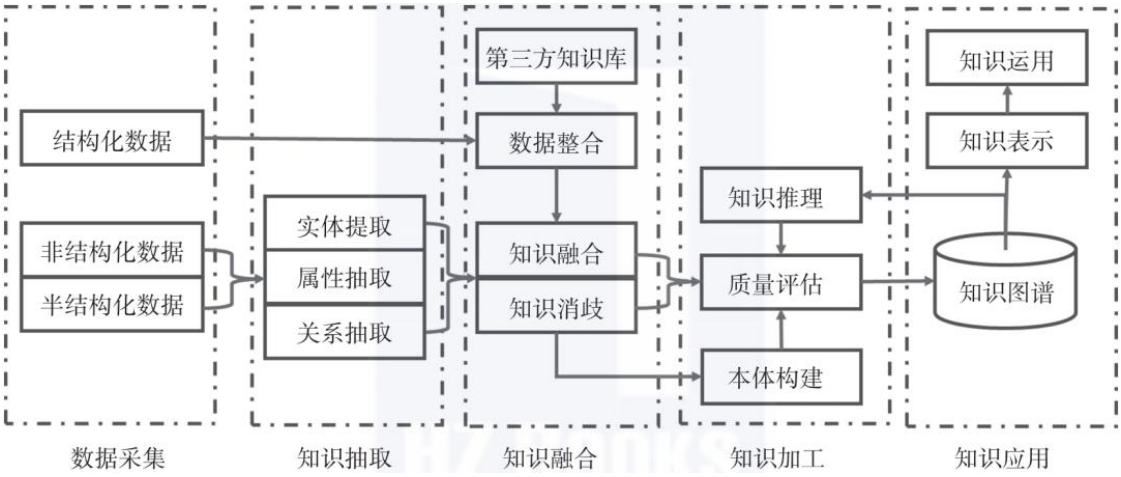
知识图谱算法总结

知识图谱系统包括知识抽取、融合、推理等，每个环节均涉及不同的算法，目前各步骤所用到的算法根据现状和需求不同存在不同挑战。



一、知识抽取

知识抽取是实现自动化构建大规模知识图谱的重要技术，其目的在于从不同来源、不同结构的数据中进行知识提取并存入知识图谱中。知识抽取的数据源可以是**结构化数据**（如链接数据、数据库）、**半结构化数据**（如网页中的表格、列表）或者**非结构化数据**（即纯文本数据）。面向不同类型的数据源，知识抽取涉及的关键技术和需要解决的技术难点有所不同。结构化数据可以从关系数据库中抽取知识，抽取工具包括 D2R、Virtuoso、Orcle SW、Morph 等。面向半结构化数据的知识抽取可基于固定模式对实体信息进行抽取，包括 abstract，infobox，category，page link 等。面向无结构化数据的知识抽取是当前知识图谱构建的技术瓶颈，关键技术包括**实体识别和链接**、**关系抽取**和**事件抽取**。

1.1 实体识别

命名实体识别是指识别文本中的命名性实体，并将其划分到指定类别的任务。常用实体类别包括人名、地名、机构名、日期等。

1.1.1 基于规则的实体识别

早期的命名实体识别方法主要采用人工编写规则的方式进行实体抽取。这类方法首先构建大量的实体抽取规则，一般由具有一定领域知识的专家手工构建。然后，将规则与文本字符串进行匹配，识别命名实体。这种实体抽取方式在小数据集上可以达到很高的准确率和召回率，但随着数据集的增大，规则集的构建周期变长，并且移植性较差。

1.1.2 基于统计模型方法的实体识别

基于机器学习的方法更加健壮和灵活，比较客观，不需要太多人工干预和领域知识，缺点是依赖人工设计特征。主要包括隐马尔可夫模型（Hidden Markov Model, HMM）、最大熵马尔可夫模型 (Maximum Entropy Markov Model, MEMM)、条件随机场（Conditional Random Fields，CRF）、支持向量机（Support Vector Machine，SVM）等。

隐马尔可夫模型（Hidden Markov Model,HMM）和条件随机场（Conditional Random Field,CRF）是两个常用于标注问题的统计学习模型，也被广泛应用于实体抽取问题。HMM 是一种有向图概率模型，模型中包含了隐藏的状态序列和可观察的观测序列。每个状态代表了一个可观察的事件，观察到的事件是状态的随机函数。HMM 模型结构如图 1-1 所示，每个圆圈代表一个随机变量，随机变量 xt 是 t 时刻的隐藏状态；随机变量 yt 是 t 时刻的观测值，图中的箭头表示条件依赖关系，HMM 模型有两个基本假设：

●在任意 t 时刻的状态只依赖于其前一时刻的状态，与其他观测及状态无关。

●任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其他观测及状态无关。

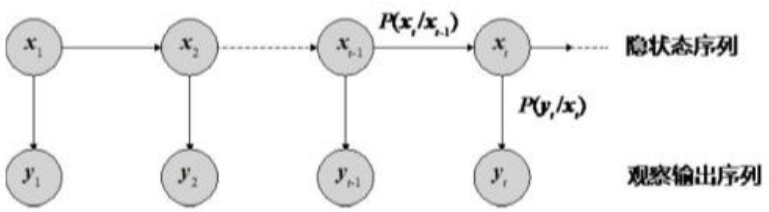


图 1-1　HMM 模型结构

在应用于命名实体识别问题时，HMM 模型中的状态对应词的标记，标注问

题可以看作是对给定的观测序列进行序列标注。基于 HMM 的有代表性的命名实体识别方法可参考文献（(1) Liu F,Zhao J,Lv B,et al.Product Named Entity Recognition Based on Hierarchical Hidden Markov Model.Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing 2005. (2) Morwal S,Jahan N,Chopra D.Named Entity Recognition Using Hidden Markov Model（HMM）.International Journal on Natural Language Computing （IJNLC）,2012,1（4）:15-23）。

CRF 是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型。在序列标注问题中，线性链 CRF 是常用的模型，其结构如图 1-2 所示。在序列标注问题中，状态序列变量 x 对应标记序列，y 表示待标注的观测序列。

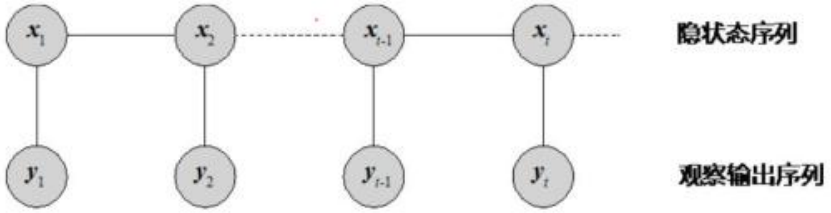


图 1-2　线性链 CRF 模型结构

给定训练数据集，模型可以通过极大似然估计得到条件概率模型；当标注新数据时，给定输入序列 y，模型输出使条件概率 P（x|y）最大化的 x*。美国斯坦福大学开发的命名实体识别工具 Stanford NER 是基于 CRF 的代表性系统[1]。

1.1.3 基于深度学习方法的实体识别

与传统统计模型相比，基于深度学习的方法直接以文本中词的向量为输入，通过神经网络实现端到端的命名实体识别，不需要太多人工干预和领域知识，缺点是需要人工标注数据，数据稀疏问题比较严重。

目前，用于命名实体识别的神经网络主要有卷积神经网络（Convolutional Neural Network,CNN）、循环神经网络（Recurrent Neural Network,RNN）以及引入注意力机制（Attention Mechanism）的神经网络。一般地，不同的神经网络结构在命名实体识别过程中扮演编码器的角色，它们基于初始输入以及词的上下文信息，得到每个词的新向量表示；最后再通过 CRF 模型输出对每个词的标注结果。

（1）LSTM-CRF 模型。图 1-3 展示了 LSTM-CRF 命名实体识别模型。该模型使用了长短时记忆神经网络 (Long Shot-Term Memory Neural Network,LSTM) 与 CRF 相结合进行命名实体识别。该模型自底向上分别是 Embedding 层、双向 LSTM 层和 CRF 层。Embedding 层是句子中词的向量表示，作为双向 LSTM 的输入，通过词向量学习模型获得。双向 LSTM 层通过一个正向 LSTM 和一个反向 LSTM，分别计算每个词考虑左侧和右侧词时对应的向量，然后将每个词的两个向量进行连接，形成词的向量输出；最后，CRF 层以双向 LSTM 输出的向量作为输入，对句子中的命名实体进行序列标注。经过实验对比发现，双向 LSTM 与 CRF 组合的模型在英文测试数据上取得了与传统统计方法最好结果相近的结果，而传统方法中使用了大量的人工定义的特征以及外部资源；在德语测试数据上，深度学习模型取得了比统计学习方法更优的结果。

