Computer Science and Application 计算机科学与应用, 2022, 12(3), 602-609
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123061

Hans汉斯

融合结构与语义信息的知识推理算法

刘明坤

北京邮电大学，北京

收稿日期：2022年2月14日；录用日期：2022年3月10日；发布日期：2022年3月17日

摘　要

知识图谱以三元组形式存储了大量的事实、知识，但同时也存在着事实缺失的问题，因此需要在图谱的已知事实基础上推理预测新的事实即知识推理。传统的知识推理算法只简单利用了知识图谱的结构信息，对知识图谱的信息挖掘不够充分。本文提出了一个融合语义和结构信息的知识推理算法，该算法在利用知识图谱的结构信息的同时，也利用了大规模文本数据中的上下文信息，能够更加准确地表示实体、关系等知识图谱的基本元素。同时针对知识推理模型训练过程中三元组负采样存在的低质量和假阴性问题，我们引入了生成对抗网络来解决这个问题。实验表明，本算法可以实现良好的知识推理效果。

关键词

知识图谱，知识推理，负采样

Knowledge Inference Algorithm Based on Combination of Structure and Context

Mingkun Liu

Beijing University of Posts and Telecommunications, Beijing

Received: Feb. 14th, 2022; accepted: Mar. 10th, 2022; published: Mar. 17th, 2022

Abstract

Knowledge graph stores a lot of facts and knowledge by triples, but there is also the problem of lack of fact. Therefore, it is necessary to infer and predict new facts, that is, knowledge inference, based on the known facts of the graph. The traditional knowledge inference algorithms only make use of the structure information of the knowledge graph, which is not sufficient to mine the information of the knowledge graph. This paper proposes a knowledge inference algorithm that combines semantic and structural information. It not only uses the structure information of knowledge graph, but also uses the context information in large-scale text data, which can more accu-

文章引用: 刘明坤. 融合结构与语义信息的知识推理算法[J]. 计算机科学与应用, 2022, 12(3): 602-609.
DOI: 10.12677/csa.2022.123061

rately represent the basic elements of knowledge graph such as entities and relationships. At the same time, because of the problem of low quality and false negativity of negative sampling of triples in the training of knowledge inference model, we introduce the GANs to solve this problem. Experiments show that this algorithm can achieve good knowledge inference effect.

Keywords

Knowledge Graph, Knowledge Inference, Negative Sampling

1. 引言

自 1989 年万维网出现以来，互联网技术经历了多次迭代，今天正朝着以语义网(Semantic Web)概念为基础的知识互联的“Web 3.0”时代发展。在这一基础上，Google 于 2012 年提出了知识图谱这一概念，并宣布以知识图谱项目为基础构建下一代智能搜索引擎。随后，这一技术引起了业界和学术界的广泛关注，并应用于医疗、教育、电商等行业，推动人工智能从感知智能向认知智能跨越。

知识图谱以三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$形式存储着大量事实、知识。随着知识图谱的不断发展，对知识图谱的推理也越来越受到人们的关注。因为无论知识图谱的规模多么庞大(可能存在数百万的实体和数亿的关系事实)，总会有一些需要的事实在其中是缺失的。如何利用知识图谱中已经存在的实体、关系或事实信息，来推理预测实体之间新的关系，已经得到了相关领域研究学者的重视。但是当前在知识表示学习领域仍面临一些挑战：1) 如何在通过知识图谱结构学习到独特的实体关系表示的同时，充分利用好在大规模文本数据中的上下文、句法和语义信息。2) 如何在知识推理模型训练过程中，解决三元组负采样过程中出现的低质量和假阴性问题。

目前的知识推理算法中，以 TransE [1]为代表的基于翻译的知识推理模型只简单利用了知识图谱的结构信息，其采用的均匀采样[1]导致负采样得到的三元组存在低质量和假阴性问题，伯努利采样[2]虽然缓解了假阴性问题，但没有很好地解决低质量问题；KG-BERT [3]则仅利用文本数据的语义信息进行建模。针对以上问题，本文提出一种融合结构与语义信息的知识推理算法 ScKGAN (Adversarial Learning based Combination of Structure and Context for Knowledge Graph Completion)。在该算法中，我们引入生成对抗网络 GANs 通过生成器(Generator)与判别器(Discriminator)的对抗性训练以提升负采样获得的三元组质量，生成器(Generator)采用具有 softmax 概率分布的知识图谱补全模型捕捉知识图谱结构信息，判别器(Discriminator)则采用双向 Transformer 预训练语言模型捕获大规模文本数据中的上下文语义信息以实现知识图谱结构与语义信息的融合。在公开数据集上进行测试表明，该算法效果表现良好。

