基于表示学习的翻译模型综述

田应彪

（大连海事大学 信息科学技术学院，辽宁省大连市 中国 116026）

摘 要 人类知识提供了对世界的正式的理解。表示实体之间结构关系的知识图谱已成为面向认知系统和人类智能的日益流行的研究方向。本文回顾了基于知识表示的翻译模型 TransE, TransH, TransR, CTransR 以及 TransD 模型，并对这些模型进行了比较。

关键词 知识表示；翻译模型；知识图谱

中图法分类号 TP311 **文献标识码** A

Overview of Translation Models Based on Representational Learning

Yingbiao Tian

(School of Information Science and Technology, Dalian maritime university, Liaoning Dalian 116026 China)

Abstract Human knowledge provides a formal understanding of the world. Knowledge graphs that represent structural relations between entities have become an increasingly popular research direction towards cognition and human-level intelligence. This paper reviews and compares the knowledge representation based translation models TransE, TransH, TransR, CTransR, and TransD.

Key words Knowledge representation; Translation model; Knowledge Graph

1、知识库简史

知识表示在逻辑和人工智能领域经历了长期的发展历史。图形知识表示的思想最早可以追溯到1956 年 Richens 提出的语义网概念，而符号逻辑知识可以追溯到 1959 年的一般问题求解器[1]。知识库首先与基于知识的系统一起用于推理和解决问题。MYCIN 是最著名的基于规则的医学诊断专家系统之一，知识库约有 600 条规则。后来，在人类知识表示社区见证了了基于框架的语言、基于规则的和混合表示的发展。大概在这个时期末，Cyc 项目开始了，项目旨在汇集人类知识。资源描述框架(RDF)和网络本体语言(OWL)相继发布，成为语义网的重要标准。随后，许多开放的知识库或本体被发表，如 WordNet、DBpedia、Y AGO 和 Freebase。Stokman 和 Vries 在 1988 年提出了一个关于图中结构知识的现代概念。然而，正是在 2012 年，知识图的概念自谷歌的搜索引擎首次推出以来获得了极大的流行，提出了一种名为 knowledge Vault 的知识融合框架来构建大规模知识图。

2、定义

通过描述一般语义表示或基本特征来给出定义的研究已经做了很多。然而，没有这种广泛接受

的正式定义。Ehrlinger 分析了现有的几种定义，提出了强调知识图推理引擎的定义 1。Wang 等在定义 2 中提出了多关系图的定义。一般将知识图定义为 $G=\{E,R,F\}$，其中 E，R，F 分别是实体，关系和事实的集合。一个事实表示为三元组$(h,r,t) \in F$。

Definition 1 (Ehrlinger and Woß [12]) 知识图获取信息并将信息集成到本体中，并应用推理机来导出新知识。

Definition 2 (Wang et al. [5]) 知识图是由实体和关系组成的多关系图，它们分别被视为节点和不同类型的边。

3、知识表示学习及翻译模型回顾

近年来，以深度学习为代表的表示学习迅速发展，在很多领域都取得了巨大进展。简而言之，表示学习将要描述的对象表示为低维稠密向量，这也被称为分布式表示，从而有效解决了数据稀疏问题，并且便于在低维语义空间中进行计算。将表示学习应用于知识图谱（Knowledge Graph, KG），即是知识表示学习。

在当前的主流知识库中，知识被存储为(h,r,t)的三元组形式，其中 h 表示头实体，r 表示联系,t 表示尾实体。知识表示学习的任务就是学习 h,r,t 的分布式表示（也被叫做知识图谱的嵌入表示(embedding)）。

目前，知识表示学习方法从实现形式上可以分为两类：基于结构的方法和基于语义的方法。基于结构的嵌入表示方法包括 TransE, TransH, TransR & CTransR, TransD 等，这类方法从三元组的结构出发学习 KG 的实体和联系的表示；基于语义的嵌入表示方法包括 NTN, SSP, DKRL 等，这类方法从文本语义的角度出发学习 KG 的实体和联系的表示。

知识表示学习从发展来看可以分成两个阶段，以 2013 年 Borders 等人受 Mikolov 发现的语义空间中词向量的平移不变现象启发，从而提出了 TransE 为分割。在 TransE 之前，有 Structed Embedding, Semantic Matching Energy 等模型，在 TransE 之后，人们在此基础上进行改进，依次提出了 TransH, TransR, TransD 等模型。本篇介绍 TransE 及其后续模型。

