情 报 学 报 ISSN 1000－0135
第29卷第6期1074－1079，2010年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000－0135
Vol.29 No.6，1074－1079 December 2010

doi：10.3772/j.issn.1000-0135.2010.06.016

一种电子产品领域命名实体识别方法研究

邹 涛
（西安电子科技大学经济管理学院，西安 710071）

摘要 本文通过研究开源自然语言处理平台 GATE 和条件随机场模型，提出一种高效的电子产品领域命名实体识别策略，为实习项目中的初步工作——通过计算机智能方法识别出电子产品领域的产品品牌、属性等命名实体提出解决方案，并为下一步可能开展的领域内自动问答系统等高层应用提供底层支撑。该方法是基于层叠模型的规则与统计相结合的新的方法，分别继承了基于规则和基于统计识别方法的优点。最终，通过分析电子产品领域自身的领域特点实现了如品牌、重量等二十余种命名实体的识别。对比实验结果表明，该系统达到了令人满意的识别效果。

关键词 命名实体识别 电子产品领域 GATE 条件随机场

Research of a NER Strategy in the Field of Electronic Products

Zou Tao
（*College of Economics & Management*，*Xidian University*，*Xi' an* 710071）

Abstract This paper presents a strategy for named entity recognition（NER）system in the filed of electronic products through studying the open-source platform for natural language processing with GATE and CRF model. The preliminary work of my internship program is to recognize the Named Entity in the areas of electronics products such as product brand and attributes of them. This paper will give a solution and the experimental results will be given as last. For the next step，the system can be carried out in areas such as automatic question answering system for high-level applications. At last，over twenty kinds of NER can be recognized by the system. Comparative experiments are done to prove its system identification results，and the system achieved satisfactory results.

Keywords named entity recognition，filed of electronic products，GATE，conditional random fields

1 引 言

随着因特网和信息产业的快速发展，大量信息以电子文档的形式出现在人们面前，人们迫切希望计算机能对电子文档的文本信息实现自动化处理。命名实体识别（Named Entity Recognition，NER）是目前文本信息自动化处理中一个尚未得到很好解决的问题。命名实体是文本中基本的信息单位，是文本中的固有名称、缩写及其他唯一标识，是正确理解文本的基础。狭义地讲，可以把命名实体分为人名、地名、组织名等。广义地讲，命名实体还可以包括时间表达式、数值表达式等。在各种应用领域，还可以根据具体的需要定义其他类型的命名实体。例如在某个具体应用中，可能需要把住址、电子信箱、电话号码、会议名称等作为命名实体。又如在本文研究中，将以电子产品价格、上市时间、重量、尺寸等各种参数作为命名实体进行识别。相对于稍显成熟的通用

收稿日期：2009 年 8 月 14 日

作者简介：**邹涛**，女，1985 年生，西安电子科技大学经济管理学院情报学专业毕业，硕士，主要研究领域：计算机情报检索。E-mail：zoutao2007@sina.com。

领域命名实体识别来说,具体领域的研究较少。

目前,命名实体识别方法主要有两种:基于规则和基于统计的方法[1]。基于规则的方法是在识别过程中,利用语言学家制定的规则和待测字符串相匹配。在基于规则的方法中,识别效果在很大程度上取决于规则的完备性和合理性。规则的制定需要专门的语言学家参与。所以规则知识的获取成为基于规则的方法的瓶颈。但优点是一旦生成规则,效率高,识别效果好。基于统计的方法利用经过人工标注的语料进行训练,语料加工时不需要语言学家参与,在经过标注好的语料训练之后,就可以获得较高的识别效果。而且在移植领域时,只需要做很少的改动。但缺点是需要大量的人工标注。基于统计的方法是目前的主流。基于统计的方法主要有隐马尔可夫模型、最大熵模型、条件随机场模型等[2]。

本文背景是基于笔者参与的项目——为政府和大众提供国内外信息技术和产品发展水平咨询服务的基础数据库。该项目将根据信息技术体系将产品科学分类,围绕各领域,通过抽取能够反映各项技术及产品发展水平的性能指标存入数据库中,再以定性研究和定量分析相结合的方法,建立衡量信息技术与产品水平的评价体系。通过建设该数据库,可以动态跟踪世界电子信息技术的最新发展水平,充分且准确地了解和掌握我国信息技术的发展现状及趋势,综合评价这些数据可以为我国的电子产品领域竞争情报分析工作提供强有力的依据。

如今信息的来源广泛,如来自 Email、互联网新闻、企业宣传手册、网站数据库等。但是这些信息是杂乱无章的,需要经过深层次定量的分析,才能增值,否则只是些经过简单整理的沉重的资料。而定量分析第一步需要的便是基础数据,该项目的首要工作便是希望通过计算机智能方法识别并抽取出电子产品领域的产品品牌、属性等命名实体,为下一步工作打下基础。本文就此提出并实现了一种利用层叠模型将基于规则和基于统计相结合的方法,进行高效的电子产品领域命名实体的识别,实验取得了良好的效果。

