  2. The judge also evaluates the quality of the arguments extracted by the agents and assigns rewards to them. Thus, the judge also acts as a critic teaching the agents to produce meaningful arguments



Judge (II)



- The judge processes each argument together with the query individually by a feed forward neural network $f : \mathbb{R}^{2(T+1)d} \to \mathbb{R}^d$

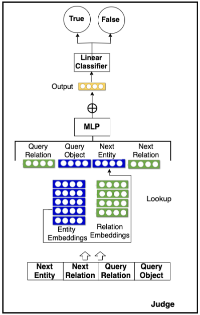
$$\boldsymbol{y}_n^{(i)} = f\left(\left[\boldsymbol{\tau}_n^{(i)}, \boldsymbol{q}^J\right]\right), \tag{4}$$

where $\boldsymbol{\tau}_n^{(i)}$ denotes the embedding of the $n - th$ argument of agent i.

- Then the judge sums the outputs for each argument up and processes the resulting sum by a linear, binary classifier.

$$t_\tau = \sigma\left(\boldsymbol{w}^\intercal \sum_{i=1}^{2} \sum_{n=1}^{N} \boldsymbol{y}_n^{(i)}\right). \tag{5}$$

Judge (III)



- The objective function of the judge for a single query q is given by the cross-entropy loss

$$\mathcal{L}_q = \phi(q) \log t_\tau + (1 - \phi(q)) (1 - \log t_\tau) \,. \qquad (6)$$

Rewards

- In order to generate feedback for the agents, the judge also processes each argument $\tau_n^{(i)}$ individually and produces a score according to

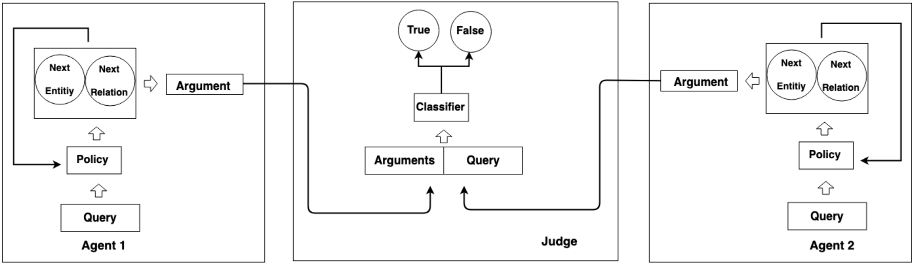
$$t_n^{(i)} = \boldsymbol{w}^\top f\left(\left[\boldsymbol{\tau}_n^{(i)}, \boldsymbol{q}^J\right]\right), \tag{7}$$

- Thus, $t_n^{(i)}$ corresponds to classification score of q solely based on the n-th argument of agent i. Since agent 1 argues for the thesis and agent 2 for the antithesis, the rewards are given by

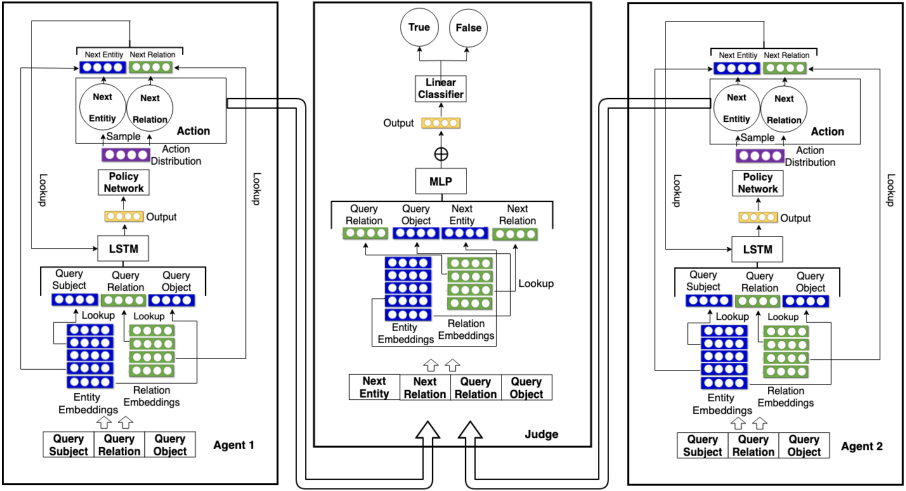
$$R_n^{(i)} = \begin{cases} t_n^{(i)} & \text{if } i = 1 \\ -t_n^{(i)} & \text{otherwise.} \end{cases} \tag{8}$$