3.1 TransE

我们用 h 表示头实体向量，用 t 表示尾实体向量，用 r 表示关系向量，TransE 模型的目标就是让 $t-h$ 尽可能地等于 r ，即 $t-h \approx r$ 。其评分函数为：

$$f_r(h,t) = \|h+r-t\|_{L1/L2} \qquad (1)$$

显然，对于正确的三元组，应该有较低的得分。在训练过程中，使用等级损失函数，这是因为在当前情况下我们没有就标签而言的监督，只有一对正确项 $(h,r,t) \in \Delta$ 和不正确项 $(h',r',t') \in \Delta'$ ，我们的目的是让正确项的得分比不正确项高。这种情况出现在我们只有正例时，知识图谱就是这种情况，我们只知道正确的三元组（golden triplet），再通过破坏一个正例来生成负例。等级损失就适用于这种情况，因此我们定义损失函数：

$$L = \sum_{(h,r,t) \in \Delta} \sum_{(h',r',t') \in \Delta'} \max\left(f_r(h,t) + \gamma - f_{r'}(h',t'),0\right) \qquad (2)$$

上式中的γ表示正例和负例得分的最小间隔(margin)，实际使用时常取$\gamma=1$。事实上，将等级损失中的得分函数替换为样本被预测为某个类别的概率，则上式的形式与多分类情形下的 hinge 损失一致。TransE 采取的生成负例三元组的方法是，将正确的三元组的头实体、尾实体、关系三者之一随机替换为其他实体或关系，从而构成负例集合 Δ' ，这种方法称为均匀采样（与后面的伯努利采样相对比）。

在代码实现中，首先选取一个正例三元组 (h,r,t) ，再从 Δ' 中采样得到一个负例三元组 (h',r',t')，然后分别计算正例得分 $f_r(h,t)$ 和负例得

分$f_{r'}(h',t')$，若 $f_r(h,t)+\gamma-f_{r'}(h',t')>0$ ，则梯度下降更新 h,r,t,h',r',t'。

显然，TransE 模型在处理复杂关系建模（一对多、多对一、多对多关系）时会遇到困难，例如，对于一对多关系（美国，总统，奥巴马）和（美国，总统，特朗普），TransE 模型会使得尾实体向量奥巴马和特朗普的表示非常相似。事实上，这是由于对于不同的关系 γ，实体向量的表示总是相同的。

3.2 TransH

TransH 方法由中山大学信科院冯剑琳团队和 MSRA 联合提出，克服了 TransE 模型的上述缺点，使得同一个实体向量在不同关系下有不同的表示。TransH 模型对于每一个关系 r，假设有一个对应的超平面（关系 r 落于该超平面上），其法向量为 w_r，且有 $||w_r||_2=1$。

类似于 TransE 模型的翻译在该超平面上进行，具体地，首先将头实体 h 和尾实体 t 投影到该超平面上得到$h_\perp$和 $t_\perp$ ，即

$$h_\perp=h-w_r^T h w_r$$
$$t_\perp=t-w_r^T t w_r \quad (3)$$

进而，我们定义得分函数为

$$f_r(h,t)=\|h_\perp+r-t_\perp\|_{L1/L2} \quad (4)$$

TransE 及 TransH 对比如图 1。可见头实体 h 和尾实体 r 有更多的表示自由。只要满足 h 与 r 在超平面的投影为$h_\perp$即可。

3.3 TransR

虽然 TransH 模型使得同一实体在不同关系下通过投影有了不同的表示，但投影之后仍然处于原来的空间 R^n 中，这里 n 表示实体向量和关系向量均为 n 维。换言之，TransH 模型假设实体和关系处于相同的语义空间中，这在一定程度上限制了它的表示能力。下面的 TransR 模型改进了这一缺陷，提高了模型的表示能力。

值得一提的是，在论文中除了提出了 TranH 模型，另一贡献是提出了基于伯努利分布的采样方法。在原来的均匀采样中，容易将错误的负例引入到训练过程中来，例如，对于正例（美国，总统，奥巴马），随机替换奥巴马为罗斯福构成（美国，总统，罗斯福）作为负例，实际上由于罗斯福也是总统，这并不是一个负例。新的采样方法的动机是，对于一对多关系，我们以更大的概率来替换其头实体，对于多对一关系，我们以更大的概率来替换其尾实体。具体地，对于包含关系 [公式] 的所有三元组，我们定义两个统计量：