2. 相关工作

2.1. 知识推理算法

随着深度学习技术在各领域的广泛深入应用，在面向知识图谱的知识推理方向上，深度学习也在发挥着越来越多的作用。例如，一类基于神经网络的推理方法将知识图谱中事实三元组表示为向量形式送

入神经网络中，通过训练神经网络不断提高事实三元组的得分，最终通过输出得分选择候选实体完成推理，这类推理方法没有很好地利用事实三元组的上下文语义信息。Socher [4]等提出了神经张量网络(Neural tensor networks, NTN)模型，这一模型使用双线性张量层替换神经网络层，实体通过连续的词向量平均表示。Neelakantan [5]等针对知识图谱中存在的多跳路径，利用循环神经网络进行建模。Graves [6]等结合记忆系统和神经网络，完成了构建可微神经计算机模型工作，经样本学习到知识后将其存储起来，进行快速知识推理。

近年来，以深度学习为代表的表示学习(知识图谱表示，又称知识图谱嵌入)技术也取得了重要的进展，其关键思想是将知识图谱中的实体和关系映射到连续的稠密低维实值向量空间中，以便简化操作，同时保留知识图谱的固有结构，进而在低维空间中高效计算实体、关系及其之间的复杂语义关联，这类算法只是利用了知识图谱的结构信息来进行实体和关系的学习。基于分布式表示的推理是知识表示学习中的一类代表性算法，因其简单高效且适应于大规模知识图谱推理的特点而不断发展。例如，Bordes [1]等提出了 TransE 模型，将所有的实体和关系表示为同一个空间下的向量，假设事实元组中头实体向量和关系向量之和应该约等于尾实体的向量。通过随机替换事实元组中的某一项来构建负例。计算元组中头向量和关系向量的和向量与尾向量的距离作为候选实体的得分。TransE 模型简单有效，基于该模型衍生出很多方法。Nickel [7]等提出的 RESCAL 模型为了捕捉实体和向量中隐含的语义，对潜在因子间的相互作用进行建模，获得了关系表示矩阵，通过计算实体向量与关系矩阵的乘积来得到元组得分。

2.2. 三元组负采样机制

负采样的思想首先在语言的概率神经模型中提出，并被称为重要性抽样。Mikolov [8]等人强调，它是 NCE 的简化版本，有利于 word2vec 的培训。负采样将密度估计问题转化为一个二元分类问题，能够区分真实样本和噪声样本，从而简化了计算并加速了训练。按照 NCE 的思想，为了提高知识推理模型的训练效率，需要大量的负样本。因此，负采样成为知识表示学习中的一个关键点。

均匀采样[1]是最常用的负采样方法之一，它通过将头部或尾部实体替换为从知识图谱实体集中均匀采样的实体来腐蚀正三元组。然而，这种产生的负三元组太容易被辨别，并且在大多数情况下对训练几乎没有贡献。与随机均匀模式下的等概率抽样不同，伯努利采样[2]在头部和尾部替换中采用不同的概率来解决假阴性问题，但没有很好地缓解三元组负采样的低质量问题。TransE SNS [9]和 NSCaching [10]尝试将负采样的候选实体收集到定制集群中。此外，受 CKRL [11]的启发，NKRL [12]提出了一种置信度感知的负采样方法。Trouillon 等人[13]进一步研究了每个正三元组产生的负三元组样本数量，得出每个正样本 50 个负样本是平衡准确性和持续时间的好选择。

