基于 Transformer 的表型-药物-分子多层次知识图谱链接预测

李光耀[1]，孙泽群[1]，胡伟[1,2⋆]

[1] 南京大学计算机软件新技术国家重点实验室
[2] 南京大学健康医疗大数据国家研究院
{gyli.nju,zqsun.nju}@gmail.com, whu@nju.edu.cn

摘要 知识图谱以三元组形式组织了大量结构化知识，然而其天然具有不完备性。为缓解该问题，链接预测任务通过对已有的三元组数据进行挖掘，从而预测出实体间的潜在未知关系。本文主要针对表型-药物-分子多层次知识图谱链接预测任务，基于 Transformer 框架设计了一种新型知识图谱嵌入模型。我们主要针对三元组的特殊结构，设计了一种新的位置编码方式。且有别于一般的掩码训练方式，我们充分利用了输出端所有位置的输出，为训练目标额外增加了一个损失项。实验结果表明，相比于目前主流的知识图谱嵌入模型，我们提出的模型效果显著，并在评测任务测试集上，取得了 MRR 指标为 0.201 的好成绩，排名第一。我们也进一步验证了本文提出的改进模块的有效性。

关键词： 知识图谱 · 链接预测 · Transformer

1 引言

知识图谱可视作由特定模式对人类知识进行组织的知识库，其中每一条知识表示为一个三元组，其有效刻画了真实世界中的各类知识，并广泛应用于各个领域中 [4]。在生物医药领域，表型药物发现是一个重要课题。通常，研究者会通过大量生物实验来确定药物治疗效果，但是这个过程会耗费大量时间。针对该问题，CCKS 2021 的第七个评测任务通过构建表型-药物-分子多层次知识图谱，利用链接预测任务来实现对疾病、症状、药物、基因、副作用等之间潜在关系的预测，从而为后续致病机理和药理作用机制研究做支撑。

链接预测任务通常采用知识图谱嵌入模型完成。知识图谱的嵌入指将实体与关系从符号空间嵌入到向量空间，同时要求实体与关系在向量空间中的

⋆ 通信作者

表示尽可能保持其在符号空间中的结构特性。知识图谱嵌入模型的设计主要包含三个部分 [12]。(i) 首先对实体、关系的向量表示进行建模，例如是在实数域空间还是在复数域空间建模；(ii) 设计评分函数，用于衡量一个三元组为正例的可能性；(iii) 训练部分，例如选择具体的损失函数。目前，知识图谱嵌入模型大致可以分为三类：平移距离模型、语义匹配模型、基于深度神经网络的模型，且这三类模型的评分函数分别基于距离、语义相似度，以及直接通过深层神经网络得到打分。

鉴于近年来 Transformer [9] 架构广泛应用于各个领域，并为诸多任务带来了性能突破。本文主要基于该框架构建了具有深层网络的嵌入模型，并应用于链接预测任务中。针对知识图谱三元组结构的特殊性，我们设计了一种新的位置编码方式，通过分别为头、尾实体设计独特的语义组合算子，将实体与关系的嵌入表示进行组合作为位置编码；同时区别于以往的掩码预测方式中只考虑掩码位置处的输出对最终预测的影响，我们考虑利用其他位置的输出表示构建训练目标，从而在预测中充分利用模型输出端所有位置的表示。通过在表型-药物-分子多层次知识图谱链接预测任务上进行实验，充分验证了我们模型的有效性。接下来，本文分别对相关工作、方法、实验以及总结进行介绍。

