浙江大学

浙大校史人物知识图谱

课程名称：知识图谱导论

题　目：浙大校史人物知识图谱

学院：计算机科学与技术学院

姓名：宋子杰　学号：22121208

姓名：周宇轩　学号：22160159

目录

浙大校史人物知识图谱 1

1 项目简介 3

1.1 浙大校史人物的关系抽取 3

1.2 人物关系的可视化 4

2 关系抽取 4

2.1 命名实体识别 4

2.1.1 简介 4

2.1.2 命名实体识别的方法介绍 6

2.2 关系抽取 6

2.3 BERT 7

2.4 DEEPKE 8

2.5 浙大校史人物关系抽取 8

3 图数据可视化 9

4 小结 12

1 项目简介

该项目构建了一个关于浙大校史人物的可视化知识图谱，将各种实体关系进行处理，用形象的方式呈现出来，使同学或参观者更容易理解浙大校史人物中的关系，加深对浙江大学校史的认识，助力宣传浙江大学，为进一步挖掘、分析、构建相关知识及它们之间的相互联系提供了便利。

主要工作分为两方面：浙江大学校史人物的关系抽取和人物关系的可视化。

1.1 浙大校史人物的关系抽取

该项目使用基于 bert 的 OpenKE 工具从浙大校史中提取到了大量人物三元组信息。文本信息来源于《名流浙大（百年求是）》、《图说浙大》、《浙江大学图史》。具体流程如下：

首先使用了 OpenKE 的 NER 功能，即命名实体识别功能，从非结构化文本中提取得到大量实体信息。例如，对于文本

“作为乡人和被器重的留洋科学家，竺可桢很快被任命为浙江大学的校长，因为由他来出任浙大校长，地缘、人缘等都一应俱备。”

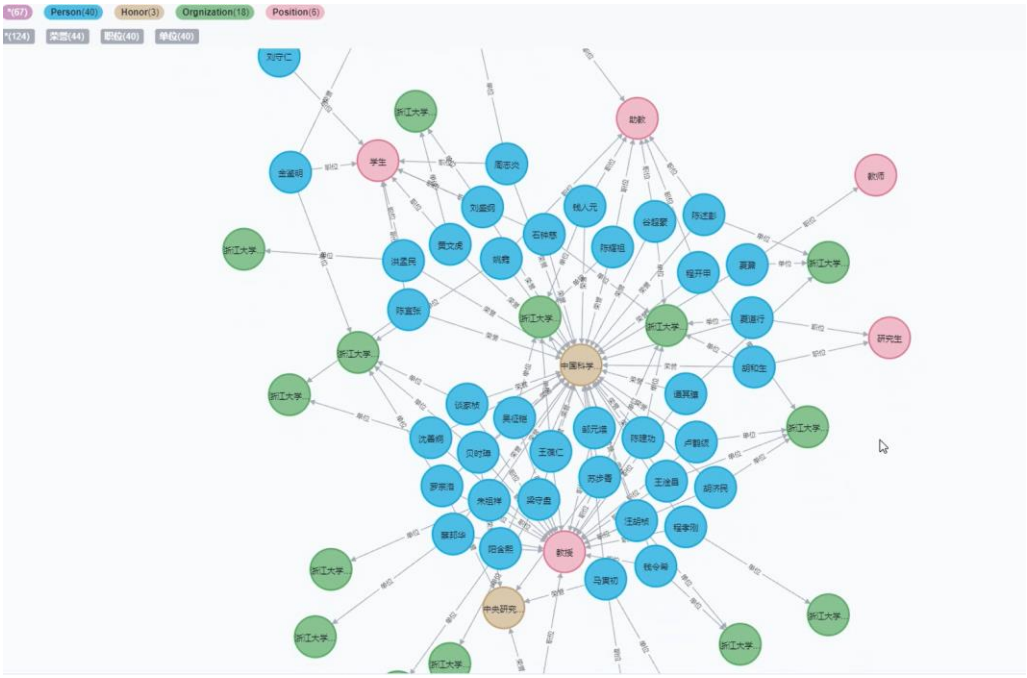
我们可以在这句话中抽取到人名“竺可桢”、机构名“浙江大学”、职位“校长”“科学家”等。接着对所有文本中的识别到的实体进行了记录。

