一种基于互联网文本的开放式实体关系抽取

/\* 传统的实体关系抽取方法是封闭领域的而且需要提前定义一个完整的关系类型体系，无法适应互联网大数据的需求，开放式实体关系抽取很好的解决了这个问题。

\*/

摘要：

随着互联网大数据的流行，开放式实体关系抽取受到了广泛的关注。传统的开放式实体关系抽取通常只考虑指示词的位置信息以及指示词的词性，本文加入了对句法结构的考量提出一种基于句法结构树的实体对最小路径距离算法。信息增益可以很好的评价指示词对不同实体对类型的区分能力，本文提出一种基于词相似度的信息增益网结构，通过近义词信息增益的补充可以得到更科学合理的评价结果，当指示词由于语料的原因不能得到理想的信息增益时也可以通过增益网中的近义词进行调整。经实验验证，本文提出的方法取得了很好的效果。

引言：

实体关系抽取是信息抽取的子任务，指从文本中抽取出实体之间的关系形成“实体-关系-实体”三元组。随着互联网的发展进入到Web2.0时代，互联网文本信息开始呈现爆炸式的增长。互联网文本数据具有海量，更新速度快，领域多样性等特点，为传统的实体关系抽取带来了挑战。开放式实体关系抽取技术[2]不需要预先定义关系类型体系，而是使用实体对上下文中出现的词语描述实体关系，具有开放领域开放关系类型的特点，可以用来处理海量的互联网文本信息。

英文的开放式实体关系抽取的相关研究已经进行了很长的时间，有了一些成熟的研究成果。。。。。。

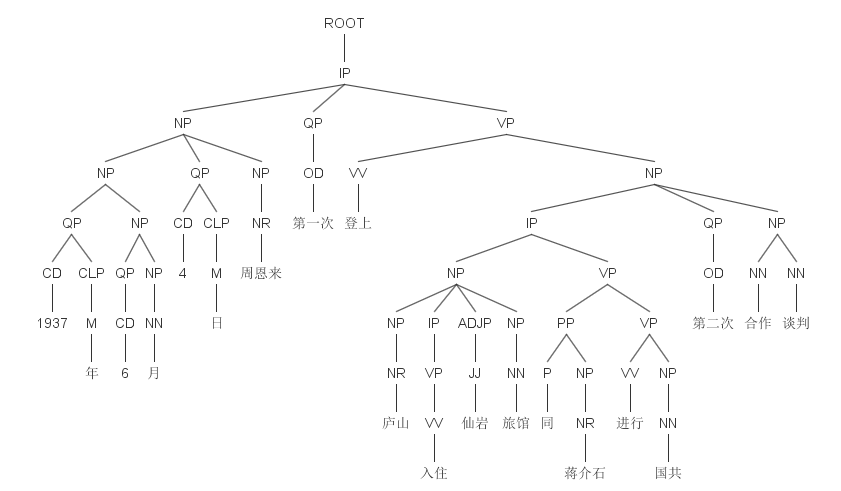
目前中文的开放式实体关系抽取的相关研究还比较少。由于中文与英文相比词语更加的灵活多变、构词更复杂且没有明显的标志，因此针对英文的研究成果很难直接应用到中文领域。/\*王莉峰\*/。针对中文文本目前主要的思路是首先根据一定的规则抽取出候选关系指示词，然后对候选关系指示词进行去噪处理得到关系词词表，最后根据词表对文本进行关系抽取。[3刘安安][4计算机研究与发展2016]提出根据词语与实体对的位置关系以及词性选取指示词，经实验验证取得了不错的效果，但这种选取方法没有考虑实体对与关系词之间的语义信息，因此会引入大量噪声。[康丽丽][地大]考虑了句子的语义信息，通过制定一系列的启发式规则提取指示词，但这种人工定义抽取规则的方法具有一定的局限性。本文提出一种基于句法结构树的方法抽取关系指示词，通过在句法树上应用最小路径距离算法计算句子中所有名词和动词与实体对的距离，得到词语和实体对的相关程度。不仅考虑了句法结构信息，而且算法是通用的不需要针对不同语料对抽取规则做调整。经实验验证。。。/\*好\*/。

/\*引出基于词相似度的信息增益网结构\*/

经过第一轮粗抽取得到的关系三元组仍旧可能出现抽取错误的情况，需要进一步的筛选。/\*比如“张三的朋友李四出生于吉林”这句话，有可能提取出张三-朋友-吉林这样的关系三元组，显然这是不合理的。\*/[刘安安]通过分析中文关系抽取语料库发现同一个指示词往往只出现在特定的关系三元组中，提出使用信息增益来评价指示词对不同实体对类型的区分能力。由于我们获得的语料库可能不够理想，不能充分体现出所有指示词在不同实体对类型上的分布，如果语料中指示词的分布不够理想，此时只用指示词的信息增益来做评价的话会出现偏差。在评估指示词对不同实体对类型的区分能力时词义相近的词可以相互补充。比如抵达和赶到，如果抵达计算得到的信息增益很高的话赶到也应该会很高。本文提出一种基于词相似度的信息增益网结构，使用NetIG评价指示词对不同实体对类型的区分能力，经实验/\*好\*/，尤其是在处理不同领域语料时。。。。

1. 复句分解

复句由两个或两个以上意义相关，结构上互不作句子成分的分句组成。在分析复句的时候，句法分析树往往不能得到理想的结果。考虑这样一个句子，“1937年6月4日，周恩来第一次登上庐山，入住仙岩旅馆，同蒋介石进行国共第二次合作谈判”，分析得到的句法树如图1所示。可以看到句法分析器将“周恩来入住仙岩旅馆”错误的分析成了“庐山入住仙岩旅馆”。