- We employ REINFORCE [12] to maximize the expected cumulative rewards

R2D2: Simple Architecture



R2D2: Detailed Architecture



Datasets

- We measure the performance of R2D2 with respect to the triple classification and the KG completion task on the benchmark datasets FB15k-237 [10] and WN18RR [3].

Dataset	Entities	Relations	Triples
FB15k-237 [10]	14,541	237	310,116
WN18RR [3]	40,943	11	93,003

Triple Classification (I)

Dataset	FB15k-237			WN18RR		
Method	Acc	PR AUC	ROC AUC	Acc	PR AUC	ROC AUC
DistMult [13]	0.739	0.78	0.803	0.715	0.815	0.758
ComplEx [11]	0.738	0.789	0.796	**0.802**	**0.887**	**0.860**
TransE [2]	0.673	0.727	0.736	0.676	0.754	0.710
TransR [5]	0.612	0.655	0.651	0.721	0.724	0.792
SimplE [4]	0.703	0.733	0.756	0.722	0.812	0.742
R2D2	**0.751**	**0.86**	**0.848**	0.726	0.821	0.808

Table: The performance on the triple classification task.

Triple Classification (II)

Dataset	FB15k-237			WN18RR		
Method	Acc	PR AUC	ROC AUC	Acc	PR AUC	ROC AUC
DistMult [13]	0.739	0.78	0.803	0.715	0.815	0.758
ComplEx [11]	0.738	0.789	0.796	**0.802**	**0.887**	**0.860**
TransE [2]	0.673	0.727	0.736	0.676	0.754	0.710
TransR [5]	0.612	0.655	0.651	0.721	0.724	0.792
SimplE [4]	0.703	0.733	0.756	0.722	0.812	0.742
R2D2	**0.751**	**0.86**	**0.848**	0.726	0.821	0.808
$\mathrm{R2D2}_{+}$	**0.764**	**0.865**	**0.857**	**0.804**	**0.909**	**0.893**

Table: The performance on the triple classification task.

Example Debates

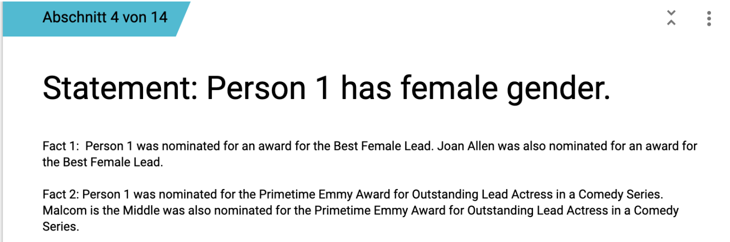
Query:	Richard Feynman $\xrightarrow{nationality}$ USA?	Nelson Mandela $\xrightarrow{hasProfession}$ Actor?
Agent 1:	Richard Feynman $\xrightarrow{livedInLocation}$ Queens $\wedge$ Queens $\xrightarrow{locatedIn}$ USA	Nelson Mandela $\xrightarrow{hasFriend}$ Naomi Campbell Naomi Campbell $\xrightarrow{hasDated}$ Leonardo DiCaprio
Agent 2:	Richard Feynman $\xrightarrow{hasEthnicity}$ Russian people $\wedge$ Russian people $\xrightarrow{geographicDistribution}$ Republic of Tajikistan	Nelson Mandela $\xrightarrow{hasProfession}$ Lawyer $\wedge$ Lawyer $\xrightarrow{specializationOf^{-1}}$ Barrister

Table: Two example debates generated by R2D2: While agent 1 argues that the query is true and agent 2 argues that it is false.

