DOI:10.13833/j.cnki.is.2012.05.019

基于规则和统计相结合的中文命名实体识别研究

潘正高

(宿州学院 信息工程学院,安徽 宿州 234000)

摘 要:介绍命名实体识别在文本信息处理领域的重要地位,分析了中文命名实体识别存在的困难,介绍中文命名实体识别的一般过程、评价标准及方法。提出了一种在构造内部规则和外部规则的同时采用概率统计的中文命名实体的识别方法,并利用这种基于规则和统计相结合的方法。实验证明该方法获得了较高的准确率和召回率,具有可行性和合理性,同时也指出了它的局限性。

关键字:命名实体;文本特征;中文命名实体;识别

中图分类号:G350 文献标识码:A 文章编号:1007-7634(2012)05-708-05

Research on the Recognition of Chinese Named Entity Based on Rules and Statistics

PAN Zheng-gao

(School of Information Engineering, Suzhou University,Suzhou 234000*,China)*

Abstract: This paper described the important position of named entity recognition in the field of text information processing, analyzed the problem of named entity recognition, described the general process, the evaluation criteria and the methods of Chinese named entity recognition. This article puts forward the recognition method of the Chinese named entity through constructing internal and external rules and adopting the statistic method. The experiments proved that this method gains higher precision and recall and has the feasibility and rationality through recognizing and testing the datum. At the same time, the limitation of this method was analyzed.

Keyword: named entity; text feature; Chinese named entity; recognition

文本是自然语言描述信息的最基本形式之一。文本的有关处理技术是自然语言信息处理研究领域的一个重要研究方向。在一篇文档中,命名实体作为重要的信息元素,通常包含了该文档的主要信息。因此,准确地识别命名实体是正确理解文档内容的关键。同时,利用命名实体识别可以从文档中提取出实体字符串,有利于在没有浏览全文的情况下快速理解文章的主要内容。命名实体识别是文本信息处理的基础性工作,研究命名实体识别的方法,提高命名实体识别的准确性,对于文本信息处理研究领域意义重大。

1 命名实体识别概述

命名实体(Named Entity, NE)[1]是指一些具体或抽象的客观实体,例如人、组织、地点、时间等。文本中的命名实体大多是以特定的专有名称出现的,例如人名、组织名、机构名、地名等,也可以是时间、数

收稿日期:2011-11-14

基金项目:国家自然科学基金资助项目(60975034);安徽省自然科学青年基金项目(10040606Q64);安徽省高校省级自然科学基金(KJ2012Z401);宿州学院科研开放平台项目(2011YKF10)

作者简介:潘正高(1978-),男,安徽六安人,硕士,讲师,主要从事Web文本挖掘、自然语言处理研究.

量的表达式等形式。命名实体识别的任务实际上就是指从文本中发现出命名实体,并确定其类别的过程。

1995年9月举行的第六届MUC(Message Understanding Conference)会议第一次引入了命名实体识别。当时给出的任务是识别出特定文本集中的专有名称和数量短语并对它们归类。1998年召开的第七届MUC会议已经将命名实体识别作为主要的研讨议题之一。此次会议将命名实体的类别确定为人名(Person)、地名(Location)、机构名(Organization)、日期(Date)、时间(Time)、百分数(Percentage)和货币(Monetary value)等七种类型。

英文命名实体的识别只要考虑词本身的特征而不需要涉及到分词问题,与中文命名实体识别相比,实现难度相对较低,测试的准确率和召回率可达到97%左右[2]。中文命名实体识别与英文命名实体识别相比存在着以下几个方面的问题:

(1) 实体缺乏明显的特征标志 英文命名实体的首字母大多大写,因此易于识别,但中文命名实体没有明显的标志,加大了识别的难度。

(2) 分词影响命名实体的识别效果。分词的错误容易导致命名实体的边界判断的错误。

(3) 不同类别的命名实体间存在歧义,主要包括边界歧义和分类歧义两个方面。边界歧义是指依据命名实体边界的不同,可以产生不同的识别结果。例如‘王文刚 去机场了’和‘王文 刚去机场了’两个句子中,对词的不同切分会影响识别结果。分类歧义是指同一个命名实体,可以将其标注为几种不同的实体类型。例如‘民主 大街’和‘争取 民主”,词语“民主’在前一个短语中表示地名,而在后一个短语中它则是名词,意义相差很远。