3. 融合结构与语义信息的知识推理算法

以 KG-BERT [3]为代表的知识推理文本编码方法，利用三元组自然语言文本的上下文、句法、语义信息来预测知识图谱的缺失部分。其优势是能尽可能克服知识图谱不完备性问题的影响，易于扩展到知识图谱中看不见的元素。但是，文本编码器学习知识图谱结构性信息的能力很差，导致这类算法缺乏对知识图谱结构性知识的掌握，无法克服由于自然语言的复杂性出现的实体歧义问题；此外，由于在进行知识图谱的链接预测任务时需要对所有可能的三元组进行推理，该方法由于采用原生 BERT，直接连接事实三元组的语义信息作为输入，致使推理代价高昂，出现了组合爆炸的问题。相比之下，TransE [1]等基于翻译的知识推理算法，纯粹是从结构化知识中进行学习，从而没有暴露出实体歧义问题。但是由于缺乏文本的上下文语义信息，它的表现仍然不佳。因此，我们需要在知识推理算法中融合图谱结构知识

和三元组上下文语义信息，以提升知识图谱推理算法的表现。

同时，上述方法在模型训练时，进行三元组负采样的方法是对于每一个正确的三元组，通过随机替换头实体或尾实体生成相应的负三元组集合。这种均匀采样方法在大多数情况下，替换的实体与损坏的正三元组无关，上述传统知识推理算法在训练过程中无法监督避免，导致形成的负三元组质量太低，不利于训练。以三元组(北京、位于、中国)为例，需替换其尾部实体“中国”，以生成对应的负三元组。利用这种方法，产生的负三元组可能是(北京、位于、草莓)或(北京、位于、牙刷)，知识推理模型仅根据不同的实体类型就可以很容易地区分这些低质量的三元组，这会减慢模型训练的收敛速度。均匀采样的另一个严重缺点在于假阴性负样本。例如用姚明取代三元组(易建联，性别，男)的头实体，得到的负三元组(姚明，性别，男)仍然是一个真实(假阴性)的事实，这会大大降低知识推理算法掌握事实的准确程度。为了缓解假阴性问题，伯努利负采样建议根据关系(即在一对多关系中给更多的机会替换头部，在多对一关系中给更多的机会替换尾部)的映射特性，用不同概率替换头部或尾部实体。但这依然无法有效缓解三元组负采样过程中出现的低质量问题。

针对上述问题，受到 GANs 的启发，本文提出了一个名为 ScKGAN 的对抗性训练框架，结构如图 1 所示，该模型包含两个模块生成器(Generator)和判别器(Discriminator)。我们使用具有 softmax 概率的 ComplEx 模型作为生成器(Generator)以学习知识图谱的结构性知识；判别器(Discriminator)则采用双向 Transformer 预训练语言模型能够同时捕获知识图谱三元组文本的上下文、句法和语义信息；生成器根据判别器的奖励学习提供高质量的阴性负三元组，判别器利用生成器中输出的负三元组学习知识图谱嵌入。

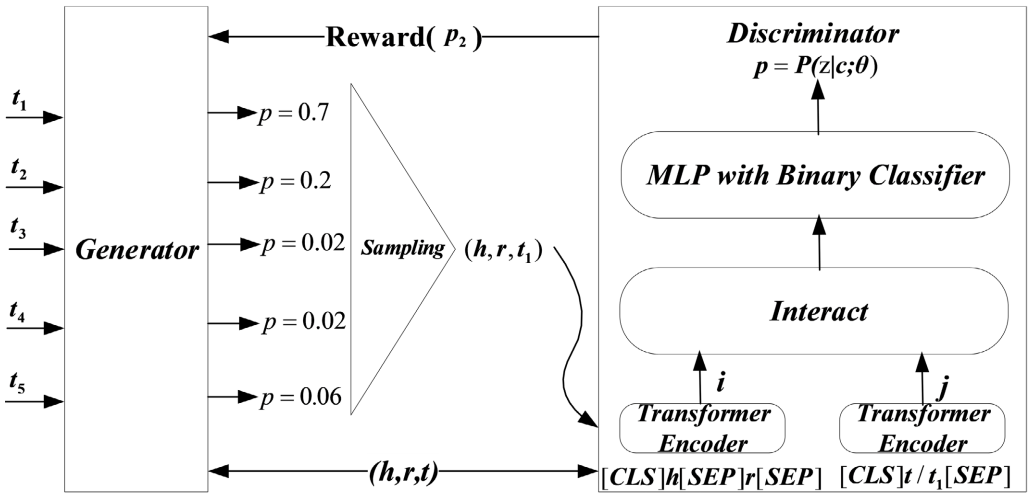

Figure 1. Flowchart of ScKGAN algorithm
图 1. ScKGAN 算法示意图

3.1. 生成器

生成器(Generator)采用具有 softmax 概率的经过预训练的 ComplEx [13]模型，以充分模拟离散 GAN 的“概率分布采样”过程。其计算一组候选负三元组的概率分布，然后从分布中抽取一个负三元组作为输出，为判别器(Discriminator)提供高质量的阴性负样本。ComplEx [13]模型打分函数为(1)，使用负对数似然损失函数进行预训练：