本文通过制定一系列规则将常见的复句分解为单句，句法分析器在处理单句时可以得到更准确的分析结果，从而可以抽取到正确的关系三元组。两种复句分解的规则：

1）主句包含完整的主谓宾结构，各个分句则缺少主语，将主句的主语补充到各个分句即可。例如，“1937年6月4日，周恩来第一次登上庐山，入住仙岩旅馆，同蒋介石进行国共第二次合作谈判”可分解为“1937年6月4日，周恩来第一次登上庐山”、“周恩来入住仙岩旅馆”以及“周恩来同蒋介石进行国共第二次合作谈判”。

2）主句与各个分句之间由起关联作用的词语连接成复句，如“但”、“并且”等，通过构建关联词词典检测这种类型的复句，将关联词去除，各个分句即可构成单句。例如，“上午，习近平抵达安徽，之后习近平前往金寨县向革命烈士纪念塔敬献花篮”可分解成“上午，习近平抵达安徽”以及“习近平前往金寨县向革命烈士纪念塔敬献花篮”。

2. 关系三元组抽取

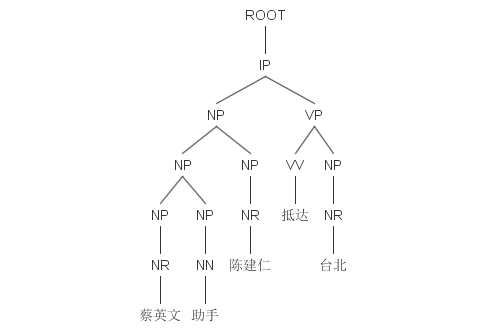
通过限制词语和实体对的位置关系获取关系指示词会带来一定的噪声，可以通过引入句法信息来达到降噪的目的。现有的引入句法信息的思路主要是通过人工定义规则来实现，规则的制定通常很复杂，实际操作起来比较困难，而且需要专业的汉语言知识，另外在处理多领域文本时移植性较差。本文提出一种简单的基于句法结构树的最短联通路径和算法来抽取关系三元组。

2.1 获取关系实体对

对于待处理的句子，首先需要获取句子中存在关系的实体对。这里是通过加入两个简单的限定，实体之间其他实体个数的限定以及实体之间词语个数的限定[2]来实现的。比如句子“蔡英文助手陈建仁抵达台北”，可以抽取出实体对“蔡英文-陈建仁”和“陈建仁-台北”。

2.2 最短联通距离和算法

获得了关系实体对以后，使用斯坦福大学的句法分析工具Stanford Parser将句子转化为句法分析树。对于每个实体对，计算句子中所有非实体词与这两个实体的最短联通距离之和，选择距离和最小的动词或者名词作为该实体对的关系指示词。图2所示为句子“蔡英文助手陈建仁抵达台北”的句法分析树。句子中共包含两个非实体词“助手”和“抵达”，对于实体对“蔡英文”和“陈建仁”，“助手”到这两个实体的最短联通距离和为5+6=11，“抵达”为6+7=13，因此得到三元组“蔡英文-助手-陈建仁”。同理可以得到三元组“陈建仁-抵达-台北”。如果采用位置限定的方法抽取关系指示词则会抽取到噪声三元组“蔡英文-抵达-陈建仁”和“陈建仁-助手-台北”。



3. 关系三元组过滤

3.1 构建关系指示词词表

基于句法分析树的最短联通距离和算法以句子为单位抽取句子中包含的关系三元组，只考虑了局部的特征，得到的关系三元组仍存在一定的噪声。[2]通过分析中文关系抽取语料库，发现可以表示实体对关系的关系指示词即正确的关系指示词往往只出现在特定的实体对类型的三元组中，如关系指示词“爸爸”、“妈妈”基本只出现在实体对类型为“人-人”的关系三元组中，也就是说正确的关系指示词可以区分不同的实体对类型。利用这一发现，[2]提出了一种基于信息增益的全局关系指示词排序方法，将每个关系指示词看做是一种特征，对于每个关系三元组该特征存在两个值，即三元组包含这个关系指示词以及不包含该关系指示词，利用信息增益的计算公式可以得到衡量关系指示词正确性的指标。

本文在[2]方法的基础上增加近义词聚集的过程来计算关系指示词的信息增益值从而构建关系指示词词表。待处理的数据集并不能保证每个关系指示词在各个实体对类型上的分布都是充分的，正确的关系指示词可能由于相关语料不够充分的原因被判断为噪声词。通过计算关系指示词之间的相似度，把相似度高的近义词聚集成一个集合，看做是一个词计算其信息增益值，集合中所有的词共享这个信息增益值进行最终的排序。这样做的另外一个好处是可以兼顾到不同领域的文本，比如“到达”和“抵达”这一对词，“到达”多用于语境比较随意的场景比如娱乐新闻，而抵达多用于正式的场合比如时政新闻，而这两个词都可以用来描述“人-地点”实体对类型。通过近义词聚集的方法可以将不同领域的文本融合在一起进行处理。

3.2 近义词聚集

3.3 相似度计算

近义词聚集需要计算关系指示词之间的相似度，[4]提出了一种基于《知网》的词语相似度计算方法。该方法采用了一种更为结构化的方式改写了《知网》中词的定义（DEF），在研究了义原的相似度计算方法、集合和特征结构的相似度计算方法的基础上提出了利用《知网》进行词语相似度计算，由实验验证取得了不错的效果。本文利用[4]中的方法计算关系指示词之间的相似度，将词义相近的词聚集在一起。