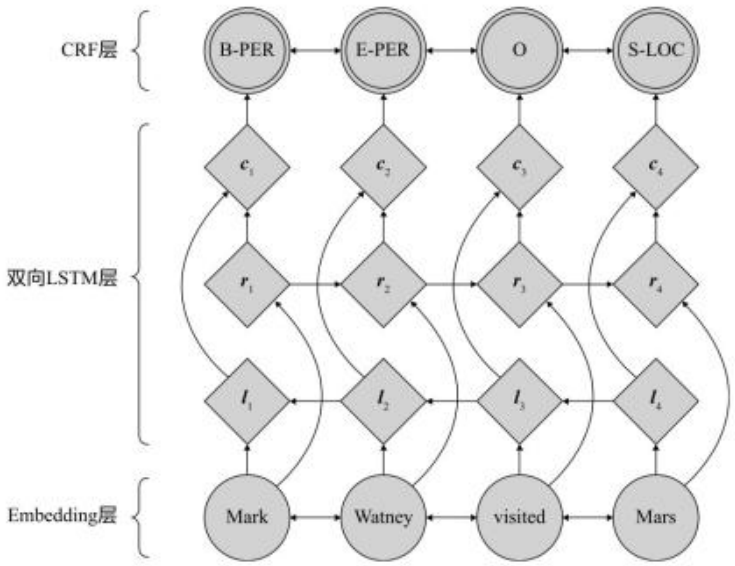


图 1-3　LSTM-CRF 命名实体识别模型

（2）LSTM-CNNs-CRF 模型。MA Xuezhe 等人发表于 ACL2016 的论文提出了将双向 LSTM、CNN 和 CRF 相结合的序列标注模型（）。并成功地应用于命名实体识别问题中。该模型与 LSTM-CRF 模型十分相似，不同之处是在 Embedding 层中加入了每个词的字符级向量表示。图 1-4 展示了获取词语字符级向量表示的 CNN 模型，该模型可以有效地获取词的形态信息，如前缀、后缀等。模型 Embedding 层中每个词的向量输入由预训练获得的词向量和 CNN 获得的字符级向量连接而成，通过双向 LSTM 和 CRF 层获得词的标注结果。LSTM-CNNs-CRF 序列标注模型框架如图 1-5 所示。在 CoNLL-2003 命名实体识

别数据集上，该模型获得了 91.2%的 F1 值。

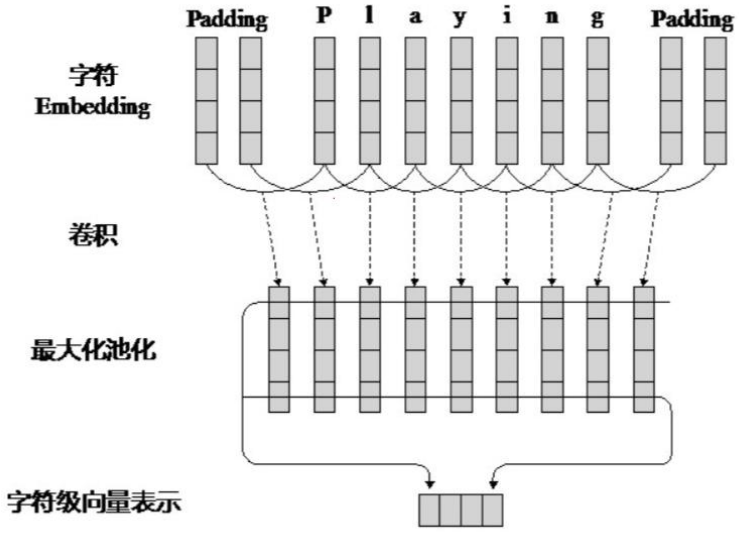


图 1-4 获取词语字符级向量表示的 CNN 模型

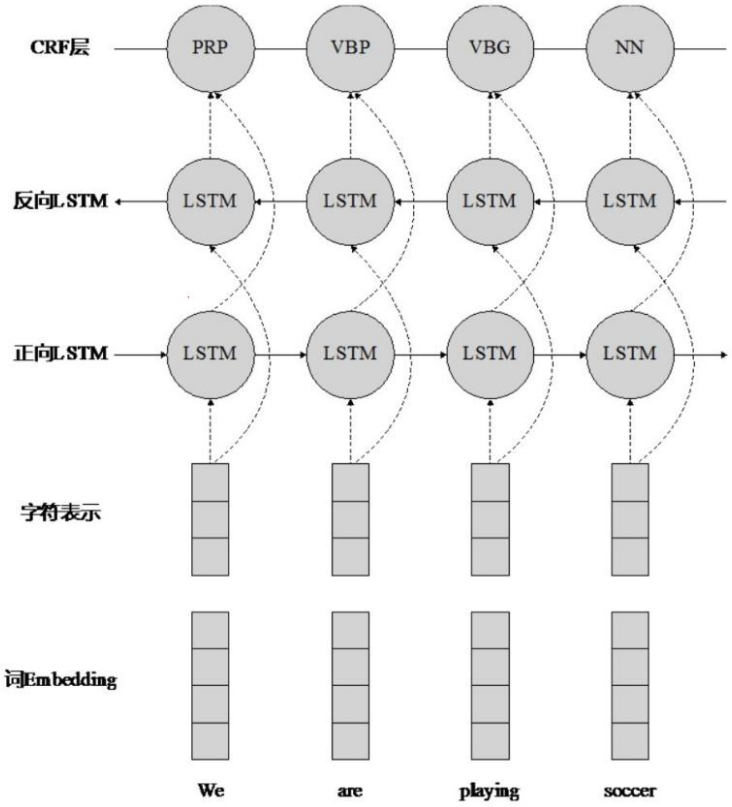
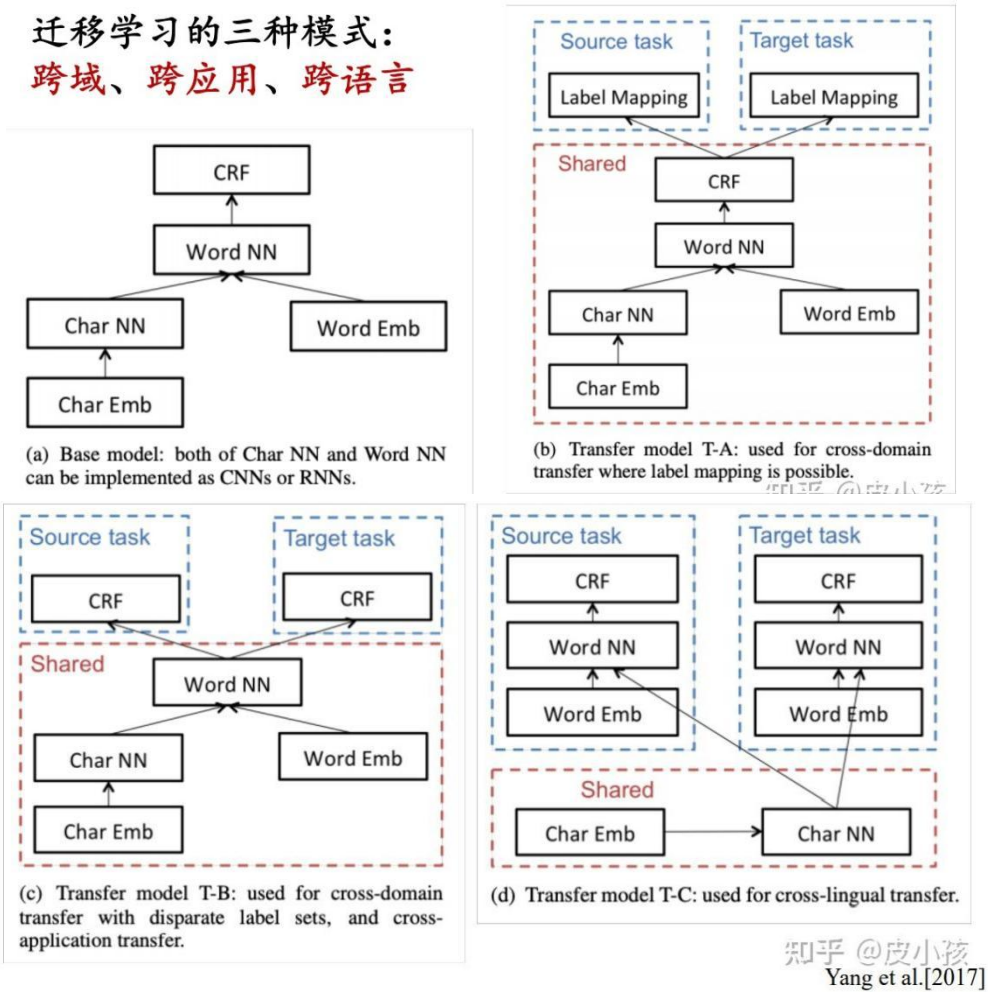