(4) 大部分的命名实体属于未登录词。汉语词汇是一个开放的集合,不可能将全部的词都放入词库。

20世纪90年代初期,国内外的一些学者开始对中文通用命名实体(如:人名、地名、组织机构名等)识别进行了一些研究。例如:清华大学的孙茂松是我国最早做中文姓名识别的,他采用统计的方法计算人名的用字概率。复旦大学的吴立德采用统计和规则相结合的方法对中文人名、组织机构名的识别进行研究,取得了较好的效果。因特尔中国研究中心的ZHANG Yi-Min和ZHOU Joe F等人采用基于记忆的学习(Memory-Based learning,MBL)算法开发了一个抽取中文命名实体及实体间关系的信息抽取系统,该系统在ACL2000上取得了较好的效果[3]。

2 中文命名实体识别

2.1 预处理

自然语言信息的特殊性决定了在进行命名实体识别前要做一些预处理工作,且中文文本的预处理对中文命名实体识别效果影响较大,在实际应用中,更应该引起重视。

(1) 中文分词。自然语言均是由字或词组成的,词是最小的有意义的语言单元。中文的词语之间没有明显的边界,因此,对于一个句子的准确切分将直接决定命名实体的正确识别。由此可见,中文词法分析中的分词是命名实体识别的第一步,也是中文信息处理的基础和关键。

(2) 词性标注。词性标注是指依据切分后的分词情况,标出每个词的类型,正确的词性标注是正确进行实体识别的前提。

本文中信息抽取时采用的是中科院计算技术研究所的词法分析系统ICTCLAS,其主要功能包括中文分词、词性标注、实体识别、新词识别等。按照973专家组评测结果,ICTCLAS分词识别率97.58%,基于角色标注的未登录词识别能取得高于90%的召回率,其中中国人名的识别召回率接近98%[4]。汉语词法分析系统ICTCLAS提供了Windows、Linux系统下的C接口和JNI接口,使用方便。

2.2 中文命名实体识别的一般过程

在汉语描述的中文文本中,命名实体的识别一般要经历下面两个阶段。

(1)在分词的同时,标注出词表中已经收集的命名实体。

(2)在此基础上,调用构建好的命名实体识别模型,对文中的尚未标记出的命名实体进行识别。如图1所示。

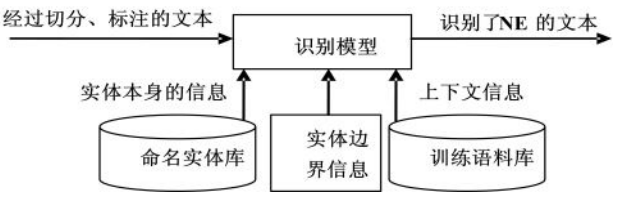

图1 命名实体识别任务

命名实体识别模型的构建应遵循如下的原则。

(1)利用实体本身信息及上下文的信息。

(2)利用词语和词类的信息。

(3)合理地使用先验知识。

实体本身信息是指实体自身的构成信息,上下文信息是指实体所在的上下文对其的约束作用。

2.3 中文命名实体识别的评价标准

通常我们是从实体的边界和实体类型两个方面来评判命名实体的正确性的。在命名实体识别过程中,实体的边界正确但类型有可能错误,反之,实体的边界错误但类型可能正确。因此,一个命名实体识别系统的结果主要有下面几种情况。

