来源：**首次公开！深度学习在知识图谱构建中的应用-阿里技术**

DeepDive系统在数据处理阶段很大程度上依赖于NLP工具，如果NLP的过程中存在错误，这些错误将会在后续的标注和学习步骤中被不断传播放大，影响最终的关系抽取效果。为了避免这种传播和影响，近年来深度学习技术开始越来越多地在关系抽取任务中得到重视和应用。本章主要介绍一种远程监督标注与基于卷积[神经网络](http://www.elecfans.com/tags/%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/)的模型相结合的关系抽取方法以及该方法的一些改进技术。

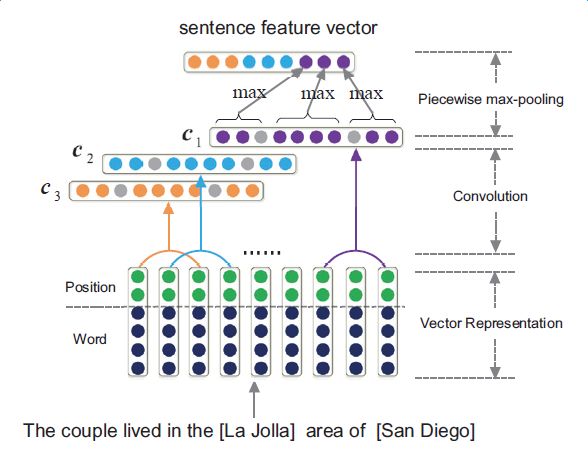
[PI](http://www.elecfans.com/tags/pi/)ecewise Convolu[ti](http://www.elecfans.com/tags/%E5%BE%B7%E5%B7%9E%E4%BB%AA%E5%99%A8/)onal Neural Networks（PCNNs）模型

PCNNs模型由Zeng et al.于2015提出，主要针对两个问题提出解决方案：

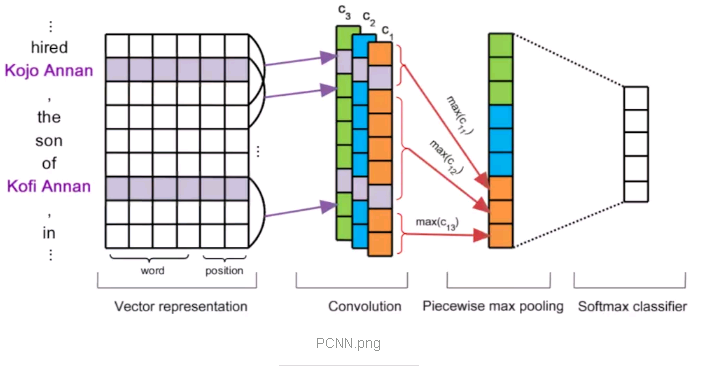
针对远程监督的wrong label problem，该模型提出采用多示例学习的方式从训练集中抽取取置信度高的训练样例训练模型。

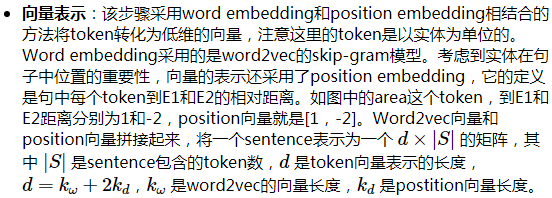
针对传统统计模型特征抽取过程中出现的错误和后续的错误传播问题，该模型提出用 piecewise 的卷积神经网络自动学习特征，从而避免了复杂的NLP过程。

下图是PCNNs的模型示意图：

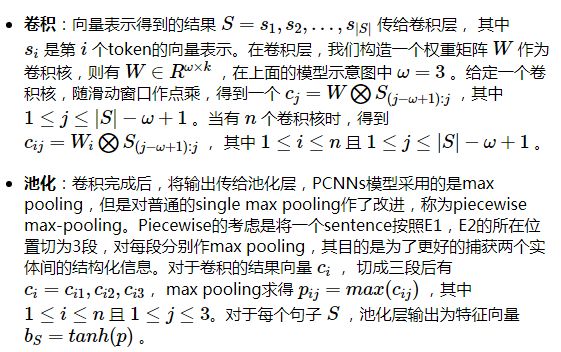
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXnyAUuZuAACiExP9bn8052.png)