图 1-5 LSTM-CNNs-CRF 序列标注模型框架

（3）基于**注意力机制的神经网络**模型。在自然语言处理领域，基于注意力机制的神经网络模型最早应用于解决机器翻译问题，注意力机制可以帮助扩展基本的编码器-解码器模型结构，让模型能够获取输入序列中与下一个目标词相关的信息。

1.1.4 基于迁移学习的实体识别

迁移学习的核心在于找到新问题和原问题之间的相似性，迁移学习属于机器学习的一个种类，与传统的机器学习方法相比，**迁移学习的训练和测试数据可以服从不同的分布**，其次**不需要足够的数据标注**，此外模型可以在不同任务之间迁移，而不需要在每个任务分别建模。基础模型是 CNN 或 RNN。

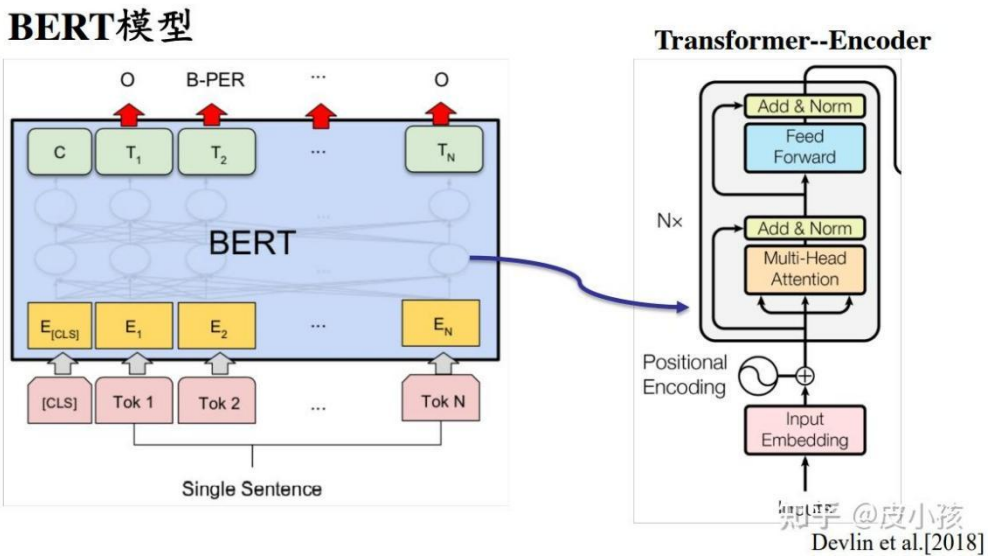


Yang et al.[2017]

1.1.5 基于预训练方法的实体识别

BERT 模型重新设计了语言模型预训练阶段的目标任务，提出了遮挡语言模型（Masked LM）和下一个句子预测（NSP）。该模型的表现 F1 值可达 92%以

上。



1.2 关系抽取

关系抽取即自动识别实体之间具有的某种语义关系 (指隐藏在句法结构后面由词语的语义范畴建立起来的关系, 句法关系: 位置关系、替代关系、同现关系), 从文本中抽取实体及实体之间的关系。关系抽取和实体抽取密切相关, 一般是在识别出文本的实体后, 再抽取实体之间可能存在的关系。当前, 关系抽取方法可以分为基于模板的方法、基于监督学习的方法和基于弱监督学习的方法。

1.2.1 基于模板的方法

使用模式（规则）挖掘关系，基于触发词/字符串等或者基于依存句法。其优点是人工规则的准确率高、可以为特定领域制定以及在小规模数据集上容易实现，构建简单；缺点是低召回率、可移植性差等。

1.2.2 基于有监督学习的实体关系抽取

基于监督学习的关系抽取方法**将关系抽取转化为分类问题**, 在大量标注数据的基础上, 训练有监督学习模型进行关系抽取。利用监督学习方法进行关系抽取的一种常见方法是训练一个层叠的二分类器(或常规的二分类器)来确定两个实体之间是否存在特定的关系。这些分类器将文本的相关特征作为输入, 从而要求文本首先由其他 NLP 模型进行标注。典型的特征有: **上下文单词、词性标注、实体间的依赖路径、NER 标注、tokens、单词间的接近距离**等。有监督关系抽取任务并没有实体识别这一子任务，因为数据集中已经标出了 subject 实体和

object 实体分别是什么，所以全监督的关系抽取任务更像是做分类任务。模型的主体结构都是特征提取器+关系分类器。特征提取器比如 CNN，LSTM，GNN，Transformer 和 BERT 等。

传统的基于监督学习的关系抽取是一种依赖特征工程的方法，近年来有多个基于深度学习的关系抽取模型被研究者们提出。深度学习的方法不需要人工构建各种特征，其输入一般只包括句子中的词及其位置的向量表示。目前，已有的基于深度学习的关系抽取方法主要包括**流水线方法**和**联合抽取方法**两大类。流水线方法将识别实体和关系抽取作为两个分离的过程进行处理，两者不会相互影响；关系抽取在实体抽取结果的基础上进行，因此关系抽取的结果也依赖于实体抽取的结果。联合抽取方法将实体抽取和关系抽取相结合，在统一的模型中共同优化；联合抽取方法可以避免流水线方法存在的错误积累问题。

基于深度学习的流水线关系抽取算法包括 CR-CNN 、Attention BLSTM，一般实验结果表明，**增加注意力层**可以有效地提升关系分类的结果。

在关系抽取问题方面，还有许多其他属于流水线方法的深度学习算法。图 1-6 列出了一些具有代表性的流水线方法在 SemEval-2010 Task 8 数据集上的结果对比 （Att-BLSTM, Att-Pooling-CNN、depLCNN+NS、DepNN、CR-CNN、CNN+Softmax）。

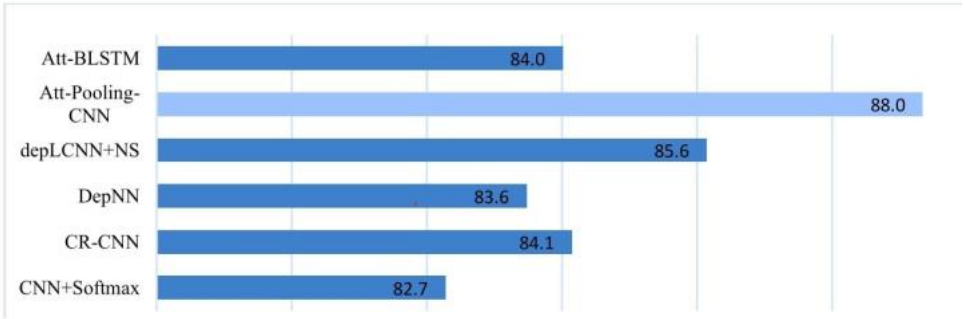


图 1-6 关系抽取模型在 SemEval-2010 Task 8 数据集 F1 值对比（%）

相关论文:

[1] Santos,Cicero Nogueira Dos,B Xiang,et al.Classifying Relations by Ranking with Convolutional Neural Networks.Computer Science,2015:132-137.

[2] Zhou P,Shi W,Tian J,et al.Attention-based Bidirectional Long Short-Term Memory Networks for Relation Classification.Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics （Volume 2: Short Papers）,2016,2:207-212.

[3] Wang L,Cao Z,de Melo G,et al.Relation Classification via Multi-Level Attention CNNs.Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016,1:1298-1307.