2 相关工作

本节首先对知识图谱、链接预测以及嵌入模型进行形式化描述，然后就相关代表工作展开介绍。

知识图谱中的三元组由实体集合 $\mathcal{E}$ 和关系集合 $\mathcal{R}$ 组成，一条三元组表示为 (h, r, t)，其中 $h, t \in \mathcal{E}$，表示头、尾实体，$r \in \mathcal{R}$ 表示关系。链接预测任务旨在通过已知的三元组集合 $\{(h, r, t)\}$，预测出潜在未知三元组中的缺失实体，例如给定头实体和关系来预测尾实体：$(h, r, ?)$，或与之相反：$(?, r, t)$。而知识图谱嵌入模型通过学出合理的映射函数 F 将实体 $\mathcal{E}$ 和关系 $\mathcal{R}$ 映射到向量空间 $\mathbb{R}^n$ 中，通过将实体的嵌入表示代入特定的评分函数中得到打分并进行排序，从而完成链接预测任务。下面，我们介绍三类知识图谱嵌入模型的代表工作。

2.1 平移距离模型

TransE [1] 首次提出平移距离的思想。其将实体建模为实数域空间上的点，关系建模为平移操作，计算头实体经过关系的平移操作后得到的表示与

尾实体之间的距离作为评分。在 TransE 基础上，后面有一大批对其进行改进的模型，例如 TransH [13]、TransR [5] 等。也有工作尝试在复数域空间进行建模，RotatE [7] 将实体建模为点，关系建模为复数域空间中的旋转操作，尽管其建模方式非常简单，但在基准数据集上，其取得了非常好的效果。

2.2 语义匹配模型

相比于第一类，该类模型的损失函数主要度量了一个三元组的语义匹配程度。RESCAL [6] 是最早提出的一个基于语义匹配的模型，其将实体建模为向量、关系建模为二维矩阵，其试图捕捉头尾实体向量中任意两个位置的交互信息。DisMult [14] 在 RESCAL 基础上进行了简化，它限制了每个关系的矩阵表示为对角矩阵，从而只捕捉头尾实体向量中对应的两两位置的交互信息。ComplEx [8] 受 DisMult 启发，通过在复数域空间建模来处理非对称关系。

2.3 基于深度神经网络的模型

近年来，随着深度学习的发展，一些知识图谱嵌入模型也试图引入这些深度神经网络模型。ConvE [2] 将头实体、关系向量拼接成二维矩阵，然后使用多层卷积网络提取特征得到一个表示，用这个表示和尾实体做内积作为该三元组的评分。RSN [3] 首先在知识图谱中进行随机游走，采样出路径，然后使用循环神经网络建模，并采用残差连接的方式，综合使用历史信息和当前三元组的输入信息进行预测。

最近，Transformer [9] 架构广泛应用于自然语言处理的各个领域，并为诸多任务带来了性能突破。作为一个新颖的深度神经网络框架，其强大的表征能力也启发了一些知识图谱嵌入模型的工作。CoKE [11] 是首个采用 Transformer 对知识图谱的结构信息进行建模的方法。它采用了类似 Bert 中的掩码预测任务，即在三元组中，对头、尾实体分别进行掩码，利用编码器得到掩码位置处的输出表示，并使用该表示进行分类，通过这个分类预测的损失来训练、优化模型。由于编码器中每一层都使用自注意力机制，掩码位置处的表示可以充分得到该三元组中其他两个字段的信息。鉴于 CoKE 在使用 Transformer 框架时没有过多地考虑知识图谱三元组结构的特殊性，我们在其基础上做进一步改进。