$$f(h,r,t) = \mathbf{Re}\left(\langle h,r,t\rangle\right) \tag{1}$$

由于生成器的输出是实体的离散索引，我们不能使用 GAN 的原始形式来生成离散样本(例如自然语

言语句或知识图谱三元组)，因为离散采样步骤防止梯度反向传播回生成器，我们调整了 GAN 的原始形式，使用基于策略梯度的强化学习训练生成器。

$\mathbb{D}^+$、$\mathbb{D}^-$集分别为正三元组集合和相对应的负三元组集合，假设生成器在给定正三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$的情况下，生成负三元组的概率分布$\mathcal{P}_G(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}'|\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$的概率分布，并通过从该分布中采样生成负三元组$(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}')$。生成器的目标可以表述为最大化以下三元组负概率的期望值：

$$\mathcal{R}_G=\sum_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})\in\mathbb{D}^+}\mathbb{E}\left[-\log(p_2)\right]\quad(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}')\sim\mathcal{P}_G(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}'|\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) \tag{2}$$

3.2. 判别器

在判别器(Discriminator)中，为了降低模型应用于链接预测需要对所有可能的三元组进行推理从而导致组合爆炸造成太过高昂的计算成本问题，具体地说，我们从基于翻译的知识图谱推理方法(TransE)中得到启发，首先将每个三元组划分为两个部分：一个是头部和关系的组合，另一个是尾部。然后，通过对其文本应用孪生文本编码器，将每个部分编码为单独的上下文表示。最后，为了在三元组中捕获实体和关系之间的上下文信息以表示三元组，以交互方式连接这两种表示，在此基础上训练二元神经分类器。

判别器(Discriminator)采用双向 Transformer 预训练语言模型，其中的交互式连接写为：

$$\boldsymbol{c}=[\boldsymbol{i},\boldsymbol{i}\times\boldsymbol{j},\boldsymbol{i}-\boldsymbol{j},\boldsymbol{j}] \tag{3}$$

$\boldsymbol{c}$ 用于表示三元组的两个部分之间的语义关系，将双向分类器应用于 $\boldsymbol{c}$ 并生成对应正负概率的二分类分布如(4)：

$$\boldsymbol{p}=P(z|\boldsymbol{c};\theta)\triangleq softmax(MLP(\boldsymbol{c};\theta))\in\mathbb{R}^2 \tag{4}$$

其中 $MLP(\cdot)$ 表示多层感知器，θ 是它的可学习参数。p_2 即是三元组为正的概率，可以作为三元组的分数来执行候选三元组排名完成链接预测任务。

使用 BCE 损失函数(5)进行优化，$\mathrm{y}_T\in\{0,1\}$ 为三元组正负标签：

$$\mathcal{L}=-\sum_{T\in\mathbb{D}^+\cup\mathbb{D}^-}\left(\mathrm{y}_T\log(1-p_2)+(1-\mathrm{y}_T)\log p_2\right) \tag{5}$$

判别器计算的奖励函数表述为：

$$R=-\log(p_2) \tag{6}$$

4. 实验

4.1. 实验设置

本文在两个数据集上进行了实验。如表 1 所示，其中 WN18RR 是公共知识图谱 WordNet 子集 WN18 的一个子集，其中涵盖了大量英语单词间的语义关系。虽然 WN18RR 较 WN18 仅仅少了 7 种关系，但是它大大增加了知识图谱表示及推理的难度。FB15k-237 则是公共知识图谱 FreeBase 子集 FB15K 的一个子集，它包含了很大量级的客观世界的事实，去掉 FB15K 中很多冗余的关系，FB15k-237 中的三元组数量约为 FB15k 的一半。