(1)正确(correct):系统识别的结果与标准结果一致。

(2)丢失(missing):系统没识别出来的命名实体,但标准结果中有。

(3)虚假(spurious):系统识别为命名实体,但标准结果中却没有。

通常用召回率和精确率来评价命名实体识别系统的性能。一个命名实体识别系统的召回率是指该系统正确识别出的结果占所有结果的比例,计算公式如下:

$$召回率=\frac{Count(正确)}{Count(正确)+Count(丢失)} \quad (1)$$

而命名实体识别的精确率是该系统正确识别的结果占所有识别结果的比例,计算公式如下:

$$精确率=\frac{Count(正确)}{Count(正确)+Count(虚假)} \quad (2)$$

在实际的应用中,为了评价命名实体识别系统的综合性能,一般采用召回率和准确率的加权及和平均值(即F指数)来作为系统性能的评测标准。系统的F值计算公式如下:

$$F=\frac{(\beta^2+1)*准确率*召回率}{\beta^2*准确率+召回率} \quad (3)$$

式(3)中的β表示召回率和准确率的相对权重。当β=1时,表示召回率和准确率二者同等重要;若β>1,则表示准确率更重要一些;若β<1,则表示召回率更重要一些。一般情况下,β值取1。

3 中文命名实体识别的常用模型及算法

3.1 基于规则的中文命名实体识别方法

基于规则的中文命名实体识别方法是利用中文命名实体的内部结构以及上下文的边界特征等信息来手工建立命名实体的识别规则。汉语自身的一些词法和语法为这些规则的构建提供了很好的依据。例如:"中国人名'张三/nh'可以利用词性信息结合中国人名姓氏用字表、名用字表来识别;复杂的人名可以利用头衔边界信息来识别,例如'王经理'、'何主席'、'老杨'、'阿扁'等。但是,这些规则依赖于设计者的直觉,主观性很强,同时,由于不同类型文档的词语分布规则存在差异,所以将某一领域文档中提取的实体识别规则应用到另一个领域时通常就会出现问题。

人工构建规则的最大缺点是代价很大。基于规则的识别系统的建立不仅需要有计算语言学背景的专家的参与,而且要付出长时间的代价。一旦将该系统应用到一个新的语言或新的领域时,往往需要付出更大的代价来修改规则。

3.2 基于统计的实体识别方法

基于统计的命名实体识别方法主要是利用大型标注语料库来训练,得出某个字作为命名实体组成部分的概率,并以此为基础来计算某个候选字段作为命名实体的概率,若大于某一阈值,则识别为命名实体。

(1)N-gram模型[5]。该模型建立在统计概率理论的基础之上的,它的前提是假设文本中某一个词的出现概率取决于它前面的若干个词出现的概率。

在N-gram模型中,对于文本中的某一个词w_i,若已知其前两个词$w_{i-2}w_{i-1}$,则可以用条件概率$p(w_i|w_{i-2}w_{i-1})$来计算w_i出现的概率。

一般来说,由n个词组成的词序列$W=w_1\ w_2\ldots\ldots w_n$,其在文本中出现的概率$p(W)$可以用概率的乘积表示,即:

$$p(W)=p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\ldots\ldots p(w_n|w_1w_2\ldots\ldots w_{n-1}) \quad (4)$$

显然,要计算出词w_n的出现概率,必须知道它前面所有词的出现概率,这样导致计算非常复杂。但如果词w_i的出现概率只与它前面的两个词$w_{i-2}w_{i-1}$有关,问题就得到了极大的简化。这样得到的模型称为三元模型(tri-gram):

$$p(W)=p(w_1)p(w_2|w_1)\prod_{i,i=3\ldots n}P(w_i|w_{i-2}w_{i-1}) \quad (5)$$

(2)基于HMM的模型。隐马尔科夫模型[6](Hidden Markov Model,HMM)作为一种描述随机过程的统计方法,能较好地捕获状态转移信息,且经典的

Viterbi算法[7]在求取最佳状态序列时表现的高效性，是的HMM在命名实体识别领域应用广泛。

HMM可以用一个五元组{S ,O, P, A, B}来表示，其中：

$S=\{S_1,\cdots, S_n\}$表示状态的集合；

$O=\{O_1,\cdots,O_m\}$表示观察值的集合；

$P=\{P_i\}$表示状态的初始概率；

$A=\{a_{ij}\}$表示从状态S_i到状态S_j的转移概率矩阵；

$B=\{b_{jk}\}$表示从状态S_j观察到O_k的发射概率矩阵。

对于给定的词性序列$W=w_1,w_2,\cdots,w_m$，实体识别的目的就是要找到一个最优的实体标注序列$T=t_1, t_1,\cdots, t_m$使得条件概率$P(T|W)$达到最大。由贝叶斯公示可得：

