第32卷 第21期 计 算 机 工 程 2006年11月
Vol.32 No.21 Computer Engineering November 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)21—0183—02 文献标识码：A 中图分类号：TP18

基于种子自扩展的命名实体关系抽取方法

何婷婷，徐 超，李 晶，赵君喆
(华中师范大学计算机科学与技术系，武汉 430079)

摘 要：命名实体间关系的抽取是信息抽取中的一个重要研究问题，该文提出了一种从大量的文本集合中自动抽取命名实体间关系的方法，找出了所有出现在同一句子内、词语之间的距离在一定范围之内的命名实体对，把它们的上下文转化成向量。手工选取少量具有抽取关系的命名实体对，把它们作为初始关系的种子集合，通过自学习，关系种子集合不断扩展。通过计算命名实体对和关系种子之间的上下文相似度来得到所要抽取的命名实体对。通过扩展关系种子集合的方法，抽取的召回率和准确率都得到了提高。该方法在对《人民日报》语料库的测试中，取得了加权平均值F-Score为0.813的效果。

关键词：命名实体；关系抽取；自学习

Named Entity Relation Extraction Method Based on Seed Self-expansion

HE Tingting, XU Chao, LI Jing, ZHAO Junzhe
(Department of Computer Science and Technology, Huazhong Normal University, Wuhan 430079)

【Abstract】 Named entity relation extraction is an important issue in information extraction. This paper proposes a special method that extracts named entity relation from large text rendezvous. It finds out the named entity pairs, which appear in the same sentences and the distances of them is under a certain value, and converts their contexts into vectors. It selects a few named entity pair instances that have the relation wanted to extract and make them as initial relation seed set. The relation seed set is extended automatically in self-study process. It gets the named entity pairs, which have the relation wanted to extract, by calculating the similarity of context vectors between named entity pairs and relation seed set. By the method of bootstrapping, the recall and precision are enhanced. It verifies the method with the PFR corpora and achieves an average weighted F-Score of 0.813.

【Key words】 Named entity; Relation extraction; Self-study

信息抽取，就是从自然文本中提取出预先指定好的信息，并给出该信息的结构化描述。命名实体包含人名、地名、机构名等名词性词语，命名实体之间关系的抽取是信息抽取中的一个重要研究课题。它与信息检索、问答系统、信息过滤有直接的关联，作为一项基础性研究，它对于自动文摘、机器翻译、内容理解、语境生成、文本分类、信息过滤以及数字图书馆建设都有重要的研究意义。绝大部分命名实体之间关系抽取的方法都是有导的学习方法。最初，人们使用的是知识工程方法(Knowledge Engineer Approach)。如Kernel方法[1]首先对句子进行浅层句法分析，然后直接使用字符串作为处理对象来计算2个对象之间的相似度。该类方法主要靠手工编制规则，使系统能处理特定知识领域的抽取问题。这种方法在某些领域取得了不错的效果，但是它要求编制规则的知识工程师对该知识领域有深入的了解，开发过程耗时耗力。

为了解决知识工程方法的不足，人们又使用了自动训练方法(Automatic Training Approach)。例如Brin方案[2]首先定义少量种子，然后通过发现种子中的共同模式来发现新的具有这种关系实体对。

Agichtein方案[3]对前后2个命名实体的类型进行限制而改进了Brin的方法。自动训练方法仅需要少量人工定义好的种子，通过系统学习获取新的规则，经训练后能处理其没有见过的新文本。

1 基于种子自扩展的命名实体关系抽取方法

本文提出的方法的主要特点体现在自学习过程中对关系种子扩展的方法上。只需要人工定义很少量的几个关系种子，通过自学习，种子不断得到扩展。在自学习过程中，把满足条件的命名实体对作为候选关系种子，然后通过让它们候选，参加对命名实体对关系的判断，来决定是不是将候选关系种子添加到关系种子集合中。