Table 1. Experiment dataset
表 1. 实验数据集

数据集	实体数量	关系数量	训练集数量	验证集数量	测试集数量
WN18RR	40,943	11	86,835	3034	3134
FB15k-237	14,541	237	272,115	17,535	20,466

实验环境：服务器使用的操作系统及版本为 Ubuntu14.04，采用 python 3.6 作为主要编程语言，并基于深度学习开源框架 Pytorch 实现了 ScKGAN。进行实验的硬件环境是型号为浪潮 P8000 的主机，其具备英特尔 Xeon(R) CPU 双物理核 32 线程处理器、256GB 内存和 NVIDIA GeForce RTX 1080Ti *2 GPU，GPU 加速库及版本为 Cuda 10.2.89。

4.2. 结果评估

本节对算法效果进行评估，参考传统知识表示及推理算法中的评价指标，本文通过链接预测任务来验证本文提出的算法的实际表现，利用训练得出的实体和关系的向量表示可以得出正负三元组的得分，再对得分进行排序来衡量推理算法的效果。本文使用 Mean Rank 和 Hits@10 两种评价标准。Mean Rank 即所有正三元组在其对应的正负三元组中的平均排名，因此 Mean Rank 的值越小代表推理算法效果越好。Hits@10 表示正三元组的排名在其对应的正负三元组中的前 10 位的比例，因此 Hits@10 的值越大代表推理算法效果越好。

依据上述评判标准，本文选择 TransE [1]模型、ComplEx [13]模型、DistMult [14]模型、TuckER [15]模型、ATTH [16]模型、KG-BERT [3]模型进行对比实验，详细结果如表 2 所示。实验结果表明，本算法在 Mean Rank 指标上表现优于其他算法，在 Hits@10 指标上表现弱于 SOTA 模型。基于上述结果，我们尝试解释原因为，1) 融合了结构与语义信息的推理算法，在生成对抗网络框架下进行三元组负采样，能够增强推理算法对正负三元组的区分能力；2) 受限于大规模文本数据上下文语义信息的高歧义性，在融合了结构信息的情况下，双向 Transformer 预训练语言模型对三元组的建模能力仍受到一定限制。

Table 2. Comparison experiment of different models
表 2. 不同模型对比实验

Method	WN18RR		FB15K-237	
	MR	Hit@10	MR	Hit@10
TransE	2365	50.5	223	47.4
ComplEx	3921	48.3	508	43.4
DistMult	3704	47.7	411	41.9
TuckER	-	52.6	-	**54.4**
ATTH	-	**55.1**	-	50.1
KG-BERT	97	52.4	153	42.0
ScKGAN(ours)	**80**	**54.1**	**120**	**52.2**

4.3. 消融研究

为了进一步验证本文所提出的模型引入结构信息以融合语义信息以及在生成对抗网络框架下进行三元组负采样的必要性，我们通过忽略生成器模型，提出了以微调判别器模型为基础实现的 KG-SBERT 算法完成消融研究。结果如表 3 所示，KG-SBERT 算法表现明显劣于 ScKGAN 算法。实验结果表明，为了解决当前知识推理模型存在的问题所采用的方法可以提高知识推理算法的表现，结合表 4 的实验结果，应用孪生文本编码器的 KG-SBERT 算法以牺牲部分性能为代价，避免了直接使用 BERT 进行编码带来的三元组组合爆炸问题，明显降低了模型训练时间，提高了模型训练速度。

Table 3. Result of ablation experiment
表 3. 消融实验结果

Method	WN18RR		FB15K-237	
	MR	Hit@10	MR	Hit@10
KG-BERT	97	52.4	153	42.0
ScKGAN(ours)	**80**	**54.1**	**120**	**52.2**
KG-SBERT(ours)	109	50.4	171	40.9

Table 4. Time of training model
表 4. 模型训练时间

Method	TRAIN TIME
KG-BERT	40 h
KG-SBERT(ours)	6 h

5. 结论

本文针对知识推理算法存在的问题，在生成对抗网络框架下，在利用大规模文本数据的上下文、文法及语义信息与捕获知识图谱结构信息的同时，缓解三元组负采样存在的低质量与假阴性问题，从而提升知识推理算法的推理效果。本文在两个数据集上进行了实验测试，实验结果表明，与之前的知识推理算法比较，本文提出的算法在 Mean Rank 上有明显的提升，在 Hits@10 上略弱于 SOTA 模型，算法的效果具备竞争力。在下一步工作中，本文将使用更多的知识图谱数据集进行实验，同时对生成器模型与判别器模型进行进一步研究，以提升算法性能。

参考文献

[1] Bordes, A., Usunier, N., Garcia-Duran, A., *et al*. (2013) Translating Embeddings for Modeling Multi-Relational Data. 27*th Conference on Neural Information Processing Systems*, Lake Tahoe, 5-10 December 2013, 2787-2795.