然后使用了 OpenKE 的 RE 功能，即关系抽取功能，从非结构化文本中抽取实体间的关系。我们已经知道有这些实体：人名“竺可桢”、机构名“浙江大学”。那么它们之间的关系是怎样的呢？

我们将每条句子中出现的实体两两配对，进行关系预测，如模型认为“竺可桢”和“浙江大学”是雇员和单位的关系，“竺可桢”和“校长”是职位的关系，同时模型还给出了该预测的可信度，我们将抽取到的关系对按可信度排序，取高于 0.9 的部分，对于结果进行了部分人工辨别，得到了基于浙大校史人物的关系三元组。

1.2 人物关系的可视化

该项目采用 Neo4j 实现人物关系的可视化。Neo4j 是一个高性能的 NOSQL 图形数据库，图形数据库也就意味着它的数据并非保存在表或集合中，而是保存为节点以及节点之间的关系，其存储的数据由节点、边及属性组成。在该项目中，将抽取得到的三元组进行整理后导入 Neo4j 中进行可视化，可视化图中的各个节点分别表示人名、机构名称、职位名称以及荣誉名称；边则代表了人名节点与机构、职位、荣誉节点之间的关系。



2 关系抽取

2.1 命名实体识别

2.1.1 简介

NER 是一个非常基础，但是非常重要的任务。 简单的理解，实体，可以认为是某一个概念的实例。例如，“人名”是一种概念，或者说实体类型，那么“蔡

英文”就是一种“人名”实体了。“时间”是一种实体类型，那么“中秋节”就是一种“时间”实体了。所谓实体识别，就是将你想要获取到的实体类型，从一句话里面挑出来的过程。

小明 在 北京大学 的 燕园 看了

PER ORG LOC

中国男篮 的一场比赛

ORG

如上面的例子所示，句子“小明在北京大学的燕园看了中国男篮 的一场比赛”，通过 NER 模型，将“小明 ”以 PER，“北京大学”以 ORG，“燕园”以 LOC，“中国男篮”以 ORG 为类别分别挑了出来。

NER 是一种序列标注问题，因此他们的数据标注方式也遵照序列标注问题的方式，主要是 BIO 和 BIOES 两种。这里直接介绍 BIOES，明白了 BIOES，BIO 也就掌握了。

先列出来 BIOES 分别代表什么意思：

B，即 Begin，表示开始

I，即 Intermediate，表示中间

E，即 End，表示结尾

S，即 Single，表示单个字符

O，即 Other，表示其他，用于标记无关字符

将“小明在北京大学的燕园看了中国男篮的一场比赛”这句话，进行标注，结果就是：

[B-PER，E-PER，O， B-ORG，I-ORG，I-ORG，E-ORG，O，B-LOC，E-LOC，O，O，B-ORG，I-ORG，I-ORG，E-ORG，O，O，O，O]

那么，换句话说，NER 的过程，就是根据输入的句子，预测出其标注序列的过程。

2.1.2 命名实体识别的方法介绍

1）HMM 和 CRF 等机器学习算法

HMM 和 CRF 很适合用来做序列标注问题，早期很多的效果较好的成果，都是出自这两个模型。两种模型在序列标注问题中应用，我们在之前的文章中有介绍，感兴趣的同学可以看下如下链接的文章：