本文提出的方法由下面几个步骤完成：

(1)找出所有出现在同一句子内的、词语距离在一定范围之内的命名实体对。对这些命名实体对的上下文进行向量化。这些命名实体对的集合是 P；

(2)人工从命名实体对集合中选出少量具有所要抽取关系的命名实体对。把这些命名实体对实例作为初始关系种子集合 S；

(3)利用关系种子集合 S 对命名实体对集合 P 进行训练。通过自学习，关系种子集合得到了扩展，扩展以后新的关系种子集合为 S'；

(4)如果新的关系种子集合 S' 等于关系种子集合 S，则通过计算命名实体集合 P 和关系种子 S 之间的相似度来得到所要抽取的命名实体对；否则，关系种子集合 S 等于新的关系种子集合 S'，跳到第(3)步。

基金项目：国家自然科学基金资助项目(60442005)；教育部科学技术研究基金资助重点项目(105117)

作者简介：何婷婷(1964－)，女，教授、博士，主研方向：自然语言处理，复杂网络；徐 超、李 晶、赵君喆，硕士生

收稿日期：2005-11-20 **E-mail：**ccnujsj@etang.com

1.1 命名实体对的识别

命名实体关系抽取之前必须首先识别命名实体，这就需要对命名实体进行标注。在同一个句子中，词语之间距离在一定范围之内的 2 个命名实体之间的关系才比较明确，所以把出现在同一句子中，并且 2 个命名实体的词语距离在一定范围之内的 2 个命名实体称为一个命名实体对。根据前后 2 个命名实体的类型，可以把命名实体对划分到不同的域中。把前一个命名实体的左边的词语数，两个命名实体之间的词语数，后一个命名实体的右边的词语数，称为上下文窗口。在存在命名实体对的句子中，把出现在上下文窗口内的命名实体对的上下文称为该命名实体对的上下文。例如，某个标注过句子为$w_{1m},\ldots,w_{11},NE_1,w_{21},w_{22},\ldots,w_{2i},NE_2,w_{31},\ldots,w_{3n}$，其中$NE_1$是一个地名，$NE_2$是一个人名。当上下文窗口大小为*left-mid-right*，如果*i*的值小于等于*mid*，则这个句子中的命名实体对为<NE_1,NE_2>，该命名实体对是在<地名，人名>这个域中的，该命名实体的上下文是

$$w_{1,left},\ldots,w_{11},\ w_{21},\ w_{22},\ldots,w_{2i}\ w_{31},\ldots,w_{3,right}$$

1.2 对上下文的向量化和手工选出的少量关系种子

命名实体对之间的关系可以由上下文来确定。通过向量空间模型[4]对命名实体对的上下文进行向量化，使之进行数学上的分析。向量空间模型具有概念简单、应用方便、利用空间相似性来逼近语义相似性等优点。

命名实体对的上下文中词语在向量中的权重的表示方法有很多。这些词语对命名实体对之间关系的描述能力是随着它们相对位置的变化而变化，本方法通过鲁松、白硕等提出的词语位置的信息增益[5]来计算词语在向量中的权重。一个词语相对前一个命名实体和相对后一个命名实体的信息增益的平均值就是该词语在命名实体对上下文向量中的权重。这样，每个命名实体对的上下文都被表示成一个向量。通过在数学上对向量的分析，来得到命名实体对之间的关系。

在找出的命名实体对中，通过人工的方法选出少量具有所要抽取关系的命名实体对，把这些命名实体对实例作为初始关系种子集合。

1.3 计算命名实体对之间关系的相似度

计算命名实体对之间关系的相似度，是通过计算它们之间上下文向量的相似度来完成的。命名实体对的上下文向量P_i与关系种子的上下文向量集合中的某个向量S_j的相似度可以通过 2 个向量之间的夹角的余弦来表示：

$$sim(P_i,S_j)=\frac{<P_i,S_j>}{\sqrt{<P_i,P_i>*<S_j,S_j>}}$$

其中，〈A,B〉表示向量 A 与向量 B 的内积。

通过计算命名实体对和关系种子之间的上下文向量的相似度，得到该命名实体对与关系种子集合的上下文向量相似度集合。取相似度集合中值最大的一个作为该命名实体对与关系种子集合之间的相似度，即