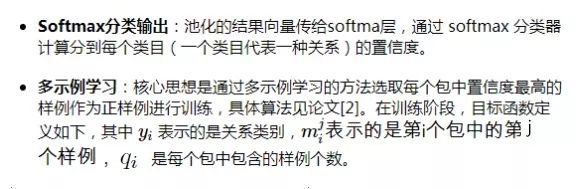
PCNNs模型主要包括以下几个步骤：

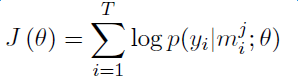


[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXnyAYpTjAACDvXf_WD4664.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXnyAWGTMAAA4WylOQYk424.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2ATziLAAEQ1-M4YLo747.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2ACP6eAABsJ4BY4oA878.jpg)

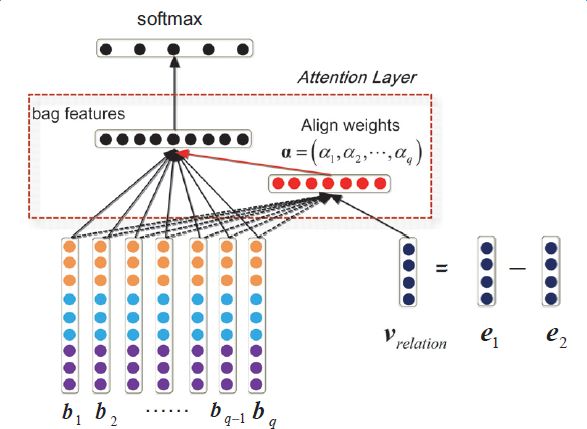
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2AQ3DuAAAfHjAGlbU958.png)

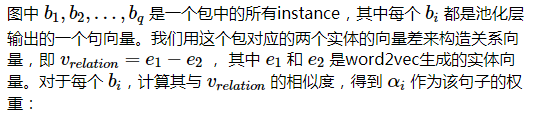
[http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2ANgwOAAAfZDSQNRU846.png](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2ANgwOAAAfZDSQNRU846.png)

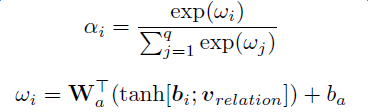
实验证明，PCNNs + 多实例学习的方法 Top N 上平均值比单纯使用多示例学习的方法高了 5 个百分点。

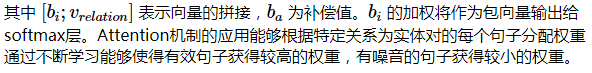
At[te](http://www.elecfans.com/tags/te/)ntion机制和其它改进

上述模型对于每个实体对只选用一个句子进行学习和预测，损失了大量的来自其它正确标注句子的信息。为了在滤除wrong label case的同时，能更有效地利用尽量多的正确标注的句子，Lin et al. 于2016年提出了PCNNs+Attention（APCNNs）算法。相比之前的PCNNs模型，该算法在池化层之后，softmax层之前加入了一种基于句子级别的attention机制，算法的示意图如下：

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn2AFIOwAACUTG56N6U325.png)

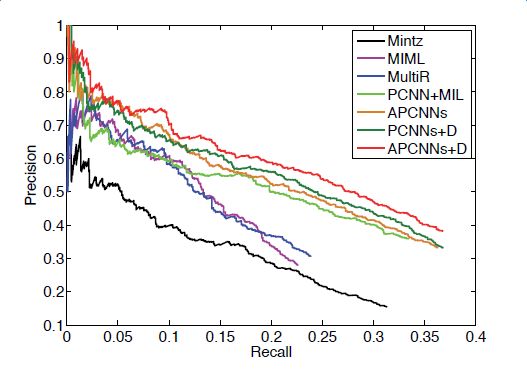
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn6AEMOEAABG_-0uCtc709.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn6ADvEfAAAx5SI_D3Y466.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn6Aea_7AAA4iYSKTfI587.png)

除了Attention机制外，还有一些其它的辅助信息也被加入多示例学习模型来改关系抽取的质量，例如在计算实体向量的时候加入实体的描述信息（Ji et al.，2017）；利用外部神经网络获取数据的可靠性和采样的置信度等信息对模型的训练进行指导（Tang et al.，2017）。

下图显示了各模型和改进算法的准确率和召回率的对比，其中[MI](http://www.hqpcb.com/quote/)ntz不对远程监督的wrong label problem做处理，直接用所有标注样例进行训练；MultiR和MIML是采用概率图模型进行示例筛选的两种多示例学习模型；PCNN+MIL是本章第一小节介绍的模型；APCNNs 在PCNN+MIL基础上添加了attention机制；PCNNs+D在PCNN+MIL基础上添加了对描述信息的使用；APCNNs+D在APCNNs基础上添加了对描述信息的使用。实验采用的是该领域评测中使用较广泛的New York Times（NYT）数据集（Riedel et al.，2010）。