$$P(T|W)=\frac{P(W,T)}{P(W)}=\frac{P(T)P(W|T)}{P(W)} \tag{6}$$

假设转移概率只与前一个状态有关，输出观察值的概率只与当前状态有关，则有：

$$P(T)=\prod_{i=1}^{m}P(t_i|t_{i-1}) \tag{7}$$

$$P(W|T)=\prod_{i=1}^{m}P(w_i|t_i) \tag{8}$$

其中：$P(t_i|t_{i-1})$表示状态t_{i-1}到状态t_i的转移概率。$P(w_i|t_i)$表示在状态t_i出现的条件下观察到w_i的发射概率。而对于一个给定的词性序列W来说，P(W)的值是确定的，可以不予考虑。所以，最终的输出T*可以表示为：

$$T*=\arg\max_{r}[\sum_{i=1}^{m}(\log P(t_i|t_{i-1})+\log P(w_i|t_i)] \tag{9}$$

HMM模型的命名实体识别过程就是以当前输入的词性为训练参数来标注最优的状态序列的过程。即在模型$\lambda=\{A,B,P\}$和观察值序列W给定的条件下找出概率最大的状态序列：

$$Q*=\arg\max_{Q}P(Q|W,\lambda) \tag{10}$$

本文采用动态规划的Viterbi算法求解最佳的状态序列，归纳后有：

$$\delta_{S+1}(j)=(\max_{i}\delta_S(i)a_{ij})\times b_j(w_{S+1}) \tag{11}$$

4 规则和统计相结合的中文命名实体识别

4.1 规则和统计相结合的实体识别方法

规则和统计相结合的方法是自然语言处理领域常用的一种信息处理方法[8]，它可以将信息的规则特征和统计特征结合起来，有利于提高系统的性能。在命名实体识别中，其优势体现在以下两个方面。

(1)通过统计概率的计算可以大幅度降低规则方法处理的复杂度，减少规则使用的盲目性。

(2)加入的识别规则可以降低系统对大规模语料库的依赖。

在命名实体中人名、地名、机构名和专有名词构成规则相对于时间和日期来说更为复杂，是我们研究的重点对象。这里的规则包括用于确定实体边界的外部规划和描述实体内部结构的内部规则。

(1)外部规则。处于具体语言环境中的命名实体，其上下文信息必然影响它的边界和类别。我们从人民日报语料库中提取的实体上下文规则，帮助命名实体的识别。

①人名的外部规则

在一个句子中，人名的前面或后面常有一些指示词，如“主席”、“教授”、“说”等，因此，这些标志性很强的指示词为我们确定实体的边界提高了很大的帮助。

据此，我们构建了两类规则：一是，若当前词的词性是nh，且前一个词为人名的前缀触发词，则当前的词为人名；二是，若当前词的词性是nh，且后一个词为人名的后缀触发词，则当前的词为人名。

②地名、机构名的外部规则

我们研究发现，地名、机构名前面的词语对其边界的确定没有太大的贡献。所以我们只选取实体后面的词语作为外部规则。

规则一：若当前词的词性为ns，且后一个词的词性为p(介词)、u(助词)、w(标点)、b(区别词)、v(动词)时，认定该词是地点名。

规则二：若当前词的词性为ni，且后一个词的词性为p(介词)、w(标点)时，认定该词是机构名。

(2)内部规则。从汉语构词的角度来说，命名实体一般是由一个或多个词构成的，其内部的构成规则我们称之为内部规则。

①人名的内部规则。

对于简单的人名，例如“姜文”，我们可以直接通过词性和外部规则来识别。对于一些复杂的人名或人名指代，我们从语料库中提取了以下几个规则：

规则一：(“小”|“老”)+(姓氏用字)，例如：小李，老张等。

规则二：(姓氏用字)+(“老”|“某”)，例如：齐老，蒋某等。

规则三：“小”|“阿”+(名用字)，例如：小娟，阿扁等。

② 地名、机构名的内部规则。

这一类规则主要体现在实体的内部词性顺序和特征词两个方面，前者体现的是实体的内部结构，后者体现的是实体的类型特征。

例如“哈尔滨/ns 工业/n 大学/n”，其内部词性序列为“/ns/n/n”，特征词为“大学”。

③ 时间和日期的内部规则。

时间与日期的差别仅在于时间的精度不同。日期(Nr)表示天一级以上的时间长度，例如，各个节假日、“二十一世纪”、“5月21日”等。时间(Nt)表示天一级以下的时间长度，例如，“9时40分”、“6日下午3点”等。我们经过研究发现规则如下：

