一种无指导的中文开放式实体关系抽取方法

摘要：

开放式实体关系抽取解决了传统的实体关系抽取需要事先定义关系类型体系、标注训练集等问题，适合处理海量的互联网文本数据，但是在中文领域目前研究的还很少。提出一种无指导的中文开放式实体关系抽取方法。首先采用基于句法分析树的最短联通距离和算法抽取句子中的关系三元组，通过复句分解，提升了句法分析树的性能；然后提出一种基于近义词聚集的全局排序方法对得到的关系指示词进行筛选得到关系词词典；最后根据关系词典对关系三元组进行过滤。经过实验验证，取得了不错的效果。

/\*课题意义，研究现状，本文贡献\*/

引言：

实体关系抽取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，是信息抽取的重要组成部分。传统的实体关系抽取需要事先定义关系类型，然而定义一个全面科学的关系类型体系是很困难的，也限制了实体关系抽取的应用场景。此外，传统的实体关系抽取还面临需要带标注的训练集、无法处理多领域语料等问题。开放式实体关系抽取使用实体对上下文的关系指示词来描述实体关系，可以处理跨领域的文本，因此越来越受到业界的关注。随着互联网的蓬勃发展，针对互联网文本数据的处理成为了一个亟待解决的问题，更是一项挑战。互联网文本具有更新速度快，领域多样性等特点，开放式实体关系抽取相比传统的实体关系抽取更适合海量互联网文本的处理任务。

英文的开放式实体关系抽取研究已经进行了很长的时间，有了成熟的研究成果[1][2][3]...。但是由于汉语言和英语之间的语言差异使得针对英文的研究成果很难直接应用到中文领域。目前中文的开放式实体关系抽取的相关研究还比较少。王莉峰[1]提出了一种领域自适应的中文实体关系抽取方法。通过最大程度地避免人工参与，解决了关系类型自动发现、抽取实体关系核心网、抽取关系三元组等问题，在人物社会关系抽取任务中取得了不错的效果。秦兵、刘安安[2]等人提出了一种无指导的开放式中文实体关系抽取方法。首先根据实体之间的距离限制和关系指示词的位置限制抽取关系三元组，然后通过对关系指示词进行全局排序以及类型排序构建关系指示词词典，最后根据词典对候选关系三元组进行过滤。经过实验，在针对互联网文本的处理任务中取得了不错的效果。[3]提出了一种基于聚类的开放式实体关系抽取方法。针对中文的句法表达，首先提出了基于实体对位置、基于句法分析和基于马尔可夫逻辑网概率模型的关系指示词提取方法。采用凝聚层次聚类，使用组平均连通作为簇相似度的度量方法，将具有相似关系的实体对聚到同一个类中，然后对每个类做标注从而得到关系三元组结果。实验结果表明，该方法取得了比较好的效果，能够获取大量有价值的实体关系。

本文总结了前人的研究成果提出一种无指导的中文开放式实体关系抽取方法。[2]采用了规则匹配的方法，通过限定词语和实体对位置关系抽取候选关系三元组。这种方法由于没有考虑句法信息会引入大量的噪声。[3][4地大]引入了句法信息，但是这种句法信息的引入是通过制定一系列规则实现的，具有一定的局限性，且规则的制定需要具备一定的汉语言专业知识。本文提出一种基于句法分析树的最短联通距离和算法获取句子中的关系三元组。句法分析树包含了关系词与实体对之间的句法关系，通过最短联通距离算法可以抽取最能描述实体对关系的关系指示词。此外针对复句的情况，本文提出了对常见的复句类型进行分解的解决方案。经过复句分解，句法分析器的性能得到了提升。相比较制定句法规则，本文提出的方法更加的简单且易实现。相比较基于规则匹配抽取关系指示词的方法，实验表明，本文提出的方法拥有更高的正确率和召回率。得到的关系三元组是以句子为单位抽取的，需要采用一种全局的统计方法对结果进行过滤。本文采用[2]提出的全局排序的方法构建关系词词典对结果进行过滤，针对全局排序的不足之处做出改进，提出一种基于近义词聚集的全局排序方法构建关系词词典。实验表明，经过近义词聚集，得到的关系词词典更具有代表性。

关系词举例表

粗提取正确率召回率对比

最终的正确率召回率对比

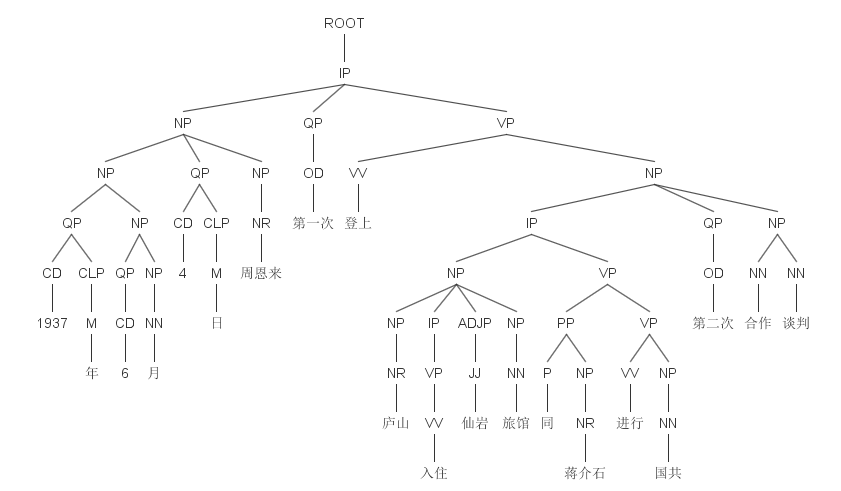
关系三元组样例

++++++++++++++++++++++++加入句法分析器的介绍-Stanford Parser！！！！！！！！！！！

++++++++++++++++++++++++types的介绍{五中类型}

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

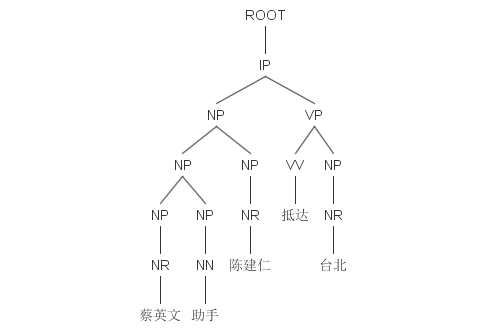
通过限制词语和实体对的位置关系获取关系指示词会带来一定的噪声，可以通过引入句法信息来达到降噪的目的。现有的引入句法信息的思路主要是通过人工定义规则来实现，规则的制定通常很复杂，实际操作起来比较困难，而且需要专业的汉语言知识，另外在处理多领域文本时移植性较差。本文提出一种简单的基于句法结构树的最短联通路径和算法来抽取关系三元组。

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。