从通用性上来看,在电子产品领域目前还没有任何的实体识别的研究,本文的研究为进一步的情报分析工作做出了贡献,比如为今后的电子产品领域问答系统等上层应用提供支持。

2 基于规则的命名实体识别

在命名实体识别研究的开始阶段,基于规则的方法占主导地位。在 MUC 命名实体评测时,参加评测的系统几乎都是基于规则的系统。基于规则的命名实体识别是通过分析命名实体的内部和外部特征,人工构造规则模板、添加专名词典来实现的。通常这个过程需要反复调试、不断优化才能达到好的识别效果,但一旦调优之后,识别效果相对统计方法要好很多。本文所用到的基于规则识别的部分将通过调用 GATE API 来实现。

GATE 项目[3]开始于 1995 年英国的谢菲尔德大学,其全称是 General Architecture for Text Engineering(文本工程通用框架)。比较理性地讲,GATE 是一个用来开发和部署用于处理人类语言的软件组件的框架。GATE 经过十几年的发展和不断完善,不仅仅是应用在研究领域,还成功地应用在很多商业领域中。在国外,很多机构都使用 GATE 作为项目基础框架部分的命名实体的识别工具。GATE 本身提供了一个英文语言分析处理组件,叫做 ANNIE。ANNIE 提供了英文分词、英文词表查询、英文分句、英文词性标注、英文抽取规则定义、英文命名实体识别和英文共指消解的功能,从而实现了英文的信息抽取,并且识别准确率等达到了非常高的水平。在中文方面其识别结果却并不让人满意,主要原因是多方面的,如中文结构的特点,词表不够专业,JAPE 规则对中文不能有效支持等。

针对这三个方面,本文系统进行了如下改进:

(1)利用中国科学院计算技术研究所在多年研究工作积累的基础上开发的汉语词法分析系统 ICTCLAS[4]来替换 GATE 中的中文分词模块;

(2)建立比较全面的针对电子产品领域通用命名实体的词表;

(3)针对领域内通用实体编写 JAPE 识别规则,提高命名实体识别的准确率。

3 基于统计的命名实体识别

基于统计的方法主要有隐马尔可夫模型、最大熵模型、条件随机场模型等[5]。

2001 年,John Lafferty 提出了一个基于统计的序列标记和分割数据的方法——条件随机场(CRFs)[6]。它的思想主要来源于最大熵模型,是一种用于标注和切分有序数据的条件概率模型。它集合了最大熵模型和 HMM 模型的优点,不仅能够综合利用包括字、词、词性在内的上下文信息,还能综合利用外部系统特征(External Feature),理论上在避

免碎片化的同时可以集成任意知识源,不管这些知识是相关的还是无关的,类似的或者迥异的。同时该模型对于长距离依赖(Long Distance Dependency)具有很好的描述能力。集合了最大熵模型和HMM模型的优点,并且避免了这些模型本身存在的一些缺点,可以有效地用于序列化标注及切分问题,在自然语言处理的一些领域,如英文词性标注、英文名词短语识别等领域均取得了比较好的效果。它在很多自然语言处理任务中有着广泛而成功的应用,如浅层句法分析、中文分词、命名实体识别等。

对于中文命名实体识别来说,可以基于两种不同粒度大小的语法元素,即基于字(Character-based)和基于词(Word-based)的两种模型。基于词的模型,可以携带有助于命名实体边界检测的分词信息,但也会引入分词的错误,同时造成条件随机场等统计模型的数据稀疏问题。而基于字的模型,则可以避免分词错误引发的问题,但也损失了一些文本特征信息。根据现有的研究比较,本文采用基于词的模型。命名实体识别实际上就是对未知序列进行类别标注,因此首先需要确定序列标注类别集。在实际的操作中,采取了BMEO的标注方式。所谓BMEO方式,是指把一个输入单元标注为实体开始B(Begin,开始)、实体中部M(Middle,内部)、实体尾部E(End,结束)和O(Other,其他)之一。参照现有的命名实体识别方面的研究成果[7],我们选择了N元特征作为条件随机场模型中的基本特征,并把上下文窗口大小设定为5个字或词。以这些基本特征训练出的条件随机场模型,通常可以达到相当不错的性能。

本模型的训练采用了网上开源的工具包条件随机场CRF++0.49。使用开源的工具包,大大方便了本文系统的开发和实验设计工作,缩短了开发周期,使我们可以把精力放在命名实体识别本身的工作上。

4 基于GATE和CRFs相结合的电子产品领域命名实体识别系统

4.1 电子产品领域命名实体的识别策略

针对电子产品领域的特点,在大范围内定义一些通用的属性参数作为命名实体,如价格、产地、尺寸、颜色等。这些命名实体是任何电子产品(如笔记本、手机、电冰箱等)都具有的。由于在写法上都有一定的特点,并且将来也不会发生改变,因此可以通过制定识别规则和编写相关代码,使用GATE开源框架来识别。这样做可以克服只使用统计方法来识别的弱点。使用统计方法进行命名实体识别对标注语料库的依赖性非常大,通常需要利用大规模的标注语料库来克服数据稀疏问题以获得比较好的性能,并且在领域移植时还需要重新标注新领域的语料库。而本文策略可以大量减少细分领域特殊命名实体的语料库标注工作。由于标注类型的减少必然大大减少训练文本的训练时间,提高了效率。另外,使用现有的GATE开源框架只需关注其识别规则,可以完全忽略其他细节问题。