规则一：数字+年|月|日[+天一级以上的时间词]，天一级以上时间词指“春季”、“第二季度”、“下半年”等。例如：98年1月，2001年，08年上半年等。

规则二：数字+“年代”，例如：60年代等。

规则三：数字+“世纪”[+数字+“年代”]，例如，十七世纪，二十世纪90年代等。

规则四：天一级时间+[天一级以下的时间词]+数字+时|分|秒|点，天一级以下的时间词指“晌午”、“凌晨”等。

除了上面提到的一些规则外，我们还使用了一些资源文件，它们包括中国人名姓氏表、常用地名词表、地名(机构名)触发词表、地名(机构名)后缀词表、机构名简称词表、时间表达式停用词表、常见节日词表、数词停用词表、常见量词单位词表等。

4.2 算法描述

(1) 从左向右扫描经分词和词性标注后的文本，若词性序列满足某一实体类别的构成规则，则产生一个该类别的候选实体。

(2) 利用该类别的外部规则进行确认，若满足规则，结束本次识别过程转向(1)；否则转(3)。

(3) 继续扫描紧跟其后的词串，按从语料库中提取的统计规则进行识别。结束本次识别过程后，若文本未结束，则转(1)。

5 实验及结果分析

本文采用1998年1月的人民日报语料前10000句作为测试语料，剩下的部分用于训练。

我们采用准确率(P)、召回率(R)和F值作为评测指标，其计算公式如下：

$$P=\frac{\text{系统标注正确的}NE\text{总数}}{\text{系统标出的}NE\text{总数}}$$

$$R=\frac{\text{系统标注正确的}NE\text{总数}}{\text{测试集中出现的}NE\text{总数}}$$

$$F=\frac{2PR}{P+R}$$

在测试集中，我们考察的几类命名实体构成情况，如表1所示。

表1 测试集的构成

实体(NE)类别	实体(NE)个数
人名(Nh)	3049
机构名(Ni)	5276
地名(Ns)	2491
日期(Nr)	1531
时间(Nt)	37
总计	12384

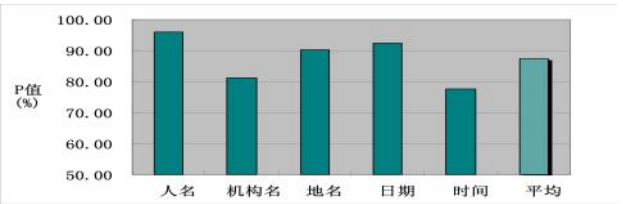

图2 不同类型NE识别结果的P值比较

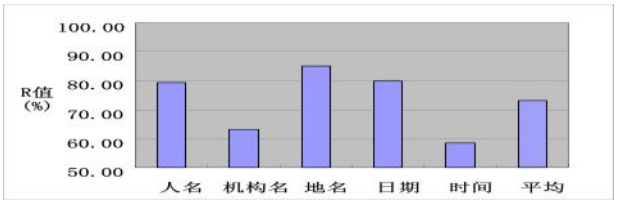

图3 不同类型NE识别结果的R值比较

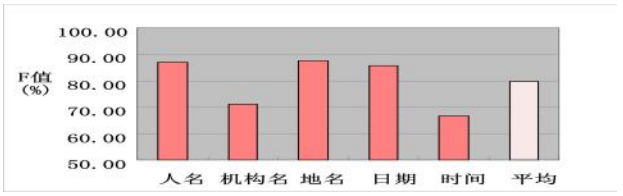

图4 不同类型NE识别结果的F值比较

从图2和图3可以看出，系统对几个类型NE识别都取得了较高的精确度，但是召回率都相对较低，原因在于识别所采用的规则限制条件严格，针对性较强，因此它对实体的识别具有较高的准确率。但是规则并不适合所有的情况，所以召回率较低。从图4可以看出，系统获得的F值较高，但是这些规则受语言环境和领域的影响较大，一旦应用对象发生变化就很难再适用。