$$\mathrm{sim}(P_i, S)=\max\{\mathrm{sim}(P_i ,S_j)\}$$

当命名实体对与关系种子集合之间的相似度大于事先设定好的阈值时，则认为该命名实体对之间的关系与关系种子之间的关系是相同的。

1.4 关系种子集合的扩展方法

由于起初仅手工选取少量关系种子，因此需要通过自学习来对关系种子集合进行扩展。把具有相同关系的命名实体对看作是一个类，则这些命名实体对的上下文向量在空间上是聚集在一起的。关系种子可以被看作是这个类在空间上的若干个质心，对关系种子集合的扩展可以看作是添加新的质心到这个类中。

在对关系种子集合进行扩展时，利用Chetan Gupta等提出的聚类的思想[6]，在对关系种子集合进行扩展时，把每一个具有该关系的命名实体对作为一个候选关系种子，然后对该候选关系种子进行评估，当候选种子通过评估，则将该候选关系种子添加到关系种子集合中，否则该候选关系种子将会被删除。对候选关系种子的评价是通过让它候选参加对其他命名实体对关系的判断来完成的，利用已有的关系种子集合确定出一些具有该关系的命名实体对，将这些命名实体对交给候选关系种子去判断，对关系种子进行扩展算法如下：

输入 候选关系种子T_i，命名实体对集合P，关系种子集合S，扩展评价大小N，相似度阈值*threshold1*，关系种子得分阈值*threshold2*。

输出 是否将候选关系种子T_i添加到关系种子集合S中。

(1)利用 S 在命名实体对集合 P 中找出 N 个具有所抽取关系的命名实体。这 N 个命名实体对组成集合 R；

(2)用候选关系种子T_i对集合R中的命名实体对的关系进行判断，抽取出M个命名实体对；

(3)如果识别率M/N大于关系种子得分阈值*thresh-old2*，则将该候选关系种子实例T_i添加到关系种子集合中；否则删除该候选关系种子。

2 实验及其结果

2.1 实验数据及评测指标

本文使用的是《人民日报》标注语料作为实验数据，该语料是对《人民日报》1998 年上半年的纯文本语料进行了词语切分和词语类型标注制作而成的。

本文中的每个词语都带有词语类型标记，标注的格式为“词语/词语类型”，即词语后面加单斜线，再紧跟词语类型标记。词与词之间用 2 个单字节空格隔开。对于实验结果的性能评测，通过以下指标对系统的性能进行评价。

$$(1)\ \text{召回率}=\frac{\text{系统抽取的命名实体对}\cap EP}{EP}$$

$$(2)\ \text{准确率}=\frac{\text{系统抽取的命名实体对}\cap EP}{\text{系统抽取命名实体对}}$$

$$(3)\ \text{F-Score}=\frac{2*\text{召回率}*\text{准确率}}{\text{招回率}+\text{准确率}}$$

其中，*EP* 表示所有出现在同一个句子中的具有所要抽取关系的命名实体对。

2.2 实验结果

在实验中，我们对在<地名，人名>这个域中的命名实体对做了测试。在实验语料中，一共抽取到出现在同一句子中的、属于<地名，人名>这个域中的命名实体对 4 528 个。在收集命名实体对上下文时，剔除掉了一些停用词和在所有上下文中出现次数总数少于 5 的词语。实验对其中的<国家，国家领导人>这类关系进行了抽取。实验中的各个参数的值如表 1 所示。不同上下文窗口下的实验结果如表 2 所示。

表 1 实验中各个参数的设置

初始关系种子数	相似度阈值	扩展评价大小	候选关系种子得分阈值
6	0.75	20	1.2 / 现有种子数

表 2 上下文窗口取不同值时系统抽取的结果

上下文窗口	候选命名实体对数	召回率	准确率	F-Score
0-4-0	1 562	56.3%	75.5%	0.645
0-6-0	2 615	79.6%	83.1%	0.813
2-6-2	2 615	71.8%	79.3%	0.753
0-8-0	2 978	73.4%	78.9%	0.761

（下转第 193 页）