1.tph：平均每个头实体对应多少个尾实体

2.hpt: 平均每个尾实体对应多少个头实体

进而，取伯努利分布的参数为 $p=\frac{tph}{tph+hpt}$ ，即以概率 p 替换三元组的头实体，以概率 $1-p$ 替换三元组的尾实体。实际上，p 反映了一对多关系的失衡程度。通过基于伯努利分布的采样，我们就降低了引入错误的负例的概率。

3.4 TransR & CTransR

TransR 模型认为，不同的关系关注实体的不同属性（实体向量的不同维度），因此不同的关系应具有不同的语义空间。TransH 模型是为每个关系假定一超平面，将实体投影到这个超平面上进行翻译；而 TransR 模型是为每个关系假定一语义空间 R^m ，将实体映射到这个语义空间上进行翻译。这里 m 表示关系向量的维度为 m。TransR 模型可以形式化描述为:

$$h_r=hM_r,t_r=tM_r \quad (5)$$
$$f_r(h,t)=\|h+r-t\|_{L1/L2} \quad (6)$$

其中，$h,t\in R^n,r\in R^m,M_r\in R^{n\times m}$ 约束条件为$\|h+r-t\|$的 L2 范数均不大于 1。CTransR 的意思是 Cluster-based TransR. CTransR 对于每一个特定的关系 r，首先根据实体对 (h,t) 进行 AP 聚类（一种不需要指定类别数的聚类方法），实际上是对实体对的差值向量 $h-t$ 进行聚类，从而将关系 r 分

解为更细粒度的子关系 r_c ，CTransR 对每个r_c分别学习相应的向量表示。形式化地，CTransR 的得分函数可以描述为

$$f_r(h,t) = \|h_{r,c} + r_c - t_{r,c}\|_2^2 + \alpha\|r_c - r\|_2^2 \quad (7)$$

上式中第二项使得 r_c尽可能地接近 r。

TransR & CTransR 模型将原来的单个语义空间分离为实体空间和关系空间，提高了模型的表示能力，然而，TransR 模型仍然存在一些缺点：

1. 在同一个关系 r 下，头、尾实体使用相同的投影矩阵M_r，而头、尾实体可能类型或属性相差很大；

2. 投影矩阵仅与关系有关；

3. 参数多，计算复杂度高。

3.5 TransD

CTransR 模型相比于 TransR 模型，实际上就是考虑了同一个关系也有不同的类型，然而，实体也有不同的类型。举个例子，在 FB15k 中，有关系 location.location.partially containedby，它可以表示山川大河被某个国家包含，也可以表示山川大河被某个城市/州包含，也可以表示国家被大洲包含，还可以表示地区被国家包含。由于实体有不同的类型，因而使用相同的映射矩阵是不合理的，而且，映射矩阵不应只与关系有关，还应与头尾实体有关。以上就是 TransD 模型的动机，这实际上是一个更细粒度的扩展模型，本质上还是由实体语义空间和关系语义空间两个空间构成。

具体地，对于一个三元组 (h,r,t)，分别定义对应的投影向量h_p,r_p,t_p，其中 p 表示投影（projection），再定义两个投影矩阵M_{rh}，M_{rt}来将实体从实体空间映射到关系空间。

$$M_{rh} = r_p h_p^T + I^{m\times n} \quad (8)$$

$$M_{rt} = r_p t_p^T + I^{m\times n} \quad (9)$$

这里，$r_p,r \in R^m$，$h,h_p,t,t_p \in R^n$ ，$M_{rh},M_{rt} \in R^{m\times n}$,$I^{m\times n}$表示单位阵，它的意思是说用单位阵来初始化投影矩阵。可见，在上式中，对头实体应用投影矩阵 M_{rh} ，它不仅与关系有关，还与头实体有关；对尾实体应用投影矩阵M_{rt}，它也不仅与关系有关，还与尾实体有关。利用这两个投影矩阵，可以得到头实体和尾实体在关系空间的投影

$$h_\perp = M_{rh},\ t_\perp = M_{rt} \quad (10)$$

从而得分函数为

$$f_r(h,t) = \|h_\perp + r - t_\perp\|_{L1/L2} \quad (11)$$

4、基于翻译模型的对比

4.1 理论对比

显然，TransE 是 TransD 的一个特例，当 m = n 且所有投影向量均为零时，TransD 就退化为了 TransE. 与 TransH 对比，显然需先令 TransD 的 m = n，再分别写出 TransH 和 TransD 的实体投影后的向量：