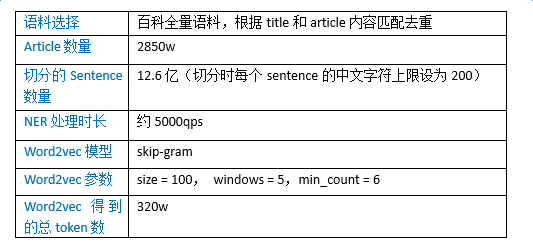
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn6AVFl6AAB2kMpazBU481.png)

深度学习方法在图谱构建中的应用进展

深度学习模型在神马知识图谱数据构建中的应用目前还处于探索阶段，本章将介绍当前的工作进展和业务落地过程中遇到的一些问题。

语料准备和实体向量化

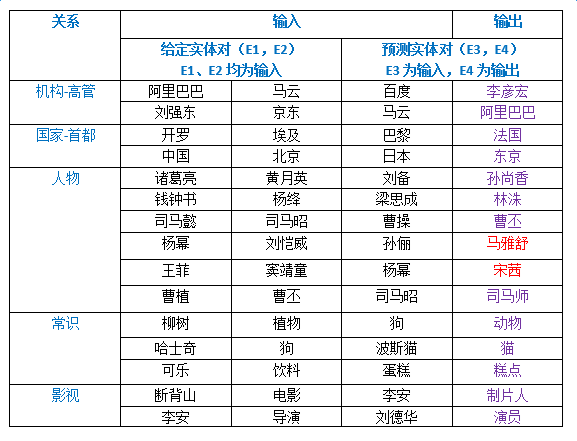
深度学习模型较大程度依赖于token向量化的准确性。与基于DeepDive方法的语料准备相同，这里的token切分由以词为单位，改为以实体为单位，以NER环节识别的实体粒度为准。Word2vec生成的向量表征token的能力与语料的全面性和语料的规模都很相关，因此我们选择百科全量语料作为word2vec的训练语料，各统计数据和模型参数设置如下表所示：

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn6ANImRAAAXyjEEHz4441.png)

为了验证词向量训练的效果，我们对word2vec的结果做了多种[测试](http://www.hqpcb.com/quote/)，这里给出部分实验数据。下图所示的是给定一个实体，查找最相关实体的实验：

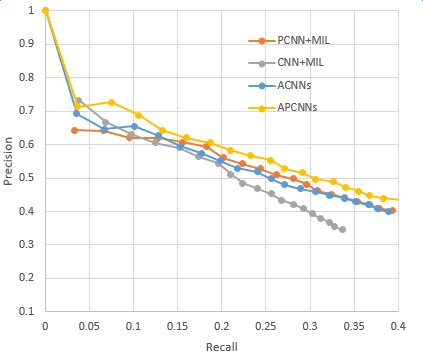
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn-AQo1nAAD6n1M9aQU505.png)

以下是给定一个实体对和预测实体对的其中一个实体，计算预测实体对中另一个实体的实验。随机选取了五种预测关系，构造了15组给定实体对和预测实体对，预测结果如下图所示，除了飘红的两个例子，其余预测均正确：

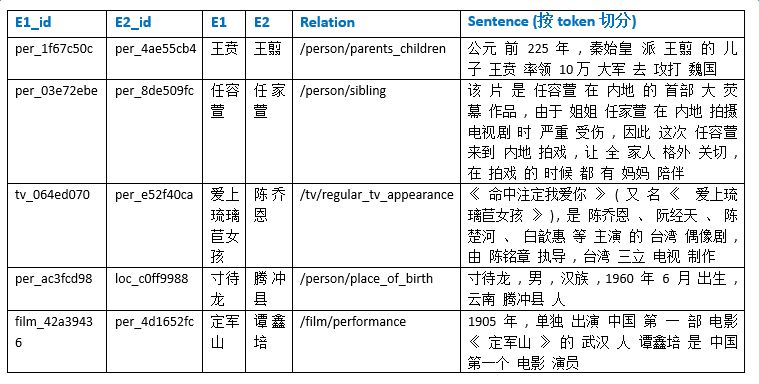
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn-AM0eiAAAqOwuMk2s409.png)