2）LSTM+CRF

目前做 NER 比较主流的方法就是采用 LSTM 作为特征抽取器，再接一个 CRF 层来作为输出层，后面我们用专门的文章来介绍这个模型。如下图所示：

3）CNN+CRF

CNN 虽然在长序列的特征提取上有弱势，但是 CNN 模型可有并行能力，有运算速度快的优势。膨胀卷积的引入，使得 CNN 在 NER 任务中，能够兼顾运算速度和长序列的特征提取，后面我们用专门的文章来介绍这个模型。

4）BERT+（LSTM）+CRF

BERT 中蕴含了大量的通用知识，利用预训练好的 BERT 模型，再用少量的标注数据进行 FINETUNE 是一种快速的获得效果不错的 NER 的方法，后面我们用专门的文章来介绍这个模型。

2.2 关系抽取

信息抽取在自然语言处理中是一个很重要的工作，特别在当今信息爆炸的背景下，显得格外的生重要。从海量的非结构外的文本中抽取出有用的信息，并结构化成下游工作可用的格式，这是信息抽取的存在意义。信息抽取又可分为实体抽取或称命名实体识别，关系抽取以及事件抽取等。命名实体对应真实世界的实体，一般表现为一个词或一个短语，比如曹操，阿里巴巴，中国，仙人掌等等。关系则刻画两个或多个命名实体的关系。比如马致远是《天净沙 • 秋思》的作者，那么马致远与《天净沙 • 秋思》的关系即是“创作”（author_of ）关系，张三是银行员工，那么张三与银行可以是“所属”(member_of)关系。

关系抽取可分为全局关系抽取与提及关系抽取。全局关系抽取基于一个很大的语料库，抽取其中所有关系对，而提及关系抽取，则是判断一句话中，一个实

体对是否存在关系，存在哪种关系的工作。

关系抽取分两步，一步是判断一个实体对是否有关第，而另一步则是判断一个有关系的实体对之间的关系属于哪种。当然这两步可变成一步，即把无关系当作关系的一种（特殊的），来进行多类别分类。

2.3 Bert

BER 的全称为 Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 masked language model（MLM），以致能生成深度的双向语言表征。BERT 论文发表时提及在 11 个 NLP（Natural Language Processing，自然语言处理）任务中获得了新的 state-of-the-art 的结果，令人目瞪口呆。

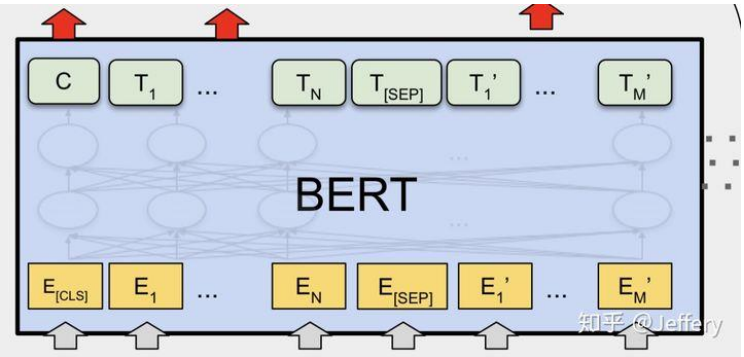
该模型有以下主要优点：

1）采用 MLM 对双向的 Transformers 进行预训练，以生成深层的双向语言表征。

2）预训练后，只需要添加一个额外的输出层进行 fine-tune，就可以在各种各样的下游任务中取得 state-of-the-art 的表现。在这过程中并不需要对 BERT 进行任务特定的结构修改。

以往的预训练模型的结构会受到单向语言模型（从左到右或者从右到左）的限制，因而也限制了模型的表征能力，使其只能获取单方向的上下文信息。而 BERT 利用 MLM 进行预训练并且采用深层的双向 Transformer 组件（单向的 Transformer 一般被称为 Transformer decoder，其每一个 token（符号）只会 attend 到目前往左的 token。而双向的 Transformer 则被称为 Transformer encoder，其每一个 token 会 attend 到所有的 token。）来构建整个模型，因此最终生成能融合左右上下文信息的深层双向语言表征。关于 Transformer 的详细解释可以参见 Attention Is All You Need 或者 The Illustrated Transformer。