[4] Xu K,Feng Y,Huang S,et al.Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling.Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,2015:536-540.

[5] Liu Y,Wei F,Li S,et al.A Dependency-Based Neural Network for Relation Classification.Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing（Volume 2:Short Papers）,2015,2:285-290.

[6] Zeng D,Liu K,Lai S,et al.Relation Classification via Convolutional Deep Neural Network.Proceedings of COLING 2014,the 25th International Conference on Computational Linguistics: Technical Papers,2014 :2335-2344.

1.2.3 基于弱监督学习的实体关系抽取

该方法优点包括：（1）减少构建标记数据所需要耗费的人力；（2）充分利用比较容易获得的无标记数据。主要包括以下几种算法：

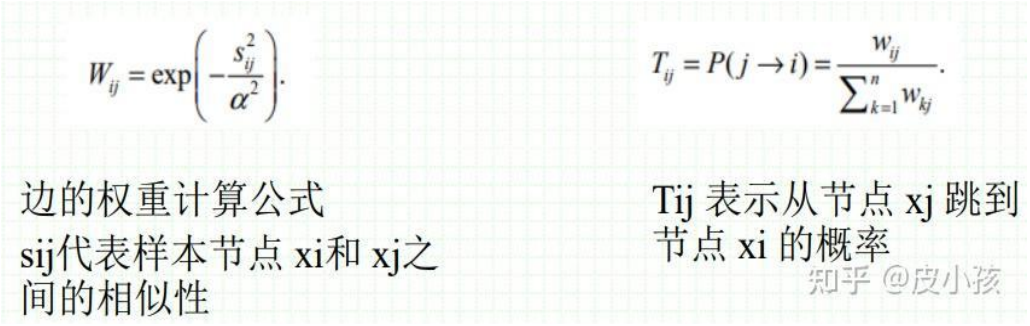
A. **Bootstrapping 算法**：首先初始化，然后进入扩张阶段，生成新的模式和新的实体，设置迭代次数。该方法的主要难点是语义漂移，缺点是对于上下文依赖的关系不友好，且无法区分细粒度的关系；优点是对于特定关系表现很好。

B. **标签传播算法**（Label Propagation Algorithm，LPA）：是一种半监督学习方法，用于向未标记样本分配标签。标签传播算法的核心思想是相似的数据应该具有相同地 label。该算法通过将所有样本通过相似性构建一个边有权重的图，然后各个样本在其相邻的样本间进行标签传播。因此标签传播算法在本质上就很适合用来做**网络结构中的社区发现**。基本步骤如下：

（1）将所有的实体对看做是图上的节点，将实体对间的距离看做是边；

（2）将一部分标注好的节点看做源头向其他节点传播，而权重值越高的边上传播的速度越快；

（3）将相似度高的节点聚在一类，类别信息通过传播过来的标注信息来判别。



C. **协同学习算法**：协同训练（Co-learning）是一类基于“**分歧**”的半监督学习方法，它最初是针对“**多视图**”数据设计的。为了更好的介绍协同训练，我们这里先介绍什么是多视图数据。在不少现实应用中，一个数据对象往往同时拥有有多个“属性集”，每个属性集就构成了一个视图。理论证明显示出，若两个视图充分且条件独立，则可以利用未标记样本通过协同训练将弱分类器的泛化性能提升到任意高。不过，视图的条件独立性在现实任务中通常很难满足，因此提升幅度不会那么大。

基本流程：

（1）使用两个不同的分类器；

（2）使用相互独立的特征在两个训练集上训练，并分别在未标注集上测试；

（3）选取置信度高的实例扩展到另一个分类器的训练集中；

（4）如此迭代若干次，当精度达到阈值时停止。

D. **远程监督算法**有一个非常重要的假设：对于一个已有的知识图谱（论文用的 Freebase）中的一个三元组（由一对实体和一个关系构成），假设外部文档库（论文用的 Wikipedia）中任何包含这对实体的句子，在一定程度上都反映了这种关系。基于这个假设，远程监督算法可以**基于一个标注好的小型知识图谱，给外部文档库中的句子标注关系标签**，相当于**做了样本的自动标注**。

远程监督关系抽取方法可以利用丰富的知识图谱信息获取训练数据，有效地减少了人工标注的工作量。但是，基于远程监督的假设，大量噪声会被引入到训练数据中，从而引发语义漂移的现象。

为了改进远程监督实体关系抽取方法，一些研究围绕如何克服训练数据中的噪声问题展开。最近，多示例学习、采用注意力机制的深度学习模型以及强化学习等模型被用来解决样例错误标注的问题，取得了较好的效果。基于句子级注意力和实体描述的神经网络关系抽取模型 APCNNs 和基于强化学习的关系分类模型 CNN-RL 均取得了较好的效果。

参考文献：

[1] Ji G-L,Liu K,He S,et al.Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions.Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence :AAAI Press,2017:3060-3066.
[2] Ji G-L,Liu K,He S,et al.Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions.Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence :AAAI Press,2017:3060-3066.

1.2.4 基于无监督学习的实体关系抽取

由于有监督和半监督机器学习方法需要事先确定关系类型, 而实际上在大规模语料中, 人们往往无法预知所有的实体关系类型。有些研究者基于**聚类**的思想, 利用无监督机器学习的方法, 尝试解决这个问题。该算法基本步骤是**首先采用某种聚类方法将语义相似度高的实体对聚为一类, 再选择具有代表性的词语来标记这类关系。**无监督的关系抽取方法最早是由 Hasegawa 在 2004 年的 ACL 会议上提出的（Discovering Relations Among Named Entities from Large Corpora [Hasegawa, 2004]），之后的很多方法多是在 Hasegawa 的基础上改进而来。一种用于**多级聚类**的无监督抽取方法（Preemptive Information Extraction using Unrestricted Relation Discovery [Shinyama, 2006]），Quan 提出了一种基于模式聚类和句子解析的无监督方法来处理生物医学关系抽取。模式聚类算法基于多项式核方法, 通过无标签数据识别交互词, 然后将这些交互词用于实体对之间的关系抽取（An unsupervised text mining method for relation extraction from biomedical literature [Quan C, 2014]）。URES 是一种主要的基于非聚类的无监督关系抽取系统，它可以完全无监督地从网页中抽取关系（URES : an Unsupervised Web Relation Extraction System [Quan C, 2006]）。

1.3 属性抽取

属性抽取要识别出实体的属性值, 而属性值结构是不确定的, 因此大部分是基于规则进行抽取，。面向的也是网页。

1.3 事件抽取

事件是指发生的事情，通常具体时间、地点、参与者等属性。事件的发生可能是因为一个动作的产生或者系统状态的改变。事件抽取是指从文本中抽取用户感兴趣的事件信息，并以**结构化的形式**呈现。例如，从恐怖袭击事件的新闻报道中识别袭击发生的地点、时间、袭击目标和受害人等信息。

一般地，事件抽取任务包含的子任务有：

●识别事件触发词及事件类型；

●抽取事件元素的同时判断其角色；

●抽出描述事件的词组或句子；

●事件属性标注；

●事件共指消解。

已有的事件抽取方法可以分为流水线方法和联合抽取方法两大类。

1.3.1 事件抽取的流水线方法

流水线方法将事件抽取任务分解为一系列基于分类的子任务，包括事件识别、元素抽取、属性分类和可报告性判别；每一个子任务由一个机器学习分类器负责实施。一个基本的事件抽取流水线需要的分类器包括：

(1) 事件触发词分类器。判断词汇是否为事件触发词，并基于触发词信息对事件类别进行分类。

(2) 元素分类器。判断词组是否为事件的元素。

(3) 元素角色分类器。判定事件元素的角色类别。

(4) 属性分类器。判定事件的属性。

(5) 可报告性分类器。判定是否存在值得报告的事件实例。

各个阶段的分类器可以采用机器学习算法中的不同分类器，例如最大熵模型、支持向量机等。

图 Pipeline 方式

主要算法包括：DMCNN（Dynamic Multi-Pooling Convolutional Neural Network），相比于传统的 max pooling 的直接对每个 feature map 做一个 max 操作来提取句子中最有用的信息, Dynamic Multi-Pooling 是将每个 feature map 根据候选元素和触发词来进行分割操作，即把每个 feature map 根据元素和触发词切分成三块，然后分别计算 max value。