[2] Wang, Z., Zhang, J., Feng, J. and Chen, Z. (2014) Knowledge Graph Embedding by Translating on Hyperplanes. *Proceedings of the* 28*th AAAI Conference on Artificial Intelligence*, Québec City, 27-21 July 2014, 1112-1119.

[3] Yao, L., Mao, C. and Luo, Y. (2019) KG-BERT: BERT for Knowledge Graph Completion. arXiv:1909.03193 http://arxiv.org/abs/1909.03193

[4] Socher, R., Chen, D., Manning, C.D., *et al*. (2013) Reasoning with Neural Tensor Networks for Knowledge Base Completion. Curran Associates Inc., Red Hook.

[5] Neelakantan, A., Roth, B. and McCallum, A. (2015) Compositional Vector Space Models for Knowledge Base Inference. arXiv:1504.06662.

[6] Graves, A., Wayne, G., Reynolds, M., *et al*. (2016) Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature*, **538**, 471-476. https://doi.org/10.1038/nature20101

[7] Nickel, M., Tresp, V. and Kriegel, H.P. (2011) A Three-Way Model for Collective Learning on Multi-Relational Data. *International Conference on International Conference on Machine Learning*, Bellevue, 28 June-2 July 2011, 809-816.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, Vol. 2, Lake Tahoe, 5-8 December 2013, 3111-3119.

[9] Qin, S., Rao, G., Bin, C., Chang, L., Gu, T. and Xuan, W. (2019) Knowledge Graph Embedding Based on Adaptive Negative Sampling. 2019 *International Conference of Pioneering Computer Scientists, Engineers and Educators*, Guilin, 20-23 September 2019, 551-563.

[10] Zhang, Y., Yao, Q., Shao, Y. and Chen, L. (2019) NSCaching: Simple and Efficient Negative Sampling for Knowledge

Graph Embedding. 2019 *IEEE* 35*th International Conference on Data Engineering* (*ICDE*), Macao (China), 8-11 April 2019, 614-625. https://doi.org/10.1109/ICDE.2019.00061

[11] Xie, R., Liu, Z. and Sun, M. (2018) Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning with Confidence. arXiv:1705.03202v2.

[12] Shan, Y., Bu, C., Liu, X., Ji, S. and Li, L. (2018) Confidence-Aware Negative Sampling Method for Noisy Knowledge Graph Embedding. 2018 *IEEE International Conference on Big Knowledge* (*ICBK*), Singapore, 17-18 November 2018, 33-40. https://doi.org/10.1109/ICBK.2018.00013

[13] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É. and Bouchard, G. (2016) Complex Embeddings for Simple Link Prediction. *Proceedings of the* 33*rd International Conference on International Conference on Machine Learning*, Vol. 48, New York, 19 June-24June, 2016, 2071-2080.

[14] Yang, B., Yih, W.-T., He, X., Gao, J. and Deng, L. (2015) Embedding Entities and Relations for Learning and Inference in Knowledge Bases. 3*rd International Conference on Learning Representations*, San Diego, 7-9 May 2015, 1-13.

[15] Balažević, I., Allen, C. and Hospedales, T.M. (2019) TuckER: Tensor Factorization for Knowledge Graph Completion. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing*, Hong Kong (China), 3-7 November 2019, 5185-5194. https://doi.org/10.18653/v1/D19-1522

[16] Chami, I., Wolf, A., Juan, D.-C., Sala, F., Ravi, S. and Ré, C. (2020) Low-Dimensional Hyperbolic Knowledge Graph Embeddings. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 6901-6914. https://doi.org/10.18653/v1/2020.acl-main.617