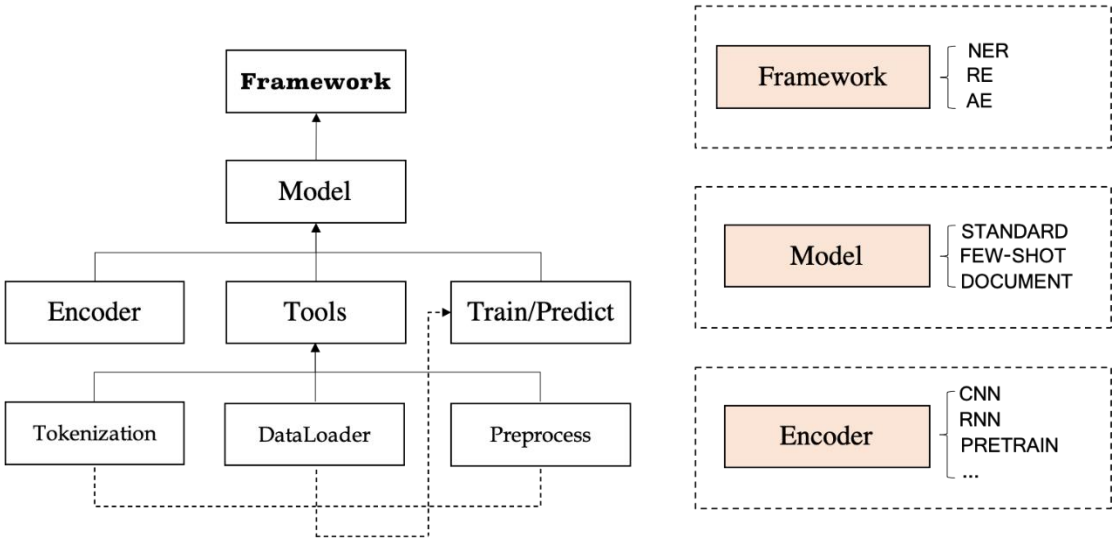
BERT 的主体结构



2.4 DeepKE

DeepKE 是一个基于深度学习的开源中文知识图谱抽取框架，是一个支持低资源、长篇章的知识抽取工具，可以基于 PyTorch 实现命名实体识别、关系抽取和属性抽取功能。

Deepke 的架构图如下所示



DeepKE 包括了三个模块，可以进行命名实体识别、关系抽取以及属性抽取任务，在各个模块下包括各自的子模块。其中关系抽取模块就有常规模块、文档级抽取模块以及低资源少样本模块。在每一个子模块中，包含实现分词、预处理等功能的一个工具集合，以及编码、训练和预测部分。

2.5 浙大校史人物关系抽取

该项目使用基于 bert 的 OpenKE 工具从浙大校史中提取到了大量人物三元

组信息。具体流程如下：

首先使用了 OpenKE 的 NER 功能，即命名实体识别功能，从非结构化文本中提取得到大量实体信息。例如，对于文本

“作为乡人和被器重的留洋科学家，竺可桢很快被任命为浙江大学的校长，因为由他来出任浙大校长，地缘、人缘等都一应俱备。”

我们可以在这句话中抽取到人名“竺可桢”、机构名“浙江大学”、职位“校长”“科学家”等。接着对所有文本中的识别到的实体进行了记录。

然后使用了 OpenKE 的 RE 功能，即关系抽取功能，从非结构化文本中抽取实体间的关系。我们已经知道有这些实体：人名“竺可桢”、机构名“浙江大学”。那么它们之间的关系是怎样的呢？

我们将每条句子中出现的实体两两配对，进行关系预测，如模型认为“竺可桢”和“浙江大学”是雇员和单位的关系，“竺可桢”和“校长”是职位的关系，同时模型还给出了该预测的可信度，我们将抽取到的关系对按可信度排序，取高于 0.9 的部分，对于结果进行了部分人工辨别，得到了基于浙大校史人物的关系三元组。