1.3.2 事件抽取的联合抽取方法

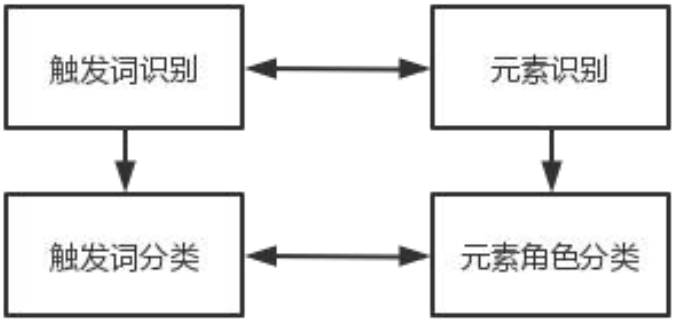


图 Joint Model 方式

JRNN（Join Recurrent Neural Network）的编码部分使用 word2vec 训练的词

向量，将实体类型作为特征并编码为向量，当在依赖关系树中存在与 wi 相连的边时，则第 i 维的值为 1。

Joint Model 相比于 Pipeline 的优势：

- 避免了误差累计传播导致模型性能下降的问题；
- 使用一个模型同时抽取出所有的事件信息；
- 使用从整体结构中学到的全局特征来提升对局部信息的预测能力。

1.3.3 基于强化学习的方法

强化学习可以通过让机器去模仿人的行为，对正确的行为给予奖励，错误的行为获得惩罚。代表性算法为 GAIL（Event Extraction with Generative Adversarial Imitation Learning，Zhang et al.[2018]），算法的核心思想是使用强化学习的思想通过 Q-Learning 的方式对句子做序列标注，标记出句子里面的 entity 和 trigger。然后使用 policy gradient 判定事件中 entity 对应的 argument role。训练过程中强化学习使用到的 reward 由 GAN 动态生成。实验证明，GAIL 比 JRNN 的性能有所提高。

二、知识融合

通过知识提取实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标。但是由于知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、层次结构缺失等问题，因此必须要进行知识的融合。大数据环境下的知识融合是面向知识服务和决策支持，以海量的多源异构数据、信息、知识为基础，利用融合算法和规则，在语义层次上组合、推理、创造出新知识的过程。

2.1 基于语义的知识融合

基于语义的方法可以细分为基于语义规则和基于本体论的方法。

2.1.1 基于语义规则的方法

基于语义规则（seman ticules）的方法主要依赖已定义的逻辑运算规则，对融合条件进行约束，进而实现知识融合。J. Gou 等描述了一组比较规则，来明确融合条件，同时通过**结果格式**和**语义规则**过滤知识对象，从而实现将知识对象分成不同粒度层次，并且指导融合过程，以规避不合逻辑的结果。Z. Feng 等使用

语义相似性来实现语义消歧，使提取的初级词汇链之间实现了知识融合。G. Wang 等出了一种基于语义元素重构的多源流程创新知识融合算法，并提出了相应的语义冲突解决规则，该算法可以有效地支持所贡献知识的初步自动融合。

基于语义规则的知识融合，较大程度地需要人工自定义规则，对于小规模数据量、特定领域还较为有效，但是面对大数据环境，尤其是网络大数据环境，在融合效率和适用性上都有所不足。

2.1.2 基于本体论的方法

本体（ontology）具有良好的概念层次结构，可支持逻辑推理，并且能够通过概念模型实现语义层知识的描述，非常适合知识的表达。很多知识融合的框架是基于本体来构建的，往往和知识元一起同属于框架的数据层。在这样的知识融合框架中，利用知识组织理论，通过**本体映射和融合**，实现知识融合。本体映射算法主要包括：SMART 算法、AnchorPROMPT 算法、基于相似度计算的算法、基于规则学习的方法、随机游走算法、基于图正则化模型、基于树结构的多策略本体映射算法、基于半自动算法生成本体互操作的映射规则等。

2.2 基于信息融合算法的知识融合

知识融合是由信息融合发展而来的，因此信息融合算法在知识融合中得到了继承和发展。在知识融合中使用的信息融合算法主要有**贝叶斯网络**和 **D-S 证据理论**。

2.2.1 贝叶斯网络

贝叶斯网络（bayesiannetwork）是一种基于概率推理的图形化网络，可用于发现数据间的潜在关系。贝叶斯网络在解决不确定性和不完整性问题上应用广泛，计算简单直接。将概率模型表示成贝叶斯知识基础，并提出基于贝叶斯的知识融合算法。

然而，贝叶斯网络在面临大规模数据时，产生的网络结构也会较大，因此，在巨大的网络结构中寻找最佳网络会使得存储空间不足，算法效率低下。此算法还要求知识之间的观测独立性、知识先验概率的可预知性，这在大数据环境下都难以满足。

2.2.2 D-S 证据理论

D-S 证据理论是关于证据和可能性推理的理论，可以消除不确定因素。D-S 证据理论是贝叶斯基础上的近一步发展，支持比贝叶斯更弱的条件。D-S 证据理

论的融合结果受知识库正确性和完整性的影响较大, 在知识融合的决策层或专家知识融合中适用性较高，但不适应于大规模数据的知识融合。

2.3 图模型方法

图模型可以是概率图、主题图或关系图。通过从其他类型数据中获得先验知识, 从而为知识分配一个概率, 可看做是图上的链路预测问题。图模型方法基于外部信息源提供的闭合知识集，因此其在大数据环境下的扩展性和适用性不强。

2.4 知识融合算法研究趋势

（1）提高知识融合的效率

未来还需开发更为高效精准的知识融合算法。虽然目前的知识融合算法能够适应知识来源的多样性，但还应进一步评估知识本身的真实性和可靠性，并且对知识的结构进行深入、清晰的分析。

（2）构建实时动态融合机制

在大数据环境，尤其是网络大数据环境下，知识时刻发生变化，并且不同来源知识的更新频率还不一致。这种动态性不仅体现在随时间而变化，还体现在不同的人、不同的外界环境、不同的背景和经验所导致的知识变化。而当前的知识融合方法主要是针对静态知识，对动态知识的融合还处于探索阶段。未来在融合过程中不但要加入知识的时间信息，开发动态融合方法，而且要构建能够实时动态反馈调整的融合机制，以应对大数据环境的动态性挑战。

（3）开展大数据实证应用研究

在当前的知识融合研究中，知识融合框架、方法研究较多，实证研究较少。在少有的实证研究中，数据规模小，数据的异构复杂程度小，使得现有融合方法在实现大数据环境知识融合的有效性上还有待进一步证明。因此，迫切需要开展真正意义上的大数据知识融合实证研究。

参考文献:

[1] 朱祥，张云秋. 近年来知识融合研究进展与趋势[J]. 图书情报工作, 2019, 63(16): 143-150.

三、知识计算

知识计算是领域知识图谱能力输出的主要方式，通过知识图谱本身能力为传统的应用形态赋能，提升服务质量效率。其中图挖掘计算和知识推理是最具代表性的两种能力，如何将这两种能力与传统应用相结合是需要解决的一个关键问题。

3.1 图挖掘计算

图挖掘计算指基于图论的相关算法，实现对图谱的探索与挖掘。图计算能力可辅助传统的推荐、搜索类应用。知识图谱中的图算法一般包括**图遍历、最短路径、权威节点分析、族群发现最大流算法、相似节点**等，大规模图上的算法效率是图算法设计与实现的主要问题。

3.2 知识推理

面向知识图谱的知识推理旨在根据已有的知识推理出新的知识或识别错误的知识。

主要任务：通过**规则挖掘**对知识图谱进行**补全与质量校验、链接预测、关联关系推理**与**冲突检测**等。

主要方法：（1）基于逻辑规则的推理；（2）基于图结构的推理；（3）基于分布式表示学习的推理；（4）基于神经网络的推理；（5）混合推理。

3.2.1 基于规则的推理

基于规则的推理通过定义或学习知识中存在的规则进行挖掘与推理。

AMIE

AMIE 是基于不完备基于不完备知识库的**关联规则挖掘算法**（Association Rule Mining under Incomplete Evidence），通过依次学习预测每种关系的规则：对于每种关系，从规则体为空的规则开始，通过三种操作扩展规则体部分，保留支持度大于阈值的候选（闭式）规则。这三种操作分别为：