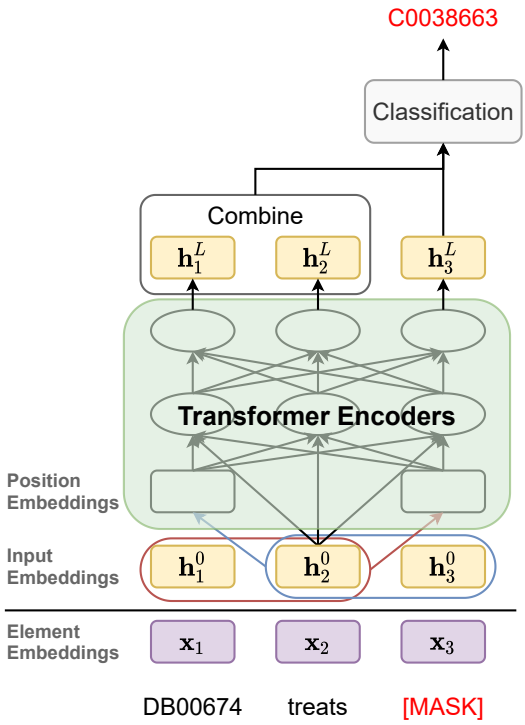

图 1: 模型框架

3 方法

本节首先介绍我们模型的整体框架，然后着重介绍提出的两点改进以及模型的训练细节。

3.1 框架

模型的整体框架如图 1所示。首先对于一个输入三元组，我们会掩盖其中的一个实体，例如图 1中掩盖了尾实体。随后三元组中每个元素根据映射表转换为对应的嵌入表示，记为 $\mathbf{h}_1^0$、$\mathbf{h}_2^0$ 和 $\mathbf{h}_3^0$。随后经过 L 层堆叠的 Transformer 编码层，每一层编码层包含自注意力模块和前馈传播网络模块，其中自注意力模块保证了每个位置上的嵌入表示充分融合了其上下文其他位置上的信息。最后输出记为 $\mathbf{h}_1^L$、$\mathbf{h}_2^L$ 和 $\mathbf{h}_3^L$。我们综合利用所有位置上的输出表示做分类任务，从而得到掩码位置上的预测实体。

3.2 位置编码

Transformer 中的核心自注意力模块如下所示：

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \tag{1}$$

其中 $\mathbf{Q} = \mathbf{H}^i\mathbf{W}^Q, \mathbf{K} = \mathbf{H}^i\mathbf{W}^K, \mathbf{V} = \mathbf{H}^i\mathbf{W}^V$。$\mathbf{H}^i$ 表示第 i 层的输入，由 $\mathbf{h}_1^i$、$\mathbf{h}_2^i$ 和 $\mathbf{h}_3^i$ 顺序拼接形成。$\mathbf{W}^Q$、$\mathbf{W}^K$ 和 $\mathbf{W}^V$ 为三个特征变换矩阵。注意到在公式 (1) 的计算中，任意两个元素之间的注意力值和它们所处的位置无关，即自注意力模块是顺序无关的。为了引入位置信息，有两类处理方式。一：绝对位置编码，在输入端每个位置处加上一个位置向量；二：相对位置编码，将相对位置的编码信息加入到注意力矩阵中。这两类编码方式都为位置信息进行单独编码，考虑到知识图谱三元组结构的特殊性，我们设计了一种新的位置编码方式，其基于对实体、关系的嵌入表示进行语义组合，从而区分每个位置。我们主要受 TransE [1] 中 $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$ 假设的启发，分别为头、尾实体设计了语义组合算子，并将组合后的表示作为对应位置的编码，如下所示：

$$\begin{aligned} \mathbf{p}_1^i &= \mathbf{h}_3^i - \mathbf{h}_2^i, \\ \mathbf{p}_3^i &= \mathbf{h}_1^i + \mathbf{h}_2^i, \end{aligned} \tag{2}$$

其中 $\mathbf{p}_1^i$、$\mathbf{p}_3^i$ 分别表示在第 i 层上的头、尾实体的位置编码。由于这两个语义组合算子的不对称性，即使不对关系所在的位置进行显示编码，我们依然能够区分三元组中的每个位置。最后我们将每一层的位置编码和输入向量相加作为编码层最终的输入表示。

3.3 输出端其他位置上的信息

在掩码式的训练方式中，通常只考虑了掩码处的输出，而忽略了其他位置处的输出对最终预测的影响。例如在图 1中，CoKE 只会使用 $\mathbf{h}_3^L$ 用作最终的预测，然而其他位置上的输出也充分融合了上下文信息，合理利用这部分信息能够有效辅助最终的预测。受 StAR [10] 启发，我们采用 3.2节中介绍的语义组合算子对输出端其他位置的嵌入表示进行语义组合，并将组合后的表示也用于预测。一方面，由于三元组结构的特殊性，语义组合后的表示一定程度也能刻画掩码处的语义信息；另一方面，通过对输出端其他位置信息的利用，可以使得训练目标的优化影响到更多的嵌入表示。