3 图数据可视化及知识查询

3.1 三元组数据整理

通过上述关系抽取得到的三元组形式如下：

```
苏步青 荣誉 中央研究院院士
苏步青 荣誉 中国科学院院士
苏步青 单位 浙江大学数学系
苏步青 职位 教授
陈建功 荣誉 中国科学院院士
陈建功 单位 浙江大学数学系
陈建功 职位 教授
```

为了更好的进行可视化，将上述三元组进行合并为：

```
苏步青 荣誉 中央研究院院士 单位 浙江大学数学系 职位 教授
苏步青 荣誉 中国科学院院士 单位 浙江大学数学系 职位 教授
陈建功 荣誉 中国科学院院士 单位 浙江大学数学系 职位 教授
```

3.2 数据导入 Neo4j 图数据库

• 安装 Neo4j 的 python 依赖包 py2neo，并在 python 导入该库。输入用户名和密码链接图数据库：

```
from py2neo import Graph
graph = Graph("http://localhost:7474",auth=("neo4j","613911z"))
```

• 创建节点 Node 和关系 Relationship：

```
from py2neo import Graph, Node, Relationship
from config import graph

with open('./data/spo-c.txt','r',encoding='utf-8') as f:
    for line in f.readlines():
        spo_array=line.strip("\n").split(" ")
        print(spo_array)
        start_node = Node('Person',name=spo_array[0])
        end_node1 = Node('Honor',hon=spo_array[2])
        end_node2 = Node('Orgnization',org=spo_array[4])
        end_node3 = Node('Position',pos=spo_array[6])

        relation1 = Relationship(start_node,spo_array[1],end_node1)
        relation2 = Relationship(start_node,spo_array[3],end_node2)
        relation3 = Relationship(start_node,spo_array[5],end_node3)
        graph.merge(start_node,'Person','name')
        graph.merge(end_node1,'Honor','hon')
        graph.merge(end_node2,'Orgnization','org')
        graph.merge(end_node3,'Position','pos')
        graph.merge(relation1,'Person','name')
        graph.merge(relation2,'Person','name')
        graph.merge(relation3,'Person','name')
```