（1）添加悬挂边：悬挂边是指边的一端是一个未出现过的变量，而另一端（变量或常量）是在规则中出现过的；

（2）添加实例边：实例边与悬挂边类似，边的一端也是在规则中出现过的变量或常量，但另一端是未出现过的常量，也就是知识库中的实体；

（3）添加闭合边：闭合边则是连接两个已经存在于规则中的元素（变量或常量）的边。

评估准则：支持度（同时符合规则体和规则头的实例数目）、置信度（支持度除以仅符合规则体的实例数目）、PCA 置信度。

优点：一是可解释性强；二是自动发现推理规则。

缺点：搜索空间大，且生成的规则覆盖度低，最终模型的预测效果也较差。

3.2.2 基于图结构的推理

路径排序算法（PRA）

PRA 是一种将关系路径作为特征的推理算法，通常用于知识图谱中的链接预测任务。因为其获取的关系路径实际上对应一种霍恩子句，PRA 计算的路径特征可以转换为逻辑规则，便于人们发现和理解知识图谱中隐藏的知识。

PRA 的基本思想是通过发现连接两个实体的一组关系路径来预测实体间可能存在的某种特定关系。

算法概述：（1）特征抽取（生成）. 特征抽取（生成并选择路径特征集合）

方法：随机游走，广度优先搜索，深度优先搜索

（2）特征计算（计算每个训练样例的特征值）

方法：随机游走概率，布尔值（出现/不出现），出现频次/概率

（3）分类器训练（根据训练样例，为每个目标关系训练一个分类器）

方法：单任务学习（为每个关系单独训练二分类器）；多任务学习（不同关系联合学习）

优点：一是可解释性强；二是自动发现推理规则

缺点：一是处理低频关系效果不好；二是处理低连通图（数据稀疏情况）的效果不好；三是当图足够大时，路径抽取工作比较费时。

3.2.3 基于分布式表示学习的推理

基于表示学习方法的中心思想是找到一种映射函数, 将符号表示映射到向量空间进行数值表示，从而减少维数灾难，同时捕捉实体和关系之间的隐式关联，重点是可以直接计算且计算速度快。

知识表示的一些背景知识：通常用三元组来表示知识。将一个三元组表示成（h,r,t），其中 h 表示头实体（head entity），r 表示关系（relation），而 t 表示尾实体（tail entity）。

我们可以用独热向量来表示这个知识。但实体和关系太多，维度太大。当两个实体或关系很近时，独热向量无法捕捉相似度。受 Wrod2Vec 模型的启发，我们想用分布表示来表示实体和关系。

常见的方法有：TransE (Translating Embedding) 系列算法、RESCAL、DistMul

等，可用于下游任务如节点分类、链接预测等。

- TransE（Translational Distance Model，转移距离模型）

基本思想：使 head 向量和 relation 向量尽可能靠近 tail 向量。通常用 L1 或 L2 范数来衡量它们的靠近程度。损失函数使用了负抽样的 max-margin 函数。TransE 只能处理一对一的关系，不适合一对多/多对一关系。

- TransH

基本思想：将关系解释为超平面上的转换操作。每个关系都有两个向量，超平面的范数向量 Wr 和超平面上的平移向量(dr)。

目标：处理一对多/多对一/多对多关系，并且不增加模式的复杂性和训练难度。

- TransR

TransR 在两个不同的空间（实体空间和多个关系空间）中建模实体和关系，并在对应的关系空间中进行转换。

综上所述，这些模型的基本思想均是利用 head 向量和 relation 向量的和来预测 tail 向量。但实体和关系可能更为复杂，寻找一种更有效地方法来表达知识具有一定的挑战性。

针对未考虑层级关系的不足，提出了实体的层级表示。TKRL(type-embodied knowledge representation learning)学习知识图谱实体和关系的表示，将层级类型信息用于映射矩阵、训练时负例的选择和评估时候选的过滤。

针对未考虑丰富语义信息的不足，SSE(semantically smooth embedding)学习知识图谱实体和关系的表示,利用实体语义类信息强制表示空间几何结构语义平滑。

TransE-NMM(TransE-neighborhood mixture modeling)，在 TransE 的基础上定义基于邻居的实体表示,引入邻居实体信息进行实体和关系的表示学习。

TEKE(text-enhanced knowledge embedding)引入文本语料中丰富的上下文信息扩展知识图谱的语义结构,学习知识图谱实体和关系的表示..

cross-KG 同时学习两个不同不同知识图谱的表示。通过映射语义相关的两个知识图谱中的实体和关系到统一的语义空间。借助更大和更稠密知识图谱的知识,促进稀疏知识图谱的表示学习，提升连接预测效果。

TransA 可用于解决最优损失函数的参数问题以及时间因素方面的不足。

新颖的时间感知知识图谱补全模型 TAE(time-aware embedding)用三元组和时间信息 预测知识图谱中的连接,即:给定三元组中的两个元素与时间区间,预测另一个元素。

自适应的鲁棒转移模型 puTransE(parallel universe TransE)进行知识图谱实体和关系的表示学习,puTransE 产生多个表示空间,每个表示空间对应一个采

样的关系和先后通过语义感知与结构感知选择机制得到的三元组集合。

3.2.4 基于张量/矩阵分解的表示推理

基于张量/矩阵分解的表示推理将(头实体,关系,尾实体)三元组看成张量/矩阵中的元素构建张量/矩阵,通过张量/矩阵分解方法进行表示学习.分解得到的向量表示相乘重构成张量/矩阵,元素值即为对应三元组有效与否的得分,可以认为得分大于特定阈值的三元组有效,或候选预测按照得分排序,选择得分高的候选作为推理结果。

RESCAL 是基于三阶张量进行表示学习，这种推理方式虽然准确率高，可解释性强，但内存占用量大，计算速度慢。针对此类问题，提出了 TRESCAL,在 RESCAL 的基础上引入实体类型信息这一关系域知识。

张量分解模型 ARE(additive relational effects)学习知识图谱三元组的隐性和观察到的模式,用一个附加项增广 RESCAL 模型(隐性模式),对应观察到的模式

RSTE(random semantic tensor ensemble)采用分治策略,从知识图谱中采样多个多样的更小规模子图张量,通过集成子图张量的 RESCAL 分解进行连接预测.RSTE 极大地降低了内存占用和运行时间,同时,通过增加子图张量的分解或使某些子图张量分解得到的结果无效,可以快速处理动态变化知识图谱的增或删操作。

3.2.5 基于空间分布的表示推理

基于空间分布的表示推理建立模型拟合知识图谱中实体和关系的空间分布特征,使得在向量表示空间中，实体和关系的空间分布尽可能地与原知识图谱一致。

高斯混合模型 TransG 从产生式的角度看待表示学习，可以学习出关系在对应三元组中最主要的隐形语义向量，指数作用拉大了其与关系的其他语义向量对三元组的贡献差距。

高斯模型 KG2E 基于密度的表示，直接建模实体和关系的确定性，在多维高斯分布的空间中学习知识图谱的表示，KG2E 可有效处理多种类型的关系。

ManifoldE(manifold-based embedding) 是基于流形的表示模型，扩展了三元组的位置从从向量空间中的一个点到一个流形结构，扩展了基于转移的头实体向量加关系向量等于尾实体向量到基于轨道的流形函数,取得了很好的连接预测效果,是目前除了神经网络方法以外最有效的方法。

3.2.6 基于神经网络的推理

利用神经网络直接建模知识图谱事实元组，得到事实元组元素的向量表示，用于进一步推理。

优点：推理能力和泛化能力较强；

缺点：相比基于分布式表示的推理，复杂度更高,可解释性更弱。

NTN(neural tensor network，神经张量网络)用双线性张量层代替传统的神经网络层，在不同的维度下，将头实体和尾实体联系起来，刻画实体间复杂的语义联系。

引入类似的神经张量网络模型预测知识图谱中新的关系,通过用从文本无监督学到的词向量初始化实体表示提升模型,甚至可以预测知识图谱中未出现实体的关系。

共享变量神经网络模型 ProjE，减少了大量参数，可处理大规模知识图谱。

目前存在的问题/难点：神经网络的可解释性问题；如何扩展其他领域中更多基于神经网络的方法到知识图谱领域中。

突破口：从一般图结构数据基于神经网络的方法迁移到知识图谱中。

3.2.7 混合推理

通过混合多种推理方法，充分利用不同方法的优势。

基于分布式表示推理的强计算能力、基于神经网络推理的强学习能力和泛化能力。目前混合推理包括混合规则与分布式表示推理混合神经网络与分布式表示推理。存在的问题：目前混合推理还停留在两种方法的浅层混合，即以一种方法为主，另一种方法为辅的推理，尚缺乏更深层次的混合模式以充分利用各方法的优势。