TransH:$h_\perp = h - w_r^T h w_r$

TransD: $h_\perp = M_{rh}h = h + h_p^T h r_p$

因为 $h_p^T h r_p = r_p h_p^T h$ ，因此为便于格式上的对比，我们将 TransD 中的$h_\perp$写成上述形式。可见，当 $m = n$ 时，TransD 与 TransH 唯一的区别在于 TransD 中的投影向量不仅与关系有关，还与实体有关。

对比 TransR 模型，TransD 为头实体和尾实体分别设置了投影矩阵，另外，注意到在 TransD 中公式经过展开之后没有矩阵-向量乘法操作，这相比于 TransR 模型降低了计算复杂度，更适用于大规模知识图谱的计算。

4.2 性能对比

评估不同知识表示学习方法的优劣的主要指标就是在知识图谱的一些典型任务上的表现，比如三元组分类（triplet classification）和链接预测（link prediction）。图二为各模型在 WordNet 和 Freebase 中各模型在两任务上的表现，TransE, TransH,

TransR, CTransR, TransD 包括了使用 bern 和 unif 采样的结果。图三为链接预测的结果。

Data sets	WN18				FB15K			
Metric	Mean Rank		Hits@10		Mean Rank		Hits@10	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
Unstructured (Bordes et al. 2012)	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL (Nickle, Tresp, and Kriegel 2011)	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE (Bordes et al. 2011)	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (linear) (Bordes et al.2012)	545	533	65.1	74.1	274	154	30.7	40.8
SME (Bilinear) (Bordes et al. 2012)	526	509	54.7	61.3	284	158	31.3	41.3
LFM (Jenatton et al. 2012)	469	456	71.4	81.6	283	164	26.0	33.1
TransE (Bordes et al. 2013)	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif) (Wang et al. 2014)	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern) (Wang et al. 2014)	401	388	73.0	82.3	212	87	45.7	64.4
TransR (unif) (Lin et al. 2015)	232	219	78.3	91.7	226	78	43.8	65.5
TransR (bern) (Lin et al. 2015)	238	225	**79.8**	92.0	198	77	48.2	68.7
CTransR (unif) (Lin et al. 2015)	243	230	78.9	92.3	233	82	44.0	66.3
CTransR (bern) (Lin et al. 2015)	231	218	79.4	92.3	199	75	48.4	70.2
TransD (unif)	242	229	79.2	**92.5**	211	**67**	**49.4**	**74.2**
TransD (bern)	**224**	**212**	79.6	92.2	**194**	91	**53.4**	**77.3**

图二 模型性能表现

Tasks	Prediction Head (Hits@10)				Prediction Tail (Hits@10)			
Relation Category	1-to-1	1-to-N	N-to-1	N-to-N	1-to-1	1-to-N	N-to-1	N-to-N
Unstructured (Bordes et al. 2012)	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE (Bordes et al. 2011)	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME (linear) (Bordes et al.2012)	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME (Bilinear) (Bordes et al. 2012)	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE (Bordes et al. 2013)	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH (unif) (Wang et al. 2014)	66.7	81.7	30.2	57.4	63.7	30.1	83.2	60.8
TransH (bern) (Wang et al. 2014)	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2
TransR (unif) (Lin et al. 2015)	76.9	77.9	38.1	66.9	76.2	38.4	76.2	69.1
TransR (bern) (Lin et al. 2015)	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1
CTransR (unif) (Lin et al. 2015)	78.6	77.8	36.4	68.0	77.4	37.8	78.0	70.3
CTransR (bern) (Lin et al. 2015)	81.5	89.0	34.7	71.2	80.8	38.6	90.1	73.8
TransD (unif)	80.7	85.8	**47.1**	**75.6**	80.0	**54.5**	80.7	**77.9**
TransD (bern)	**86.1**	**95.5**	**39.8**	**78.5**	**85.4**	**50.6**	**94.4**	**81.2**

图三 链接预测

参考文献

[1] Bordes A., Usunier N., Garcia-Dur an A. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of NIPS.* pags:2787-2795.

[2] Wang Z., Zhang J., Feng J. and Chen Z. 2014.Knowledge graph embedding by translating on hyperplanes. In Proceedings of AAAI. pags:1112-1119.

[3] Lin Y., Zhang J., Liu Z., Sun M., Liu Y., Zhu X.2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of AAAI.

[4] Ji G , He S , Xu L , et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]// Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing. 2015.

[5] 刘知远，孙茂松，林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2):247-261.

《智能信息处理》课程作业

基于表示学习的翻译模型综述

田应彪

作业	分数
得分	

2020 年 12 月 10 日