3.3 知识图谱可视化及知识查询

• 通过 Neo4j 本地网页端查看知识图谱可视化结果：

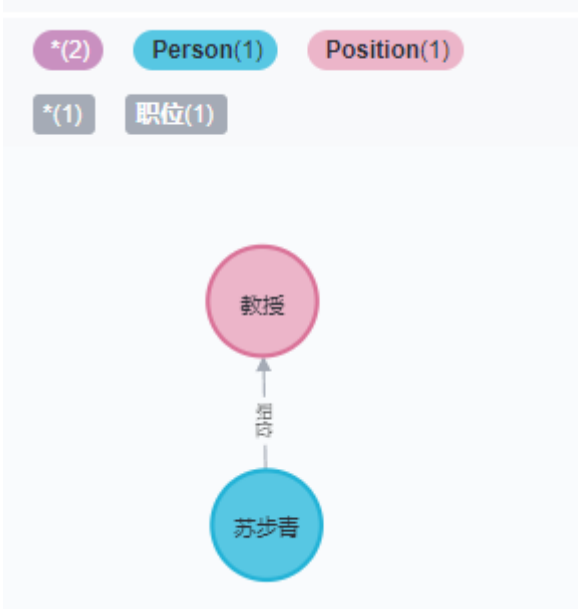
• 知识查询

通过 Cypher 语言，我们可以实现简单的知识查询，例如：

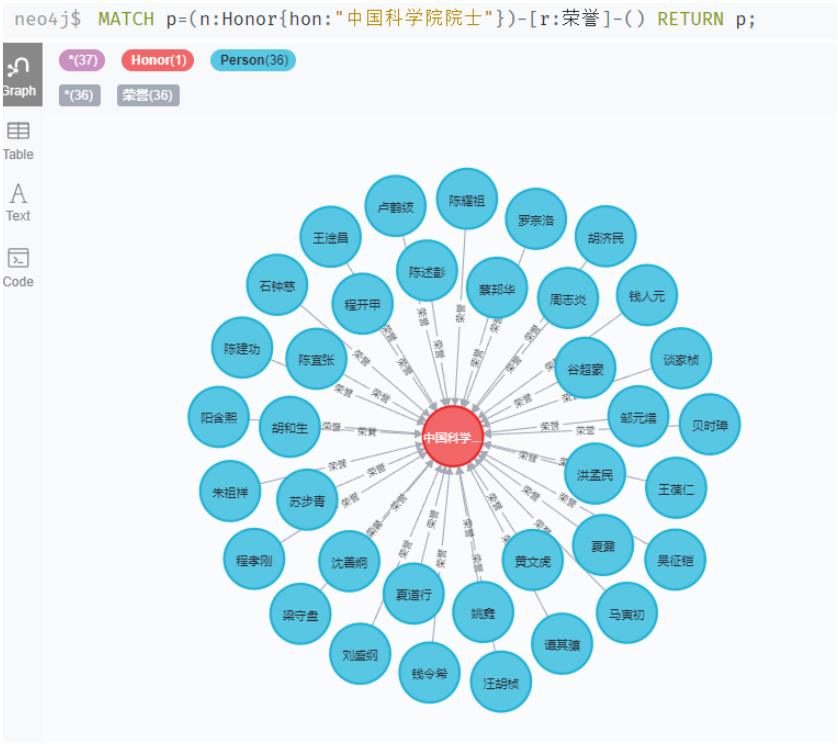
查询苏步青的职位，可以输入查询语句：

```
match p=(n:Person{name:'苏步青'})-[r:`职位`]-() return p;
```

获得结果：



查询所有获得中国科学院院士荣誉的校友，可以输入查询语句：
match p=(n:Honor{hon:"中国科学院院士"})-[r:荣誉]-() return p;
获得结果：



4 小结

该项目构建了一个关于浙大校史人物的可视化知识图谱。具体来说，该项目使用了图形学、信息可视化技术理论并利用可视化的图谱形象地展示浙大校史人物关系的结构。它把复杂的浙江大学校史人物关系通过数据挖掘、信息处理、和图形绘制而显示出来，同时人物关系被储存于图数据库中，利用图数据库的特点可以做一些知识查询，为浙江大学校史研究提供了切实的、有价值的参考，让同学和参观者对浙江大学校史有一个更清晰、更直接的认识。

基于该系统我们还可以做很多的拓展：

1. 为该系统提供一个问答系统，同学们只需搜索问题即可得到浙大校史人物的答案。问答系统将包括三个方面：问题理解与分析、信息检索及答案返回。在答案返回时还可以套上各种类型问题的回答模板，以免直接

返回答案过于晦涩。

2. 为该系统添加推理功能，可以自动推理人物关系之间的联系，得到我们尚不知道或是尚未重视的知识。

当然该系统也有一些不足：

1. 抽取到的人物关系信息还不够丰富，这与模型训练信息不足有一定的关系。
2. 使用文本的限制较大，对于类似于百度百科的文本效果较好，而对于复杂表达的语句效果欠佳，需要部分人工辨别。
3. 做知识问答系统时首先需要对用户输入的问题文本进行分词、词性标注及返回实体，并将实体返回到图数据库中搜索答案。一些工具或模型在对短词语的分词结果较好，而对于长词语效果不佳。如本图谱中涉及到的“中国科学院院士”一词，用 LTP 工具会将其分为“中国”、“科学”、“院”、“院士”，这样我们的系统可能无法做相关的问题回答。后续可以采用一些对长词语分词效果较好的模型或工具。