总体而言，（1）基于分布式表示的单步推理中细粒度建模空间分布的方法,由于充分考虑了知识图谱的空间分布特征,表达推理能力强.针对该类方法的研究还比较少,有待进一步细致挖掘知识图谱的空间分布特征,探索更多的建模方法。（2）基于神经网络的推理和混合推理是推理效果相对较好的方法，有待进一步的研究工作；（3）如何混合多种互补的方法，进一步提高推理能力，有待进一步研究。

四、知识图谱嵌入

知识图谱（KG）是由实体(节点)和关系(不同类型的边)组成的多关系图。每

条边都表示为形式(头实体、关系、尾实体)的三个部分，也称为事实，表示两个实体通过特定的关系连接在一起，例如(AlfredHitchcock, DirectorOf, Psycho)。虽然知识图谱（KG）在表示结构化数据方面很有效，但是这类三元组的底层符号特性通常使 KGs 很难操作。为了解决这个问题，提出了知识图谱嵌入，即将实体和关系转化为连续的向量空间，从而简化操作，同时保留 KG 原有的结构。那些实体和关系嵌入能进一步应用于各种任务中，如 **KG 补全**、**关系提取**、**实体分类**和**实体解析**。

知识图谱嵌入具体来说是将知识库中的知识表示为低维稠密的实体向量, 即 Embedding。知识图谱是由实体和关系组成，通常采用三元组的形式表示，【head(头实体)，relation(实体的关系)，tail(尾实体)】，简写为(h, r, t)。知识图谱嵌入任务就是学习 h，r，t 的分布式表示。

表 1 基本符号定义

符号	描述	符号	描述
G	知识图谱	S	事实集合
(h,r,t)	事实三元组	$(\mathbf{h},\mathbf{r},\mathbf{t})$	嵌入三元组
$r\in R, e\in E$	关系集合与实体集合	$f_r(h,t)$	评分函数
$\sigma(\cdot), g(\cdot)$	非线性激活函数	$\mathbf{M}_r$	映射矩阵
L	损失函数	¡ ^{d}	d 维实值空间
£ ^{d}	d 维复数空间	$\mathbb{H}^d$	d 维超复数空间
$\mathbb{T}^d$	d 维环面空间	$\mathbf{t}\otimes\mathbf{r}$	Hamilton product
$\mathbf{h}\circ\mathbf{t}$	Hadmard product	$R_e(\cdot)$	取复数值的实部
$\mathbf{h}\star\mathbf{t}$	循环相关	concat(),$[\mathbf{h},\mathbf{r}]$	向量/矩阵连接
ω	卷积滤波器	*	卷积操作
$[\mathbf{h}]_i$	向量 **h** 的第 i 项	$[\mathbf{M}_r]_{i,j}$	矩阵 $\mathbf{M}_r$ 的第 ij 项

2.1 融合事实信息

我们将这种嵌入技术大致分为两类：平移距离模型和语义匹配模型。前者使用**基于距离**的评分函数，后者使用**基于相似度**的评分函数。

2.1.1 平移距离的模型

平移距离模型利用了基于距离的评分函数, 通过两个实体之间的距离对事实的合理性进行度量。

（1）TransE 及其扩展

表示学习在自然语言处理领域受到广泛关注起源于 Mikolov 等人于 2013 年

提出的 word2vec 词表示学习模型和工具包。利用该模型，Mikolov 等人发现词向量空间存在平移不变现象。

受到该**平移不变现象**的启发，Border 等人提出了 TransE 模型，将知识库中的关系看作实体间的某种平移向量。对于每个事实三元组(h,r,t)，TransE 模型将实体和关系表示为同一空间中，把关系向量 r 看作为头实体向量 h 和尾实体向量 t 之间的平移即 h+r≈t。比如：对于给定的 2 个事实(姜文，导演，邪不压正)和(冯小刚，导演，芳华)，除了可以得到：姜文+ 导演≈邪不压正和冯小刚+导演≈芳华，还可以通过平移不变性得到：邪不压正 - 姜文 ≈芳华 – 冯小刚，即得到两个事实相同的关系（DirectorOf）的向量表示。我们也可以将 r,看作从 h 到 t，的翻译，因此 TransE 也被称为翻译模型，如图 1（a）所示，对于每一个三元组（h,r,t）TransE 希望：h+r≈t，评分函数在表 1 中所示。

虽然 TransE 模型的**参数较少，计算的复杂度显著降低，并且在大规模稀疏知识库上也同样具有较好的性能与可扩展性**。但是 TransE 模型**不能用在处理复杂关系上**。

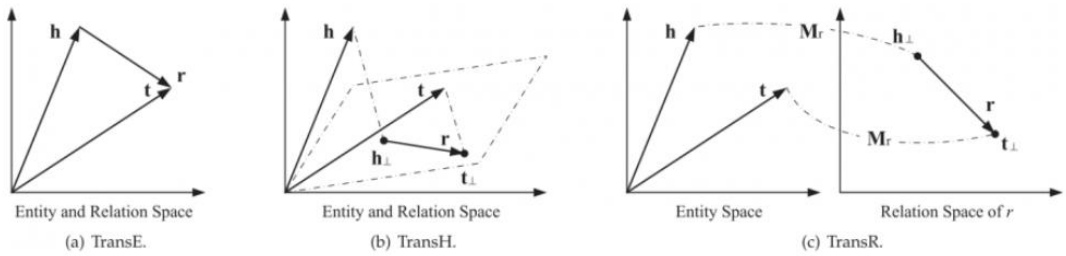


图 4-1 TransE，TransH 和 TransR 的简要说明

为了解决 TransE 模型在处理一对多 、 多对一 、多对多复杂关系时的局限性，TransH 模型**让一个实体在不同的关系下拥有不同的表示**。如图 1（b）所示，对于关系 r，TransH 模型同时使用平移向量 r 和超平面的法向量 w_r 来表示它。

虽然 TransH 模型使每个实体在不同关系下拥有了不同的表示，它仍然假设实体和关系处于相同的语义空间中，这一定程度上限制了 TransH 的表示能力。TransR 模型则认为，**一个实体是多种属性的综合体，不同关系关注实体的不同属性**。TransR 认为不同的关系拥有不同的语义空间。对每个三元组，首先应将实体投影到对应的关系空间中，然后再建立从头实体到尾实体的翻译关系。如图 4-1(c)所示是 TransR 模型的简单示例。

虽然 TransR 模型较 TransE 和 TransH 有显著改进，它仍然有很多缺点:

A. 在同一个关系:下，头、尾实体共享相同的投影矩阵。然而，一个关系的头、尾实体的类型或属性可能差异巨大.例如，对于三元组(美国，总统，奥巴马)，美

国和奥巴马的类型完全不同，一个是国家，一个是人物。

B. 从实体空间到关系空间的投影是实体和关系之间的交互过程，因此 TransR 让投影矩阵仅与关系有关是不合理的。

C. 与 TransE 和 TransH 相比，TransR 由于引入了空间投影，使得 TransR 模型参数急剧增加，计算复杂度大大提高。

为了解决这些问题，TransD 模型设置了 2 个分别将头实体和尾实体投影到关系空间的投影矩阵 Mr1 和 Mr2。

通过增加一个稀疏度参数向量解决异构多元关系的 TranSparse 是通过在投影矩阵上强化稀疏性来简化 TransR 的工作。它有两个版本:TranSparse (共享)和 TranSparse (单独)。前者对每个关系 r 使用相同的稀疏投影矩阵，后者对于头实体和尾实体分别使用 2 个不同的投影矩阵 M_r1(theta_r1)和 M_r2(theta_r2)。

除了允许实体在涉及不同关系时具有不同的嵌入之外，提高 TransE 模型性能可以从降低 h+r≈t 的要求研究开始。TransM 模型将为每个事实（h,r,t）分配特定的关系权重 theta_r，定义的评分函数如表 1 所示。**通过对一对多、多对一和多对多分配较小的权重**，TransM 模型使得 t 在上述的复杂关系中离 h+r 更远。ManifoldE 模型则是对于每个事实三元组（h,r,t）将 h+r≈t 转换为为（h+r-t）的 L2 范式约等于 theta_r 的平方。同样地，ManifoldE 把 t 近似地位于流形体上，即一个以 h+r 为中心半径为 theta_r 的超球体，而不是接近 h+r 的精确点。评分函数如表 1 所示。