模型选取与训练数据准备

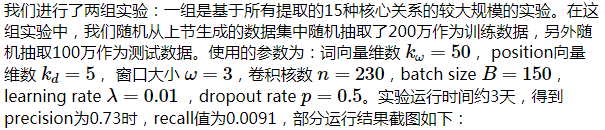
具体应用中我们选择采用APCNNs模型。我们在NYT标准数据集上复现了上一章提到的几种关键模型，包括CNN+MIL，PCNN+MIL，CNNs（基于Attention机制的CNN模型）和APCNNs。复现结果与论文中给出的baseline基本一致，APCNNs模型的表现明显优于其它模型。下图是几种模型的准召结果对比：

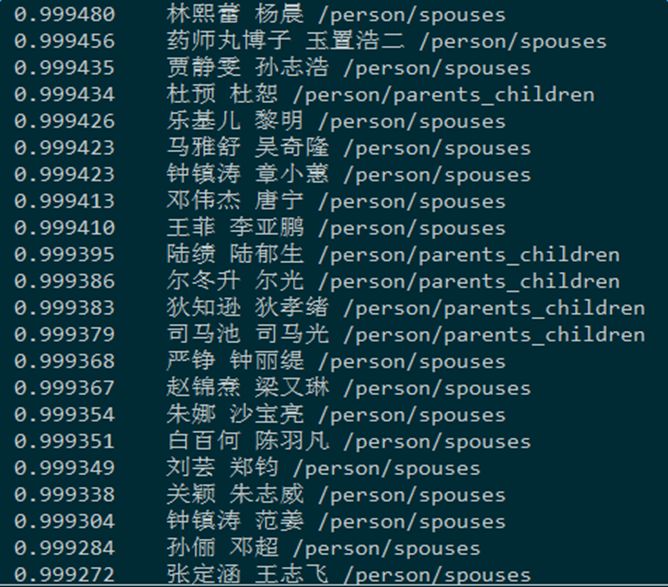
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn-ANdxQAABjCW30ra0285.png)

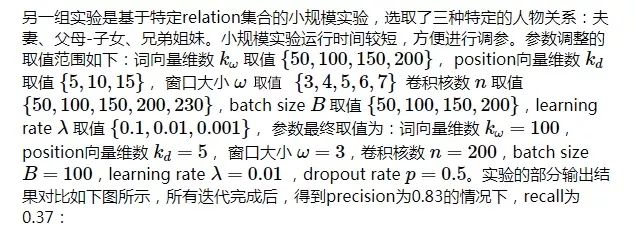
为了得到丰富的训练数据，我们取知识图谱中建设相对完善的人物、地理位置、组织机构、电影、电视、图书等领域下的15个核心关系，如电影演员、图书作者、公司高管、人物出生地等，对照百科全量语料，产出relation值为15个关系之一的标注正例，合计数目在千万量级，产出无relation值标注（relation值为NA）的示例超过1亿。

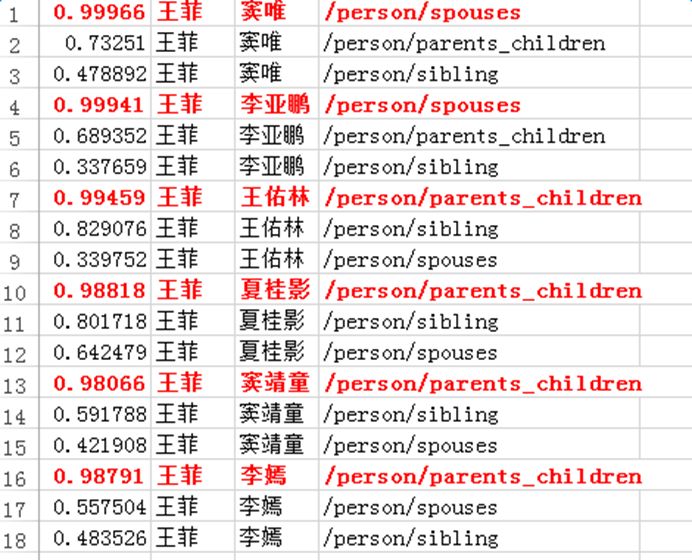
[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXn-AGzWIAAE1Kw8Mq24674.png)

应用尝试和问题分析

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXoCAKKD5AABonsKpqwg730.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXoCAcsDMAAFnUlhi8vE394.png)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXoCAViunAADQrjiwWLk181.jpg)

[](http://file.elecfans.com/web1/M00/4C/22/o4YBAFqzXoCAKvF0AAGKyXrIDKs813.png)