6 结 语

中文命名实体识别是自然语言 (下转第786页)

时，建立土地供给信息与我国住宅市场信息对接机制，确保土地供给信息能够准确、及时、全面地传递给信息使用者，最终实现我国住宅市场平稳健康有序的发展。

⑷ 土地供给信息反映出我国房地产开发商存在囤地待开发状况。这就需要以土地供给信息作为依据对住宅市场进行调控。加大对房地产开发商的监管力度，确保土地及时、有效地投入开发，顺利地转化为住宅商品。建立土地供给信息预警监测机制，对土地供给信息利用指标体系，设定相应的安全阀值，当设定的指标超过阀值时，对土地供给过冷或者过热进行综合判断，从而调整土地供给对土地市场进行有效地调控和引导。

参考文献

1 丁 军. 土地出让方式转变推高房价的机制分析[J].中国城市经济河海大学学报，2011，(11):232-234.

2 全诗凡，张 洪.土地供应政策对房价的影响分析——基于空间面板计量的实证分析[J].中国证券期货当代经理人，2006，(1):194-195.

3 郑娟尔.基于Panel Data模型的土地供应量对房价的影响研究[J].中国土地科学，2009，(12):777-780

4 梁云芳，高铁梅. 中国房地产价格波动区域差异的实证分析[J].经济研究，2007，(8):133-142.

5 王 斌，高 波.土地财政、晋升激励与房价棘轮效应的实证分析[J].南京社会科学，2011，(5):28-34

6 高炳华.住宅市场信息与住宅价格规制[J].华中师范大学学报，2011，(11):38-42.

7 卢佳平，张 红.基于价格信息传递的住宅市场空间效率[J].清华大学学报，2008，(12):2041-2043.

8 任荣荣，刘洪玉.土地供应对住房价格影响机理——对北京市的实证研究[J].价格理论与实践，2007，(10):40-41.

9 孙 巍，李 何，李 佳.收入差距、房价及利率政策对消费行为的影响[J].黑龙江社会科学，2008，(6):81-83.

(责任编辑 刘凤琴)

(上接第712页)

处理的一个基础研究领域。中文命名实体识别相对于英文命名实体识别存在着边界模糊、构成规则复杂等特点。本文在认真分析语料的基础上，构造了有利于中文命名实体识别的几个规则，并将这些规则与统计特征结合起来应用到中文命名实体识别中去。实验结果显示，命名实体识别的平均准确率达到87%以上，其中，人名的识别准确率达到近97%，可见实验时构造的规则最利于人名的识别。

参考文献

1 李保利，陈玉忠，俞士汶.信息抽取研究综述[J].计算机工程与应用，2003，(10):1-5.

2 愈鸿魁，张华平. 基于层叠隐马尔科夫模型的中文命名实体识别[C].北京：全国网络与信息安全技术研讨会，2005.

3 Zhang Y M, Zhou J F. A Trainable Method for Extracting Chinese Entity Names and Their Relations[A]. In Proceedings of the Second Chinese Language Processing workshop[C]. Hong Kong：Intel China Research Center, 2000.

4 李向阳，张亚非. 一种网上图书信息抽取方法[J]. 情报学报，2004,23 (6):655-661.

5 周 强.规则与统计相结合的汉语词类标注方法[J].中文信息学报，1995，(2):1-10.

6 刘 群，张华平，俞鸿魁，等. 基于层次隐马模型的汉语词法分析[J].计算机研究与发展，2004，(8):1421-1429.

7 刘 芳. 基于统计的汉语组块分析[J]. 中文信息学报，2000,14(6):28-32.

8 余肖生，孙 珊. 基于信息抽取的文本知识挖掘模型研究[J].情报科学，2010,28(5):776-778,792.

(实习编辑 赵红颖)