TransA 模型为每个关系 r 引入一个对称的非负矩阵 Mr，并使用**自适应马氏距离**定义评分函数，评分函数如表 1 所示。通过学习距离度量 Mr, TransA 在处理复杂关系时更加灵活。

（2）高斯嵌入模型

知识库中的关系和实体的语义本身具有**不确定性**, 而过去模型中都忽略这个因素。因此，KG2E 使用高斯分布来表示实体和关系。其中高斯分布的均值表示的是实体或关系在语义空间中的中心位置, 而高斯分布的协方差则表示该实体或关系的不确定度。

通常一个关系会对应多种语义，因此 TransG 模型通过**考虑关系 r 的不同语义**（即每种语义用一个高斯分布来刻画），形成多个高斯分布，能够区分出正确和错误实体。

（3）其他距离模型

非结构化模型(UM)是 TransE 的一个简单版本，通过设置所有的 r=0，得到一个评分函数如表 1 所示。

显然，它不能区分不同的关系。结构嵌入(SE)通过使用两个独立的矩阵

M_r_1 和 M_r_2 为每个关系 r 对头尾实体进行投影, 得到的评分函数如表 1 所示。

2.1.2 语义匹配模型

语义匹配模型利用基于相似性的评分函数。它们通过匹配实体的潜在语义和向量空间表示中包含的关系来度量事实的可信性。

（1）RESCAL 模型及其扩展

RESCAL 模型(又称双线性模型)通过使用一个向量表示每个实体来获得它的潜在语义。每个关系都表示为一个矩阵，该矩阵对潜在因素之间的成对交互作用进行了建模。DistMult 模型通过将映射矩阵 Mr 限制为对角矩阵来简化 RESCAL，然而这种过度简化的模型只能处理对称的关系，这显然对于一般的 KGs 是不能完全适用的。HolE 模型将 RESCAL 的表达能力与 DistMult 的效率和简单性相结合。ComplEx 模型通过引入复值嵌入来扩展 DistMult，以便更好地对非对称关系进行建模。ANALOGY 模型扩展了 RESCAL，从而进一步对实体和关系的类比属性进行建模, 例如, AlfredHitchcock 之于 Psycho, 正如 JamesCameron 之于 Avatar。

（2）基于神经网络匹配

语义匹配能量模型 SME 采用神经网络结构进行语义匹配。给定一个事实三元组（h,r,t），它首先将实体和关系投影到输入层中的嵌入向量。然后，将关系 r 与头实体 h 组合得到 g_u(h,r)，并与尾实体 t 组合，得到隐藏层中的 g_v(t,r)。则该事实的分数最终由它们的点积定义为匹配的 g_u 和 g_v。

神经张量网络模型(NTN)是另外一种神经网络结构，给定一个事实，它首先将实体投影到输入层中的嵌入向量。然后，将这两个实体 h,t 由关系特有的张量 M_r(以及其他参数)组合，并映射到一个非线性隐藏层。最后，一个特定于关系的线性输出层给出了评分。尽管 NTN 是迄今为止最具表达能力的模型，但是由于它的每个关系需要 O(d^2*k)个参数，并且不能简单有效地处理大型的 KGs。

多层感知机 MLP 是一种更简单的方法, 在这种方法中, 每个关系(以及实体)都是由一个向量组合而成的。给定一个事实（h,r,t）将嵌入向量 h、r 和 t 连接在输入层中，并映射到非线性的隐藏层。然后由线性输出层生成分数。

神经关联模型 NAM 使用“深度”架构进行语义匹配，给定一个事实，它首先将头实体的嵌入向量和输入层中的关系连接起来，从而给出 z_0=[h,r]。然后输入 z_0 输入到一个由 L 个线性隐层组成的深神经网络中。

（3）其他方法

除上述模型外，还有其他学习头尾实体对的表示。具体地，给定一个三元组（h,r,t），关系 r 可以表示为一个向量 r，实体对（h,t）可以用另外一个向量 p

表示。该事实的合理性可以通过 r 和 p 内积进行度量。然后，通过最小化成对排序损失来学习这些向量表示，类似于在 Eq(2)中的定义。这种实体对表示特别适用于关系提取，其目的是确定一对实体之间可能存在的关系。同样地，头实体 h 可以表示为一个向量 h，实体对（r,t）可以用另外一个向量 p 表示。然而，这种方法也有其缺点。比如，如果头-尾实体对（h_1,t）和（h_2,t）通过不同的向量表示进行建模，则它们共享的相同的尾实体信息将会丢失。而且，也无法有效地发现未配对实体(如 h3 和 t)之间的关系。此外，它还导致了空间复杂度的增加，因为每个实体对都需要计算一次向量表示，它总共需要 O(n^2d+md)个参数。

2.2 融入附加信息

多源信息提供了知识图谱中三元组事实以外的信息,能够帮助构建更加精准的知识表示,仅使用事实进行 知识图谱嵌入的方法忽略了蕴含在多源信息中的丰富知识,例如:实体类别信息（语义平滑嵌入(SSE)、融合类型的知识表示学习(TKRL)）、文本描述信息（DKRL、TEKE）、关联路径、逻辑规则（KALE）以及其他信息（实体属性、时序信息、图结构）等，充分利用这些多源信息对于降低实体与关系之间的模糊程度，进而提高推理预测的准确度至关重要。

2.3 动态知识图谱嵌入

当前 KGE 的研究主要集中于静态知识图谱,其中事实不会随时间发生变化,例如:TransE,TransH,TransR, RESCAL 等等.但是,在实际应用中,知识图谱通常是动态的,例如 Twitter 中的社交知识图,DBLP 中的引文知识 图等,其中事实随时间演变,仅在特定时间段内有效.以往的静态 KGE 方法完全忽略了时间信息,这使得静态 KGE 方法无法在这些实际场景中工作.因此,有必要设计一种用于动态知识图谱嵌入的方法。表为 TDG2E 模型与其他动态 KGE 方法对比。

表 TDG2E 模型与其他动态 KGE 方法对比

比较方法	存在的问题	TDG2E 优势	TDG2E 采用的解决办法
HyTE, Flexible Translation	独立学习不同子 KG,不能显式建模动态 KG 演化过程	同时保留当前子 KG 的结构信息与动态 KG 的时间演化模式	利用基于GRU的模型捕获动态KG中相邻子 KG 的依赖关系;引入辅助损失,利用先前的结构信息监督后续超平面的学习过程
HyTE, Flexible Translation, t-TransE	没有完全解决 KG 中时间戳分布不平衡的问题	解决了动态 KG 面临的时间不平衡问题	在 GRU 中设计时间间隔门,引入相邻子 KG 之间的时间间隔

学习资料:

技术动态 | 「知识图谱嵌入技术研究」最新 2022 综述_开放知识图谱的博客-CSDN 博客

知识图谱嵌入(KGE)：方法和应用的综述_人工智能学家的博客-CSDN 博客

[金山文档] 知识图谱: 方法、实践与应用 (王昊奋　漆桂林　等) (Z-Library).epub

[金山文档] 知识图谱：概念与技术 (肖仰华 等) (Z-Library).pdf

[金山文档] 知识图谱标准化白皮书 (中国电子技术标准化研究院) (Z-Library).pdf

参考文献:

[1] Wang, Z. Mao, B. Wang and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724-2743, 1 Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

[2] 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(02):247-261.

[3] GUO Shu,WANG Quan,WANG Bin,et al.Semantically Smooth Knowledge Graph Embedding.Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,2015:84-94.

[4] LIN Yankai,LIU Zhiyuan,LUAN Huanbo,et al.Modeling Relation Paths for Representation Learning of Knowledge Bases.EMNLP,2015:205-714.

[5] WANG Zhen,ZHANG Jianwen,FENG Jianlin,et al.Knowledge Graph and Text Jointly Embedding.EMNLP,2014:1591-1601.

[6] GUO Shu,WANG Quan,WANG Lihong,et al.Knowledge Graph Embedding with Iterative Guidance from Soft Rules.Thirty-Second AAAI Conference on Artificial Intelligence,2018.