APCNNs模型在辅助知识图谱数据构建中目前还处于尝试阶段。就运算能力而言，APCNNs模型相比DeepDive系统更有优势，能在大规模语料上同时针对多个关系进行计算，且迭代更新过程无需人工校验交互。但在业务落地过程中，我们也遇到了一些问题，总结如下：

大规模实验耗时过长，给参数的调整和每一次算法策略上的迭代增加了难度

目前学术界通用的测试语料是英文的NYT数据集，相同的模型应用于中文语料时，存在准召率对标困难的问题

深度学习的过程人工难以干预。假设我们要预测（杨幂，刘恺威）的婚姻关系，但从最初的基于大规模语料的词向量生成开始，如果该语料中（杨幂，刘恺威）共现时的主导关系就不是婚姻关系，而是影视剧中的合作关系（如“该片讲述杨幂饰演的夏晚晴在遭遇好友算计、男友婚变的窘境下，被刘恺威饰演的花花公子乔津帆解救，但却由此陷入更大圈套的故事。”），或基于某些活动的共同出席关系（如“杨幂与刘恺威共同担任了新浪厦门爱心图书馆的公益大使”），则在attention步骤中得到的关系向量就会偏向合作关系，这将导致计算包中每个句子的权值时，表达婚姻关系的句子难以获得高分，从而导致后续学习中的偏差。

深度学习模型的结果较难进行人工评测，尤其对于知识图谱中没有出现的实体对，需要在大规模的中间过程矩阵中进行匹配和提取，将权重矩阵可视化为包中每个句子的得分，对计算资源和人工都有不小的消耗。

总结与展望

基于DeepDive的方法和基于深度学习的方法各有优势和缺陷，以下从4个方面对这两种方法进行总结和对比：

1、 语料的选取和范围

Deepdive可适用于较小型、比较专门的语料，例如历史人物的关系挖掘；可以针对语料和抽取关系的特点进行调整规则，如婚姻关系的一对一或一对多，如偏文言文的语料的用语习惯等。

APCNNs模型适用于大规模语料，因为attention机制能正常运行的前提是word2vec学习到的实体向量比较丰富全面。

2、 关系抽取

Deepdive仅适用于单一关系的判断，分类结果为实体对间某一关系成立的期望值。针对不同的关系，可以运营不同的规则，通过基于规则的标注能较好地提升训练集的标注准确率。

APCNNs模型适用于多分类问题，分类结果为relation集合中的关系得分排序。无需针对relation集合中特定的某个关系做规则运营。

3、 长尾数据

Deepdive更适用于长尾数据的关系挖掘，只要是NER能识别出的实体对，即使出现频率很低，也能根据该实体对的上下文特征做出判断。

APCNNs模型需要保证实体在语料中出现的次数高于一定的阈值，如min\_count>=5，才能保证该实体有word2vec的向量表示。bag中有一定数量的sentence，便于选取相似度高的用于训练

4、 结果生成与检测

Deepdive对输出结果正误的判断仅针对单个句子，同样的实体对出现在不同的句子中可能给出完全不同的预测结果。测试需要结合原句判断结果是否准确，好处是有原句作为依据，方便进行人工验证。

APCNNs模型针对特定的实体对做判断，对于给定的实体对，系统给出一致的输出结果。对于新数据的结果正确性判断，需要结合中间结果，对包中被选取的句子集合进行提取和验证，增加了人工[检验](http://www.hqpcb.com/zhuoluye9)有的难度。

在未来的工作中，对于基于DeepDive的方法，我们在扩大抓取关系数目的同时，考虑将业务实践中沉淀的改进算法流程化、平台化，同时构建辅助的信息增补工具，帮助减轻DeepDive生成结果写入知识图谱过程中的人工检验工作，例如，对于婚姻关系的实体对，我们可以从图谱获取人物的性别、出生年月等信息，来辅助关系的正误判断。

对于基于深度学习的方法，我们将投入更多的时间和精力，尝试从以下几方面促进业务的落地和模型的改进：

将已被DeepDive证明有效的某些改进算法应用到深度学习方法中，例如根据关系相关的关键词进行过滤，缩小数据规模，提高运行效率。

将计算中间结果可视化，分析attention过程中关系向量与sentence选取的关联，尝试建立选取结果好坏的评判机制，尝试利用更丰富的信息获得更准确的关系向量。

考虑如何突破预先设定的关系集合的限制，面向开放领域进行关系抽取，自动发现新的关系和知识。

探索除了文本以外其它形式数据的关系抽取，如表格、音频、图像等。