下面具体介绍我们模型的训练优化目标。当掩码位置处为尾实体时，我们会分别计算掩码位置的预测输出 $\mathbf{u}$ 以及对输出端其他位置表示进行组合得到的预测输出 $\mathbf{v}$，其计算过程如下所示：

$$\begin{aligned} \mathbf{u} &= \text{softmax}(\text{FC}(\mathbf{h}_3^L) \cdot \mathbf{E}), \\ \mathbf{v} &= \text{softmax}(\text{FC}(\mathbf{h}_1^L + \mathbf{h}_2^L) \cdot \mathbf{E}), \end{aligned} \tag{3}$$

其中 FC() 表示一层全连接网络，$\mathbf{E}$ 表示所有嵌入表示组成的矩阵。这里我们分别对掩码位置处的嵌入表示和输出端其他位置处的嵌入表示的组合做进一步映射变换，以提高表达能力。两个预测出的嵌入表示都通过内积计算和所有实体的相似度，从而得到预测概率分布，即 $\mathbf{u}$ 和 $\mathbf{v}$。我们采用交叉熵函数计算预测分布和真实概率分布之间的损失，最终的训练目标如下所示：

$$\mathcal{L} = \mathcal{L}_1 + \beta \cdot \mathcal{L}_2, \tag{4}$$

其中 $\mathcal{L}_1 = \text{cross\_entropy}(\mathbf{y}, \mathbf{u}), \mathcal{L}_2 = \text{cross\_entropy}(\mathbf{y}, \mathbf{v})$。$\mathbf{y}$ 表示真实的概率分布，其应该是一个独热编码，为了避免过拟合，我们对其进行平滑。注意到我们使用了一个超参数 β 来控制使用其他位置信息进行预测造成损失的影响程度。

4 实验

4.1 数据集

我们使用 CCKS 2021 表型-药物-分子多层次知识图谱链接预测任务提供的数据集，其中训练集中包含 74085 个实体和 7 个关系，以及 1069113 条三元组。官方还提供了未公开的测试集。我们从训练集中随机选出 5000 条三元组作为验证集，剩余的三元组用于训练。下面如果不额外说明，所有的实验结果都为在这验证集上的测试结果。表 1记录了具体的实验数据统计情况。

4.2 实验设置

我们使用网格搜索，根据在验证集上的 MRR 选择超参数。其中 Transformer 编码层的层数，我们在 $\{4, 5, 6, 7, 8\}$ 中进行选择；嵌入表示的维度在 $\{256, 512, 1024\}$ 中选择；自注意力头个数在 $\{2, 4, 8\}$ 中选择；Dropout 在

表 1: 实验数据统计情况

实体数量	关系数量	训练集大小	验证集大小	测试集大小
74085	7	1064113	5000	64666

$\{0.1, 0.2, 0.3, 0.4\}$ 中选择。我们选用 Adam 优化器，学习率在 $\{0.0001, 0.0005\}$ 中选择；批大小在 $\{256, 512, 1024, 2048, 4096\}$ 中选择。表 2记录了我们最终确定的超参数设置情况。

表 2: 超参数设置

编码层层数	嵌入维度	自注意力头个数	Dropout	学习率	批大小
5	1024	2	0.1	0.0001	1024

4.3 实验结果

我们选取了一些具有代表性的知识图谱嵌入模型作为基线方法，包括：TransE、DisMult、ComplEx、ConvE、RotatE 以及 CoKE。表 3记录了所有模型在我们随机划分的验证集上的测试结果。注意到我们提出的模型在所有指标上都取得了最优的效果。同时相比于平移距离模型、语义匹配模型以及基于某些深度神经网络的模型，基于 Transformer 框架的嵌入模型效果显著，例如 CoKE 在 MRR 指标上比 RotatE 高了 0.015。我们认为这是由于 Transformer 框架有利于对实体、关系的上下文信息进行建模，从而能够更好地应对知识图谱中复杂的关系模式。而我们的方法相比于 CoKE 能获得进一步的提升，一定程度证明了我们提出的改进模块的有效性。