- Agents often have difficulties finding meaningful evidence if they are arguing for the false position.
- For many arguments most of the relevant information is already contained in the first step of the agents.
- Relevant information about the neighborhood of entities can be encoded in the embeddings of entities. While the judge has access to this information through the training process, it remains hidden to users.

User Study

- Online quiz consisting of ten rounds: Each round is centered around a around a person from FB15k-237.
- Along with a query (which can be true or false) we present the users six arguments extracted by the agents in randomized order.
- Based on these arguments the respondents are supposed to judge whether the statement is true or false.


Abschnitt 4 von 14

Statement: Person 1 has female gender.

Fact 1: Person 1 was nominated for an award for the Best Female Lead. Joan Allen was also nominated for an award for the Best Female Lead.

Fact 2: Person 1 was nominated for the Primetime Emmy Award for Outstanding Lead Actress in a Comedy Series. Malcom is the Middle was also nominated for the Primetime Emmy Award for Outstanding Lead Actress in a Comedy Series.

User Study

- Based on 44 participants (109 invitations were sent) we find that Based on a majority vote nine out of ten questions were classified correctly.
- In addition, we asked the respondents to rate their confidence: We found that when users assigned a high confidence score to their decision ('rather certain' or 'absolutely certain') the overall accuracy of their classification was 89%. The accuracy dropped to 68.4% when users assigned a low confidence score ('rather uncertain' or 'absolutely uncertain').

Conclusion

- We proposed R2D2, a new approach for KG reasoning based on a debate game between two opposing reinforcement learning agents.
- R2D2 outperforms all baselines in the triple classification setting on the benchmark datasets WN18RR and FB15k-237.
- The results of our survey indicate that the arguments are informative and that the judge is aligned with human intuition.

Thank you!

marcel.hildebrandt@siemens.com

References I

[1] A. Bordes, X. Glorot, J. Weston, and Y. Bengio.
A semantic matching energy function for learning with multi-relational data.
Machine Learning, 94(2):233–259, 2014.

[2] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko.
Translating embeddings for modeling multi-relational data.
In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[3] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel.
Convolutional 2d knowledge graph embeddings.
In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

References II

[4] S. M. Kazemi and D. Poole.
Simple embedding for link prediction in knowledge graphs.
In *Advances in Neural Information Processing Systems*, pages 4284–4295, 2018.

[5] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu.
Learning entity and relation embeddings for knowledge graph completion.
In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[6] M. Nickel, V. Tresp, and H.-P. Kriegel.
A three-way model for collective learning on multi-relational data.
In *ICML*, volume 11, pages 809–816, 2011.

References III

[7] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling.
Modeling relational data with graph convolutional networks.
In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[8] R. Socher, D. Chen, C. D. Manning, and A. Ng.
Reasoning with neural tensor networks for knowledge base completion.
In *Advances in neural information processing systems*, pages 926–934, 2013.

[9] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang.
Rotate: Knowledge graph embedding by relational rotation in complex space.
arXiv preprint arXiv:1902.10197, 2019.

References IV

[10] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon.
Representing text for joint embedding of text and knowledge bases.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, 2015.

[11] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard.
Complex embeddings for simple link prediction.
In *International Conference on Machine Learning (ICML)*, volume 48, pages 2071–2080, 2016.

[12] R. J. Williams.
Simple statistical gradient-following algorithms for connectionist reinforcement learning.
Machine learning, 8(3-4):229–256, 1992.

References V

[13] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng.
Embedding entities and relations for learning and inference in knowledge bases.
In *International Conference on Learning Representations*, 2015.