为了进一步验证两个改进模块的作用，我们展开了消融实验，结果如表 4所示。注意到，在去掉我们设计的位置编码后，模型的性能大大下降，这进一步表明了对于 Transformer 这种框架而言位置编码的重要性。而在训练目标中去除 $\mathcal{L}_2$ 这个损失项，即只使用掩码位置上的输出做预测，性能也出现了一定程度的下降，这表明了充分利用输出端所有位置的信息是有益的。同时将 “w/o $\mathcal{L}_2$” 的结果和 CoKE 对比，可以发现相比于 CoKE 中的绝对位

表 3: 验证集上的测试结果

模型	Hits@1	Hits@3	Hits@10	MRR
TransE [1]	0.055	0.108	0.201	0.105
DisMult [14]	0.091	0.175	0.299	0.161
ComplEx [8]	0.098	0.195	0.333	0.175
ConvE [2]	0.102	0.190	0.308	0.172
RotatE [7]	0.104	0.210	0.343	0.185
CoKE [11]	0.117	0.227	0.363	0.200
本文提出的模型	**0.128**	**0.243**	**0.386**	**0.213**

表 4: 消融实验

模型	Hits@1	Hits@3	Hits@10	MRR
本文提出的模型	**0.128**	**0.243**	**0.386**	**0.213**
w/o 位置编码	0.106	0.207	0.329	0.183
w/o $\mathcal{L}_2$	0.124	0.236	0.371	0.208

置编码方式，我们提出的位置编码更好地适用于对知识图谱三元组位置的建模，效果也更好。

我们也尝试了模型融合。我们选取了 13 组不同参数下的模型，计算这些模型前 100 个预测出的实体和对应的预测分数，然后对这些预测结果按相同实体分数相加的方式进行融合，最后将分数降序排列并选出前 10 个作为最终的预测结果。经过模型融合后，我们在验证集上，MRR 指标能够达到 0.215。最终，我们将这个版本提交到评测网站上，得到测试集上 MRR 指标为 0.201 的最好成绩。

5 总结

我们在 CCKS 2021 表型-药物-分子多层次知识图谱链接预测任务上，基于 Transformer 改进设计了一种新的知识图谱嵌入模型。我们主要提出了两点改进。一：针对 Transformer 中的自注意力模块是顺序无关的，提出了一

种新的位置编码方式。二：有别于以往掩码预测中只考虑掩码输出对预测的影响，我们综合利用了所有位置的输出表示用于最终的预测。我们提出的模型在评测数据集上取得了最好的成绩。同时我们也在本地验证集上验证了我们提出的两个改进模块的有效性。未来，我们考虑将实体属性信息融入到我们的模型框架中，从而进一步提升链接预测任务的效果。

参考文献

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems **26** (2013)
2. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: Thirty-second AAAI conference on artificial intelligence (2018)
3. Guo, L., Sun, Z., Hu, W.: Learning to exploit long-term relational dependencies in knowledge graphs. In: International Conference on Machine Learning. pp. 2505–2514. PMLR (2019)
4. Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems (2021)
5. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
6. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: Icml (2011)
7. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197 (2019)
8. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: International conference on machine learning. pp. 2071–2080. PMLR (2016)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
10. Wang, B., Shen, T., Long, G., Zhou, T., Wang, Y., Chang, Y.: Structure-augmented text representation learning for efficient knowledge graph completion. In: Proceedings of the Web Conference 2021. pp. 1737–1748 (2021)

11. Wang, Q., Huang, P., Wang, H., Dai, S., Jiang, W., Liu, J., Lyu, Y., Zhu, Y., Wu, H.: Coke: Contextualized knowledge graph embedding. arXiv preprint arXiv:1911.02168 (2019)
12. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017)
13. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 28 (2014)
14. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)