针对电子产品细分领域定义的区别于其他细分领域的命名实体,再使用统计方法进行识别。这些命名实体有两个特性:第一是每个细分领域定义的命名实体类型不同,如手机领域的手机品牌、手机操作系统等;而对于笔记本领域,可能需要识别的就是显卡类型等。第二是这些命名实体的写法是不断变化的,比如手机的品牌可能不断推陈出新,笔记本的显卡类型可能不断有新产品。使用统计方法可以弥补单纯使用基于规则方法的缺陷,可以利用训练文本上下文,上下文词性等信息识别出从未出现过的品牌等命名实体。并且避免了切换细分领域时也需重新标注通用命名实体等类型,极大地减少了工作量。

如果仅是这样简单的组合,在使用统计方法识别细分领域命名实体时会出现一些缺陷。由于使用CRF++工具包时,使用的是基于词的模型,很多时候由于分词错误,或者训练文本过分稀疏,导致命名实体识别的遗漏。而对于电子产品细分领域的命名实体来说,利用规则识别出的诸如产品重量、上市时间等命名实体的标注对后续需要识别的细分领域命名实体的识别是有很大帮助的。比如手机品牌或者手机型号常出现在手机重量之前的几个词内,但是如果由于分词错误导致测试文本中表示手机重量的那几个词的分词与训练文本中大多数的分词不一致,从而导致词性标注的不一致,将会出现手机品牌等无法识别的情况。如果采用本文的层叠模型,在使用规则识别后,将送入统计识别模块的训练及测试文本经过分词、词性标注后,都先利用规则识别模块识别出基本的命名实体,再将识别出的命名实体的组成词进行合并,用新的方式进行标注,这样将会在一定程度上改善识别效果。

因此,本文创新采用的层叠模型加入规则模块,首先识别出一部分命名实体,这将会减少一部分统

计识别的遗漏。

4.2 系统实现

本文通过开发并调用开源软件 GATE API 和在细分领域中使用条件随机场模型工具包,并运用层叠模型有效地结合两种方法,在电子产品领域实现定义的命名实体的识别。其系统架构如图 1 所示。

首先,和英文及其他欧洲语言不同,中文的词与词之间没有空格分开,所以对许多中文处理任务来说,分词是第一步。中国科学院计算技术研究所研制出的汉语词法分析系统 ICTCLAS,主要功能包括中文分词、词性标注、新词识别等。分词后的结果可以作为 GATE 规则识别接口的输入,以克服 GATE 中文分词模块的缺陷。

第二步,在 GATE 框架下编写电子产品领域通用命名实体识别的规则,以及增加规则相关的语料文本。具体添加了识别如产品价格(money .jape)、产品重量(weight .jape)、颜色(bodycolor .jape)、上市时间(date .jape)等规则,money_prefix .lst、year .lst 等实体、实体前后缀或实体前后经常出现的动词等语料文本。

以价格为例(money .jape),根据语义分析定义如下几种情况:

(1)价格前经常出现的动词(有或无)+数字+价格单位,如 3000 元;

(2)价格前经常出现的动词+数字,如定价 5000 元;

(3)价格前经常出现的动词+别的词(数量为 3 个以内)+数字,如定价 约为 5000;

(4)数字+别的词(两个以内)+价格单位,如 5000 多美元。

还有 4 种对应于上面的大写数字的规则,如三千美元。制定的规则举例如下:

```
Rule: NumberTokensMoneySuffix
//数字+别的词(两个以内)+价格单位,如5000多美元
(({Token.kind==number})  //识别的词的类别为数字
({Token.kind==punctuation} //标点符号
{Token.kind==number})?
):tag
(({Token})?    //出现一个词或不出现
 ({Token})?
 {Lookup.majoType == money_suffix} //词的查找类型在定义的money_suffix.lst列表中
)-->
:tag.Money={rule=NumberTokensMoneySuffix} //标注这样的命名实体类型为Money。
```

第三步,将需要识别的电子产品细分领域(如手机领域)的训练文本也通过 ICTCLAS 分词。标注好后,进入规则识别模块,识别出通用领域的命名实体。然后合并这些词,并改写这些实体的词性列的标注。流程如表 1 所示。

随后,选取基于词的 CRF 特征模板进行训练,将训练模型存放在系统中。

第四步,进入统计识别模块,将分词后的文本,同样经过上述的规则模块修正后,使用训练模型进行识别。初步识别结果如下:

诺基亚 /nz CM (CM 代表的意义是公司类型命名实体的中部)

推出 /d O

滑 /nx NB (NB 代表的意义是手机类型命名

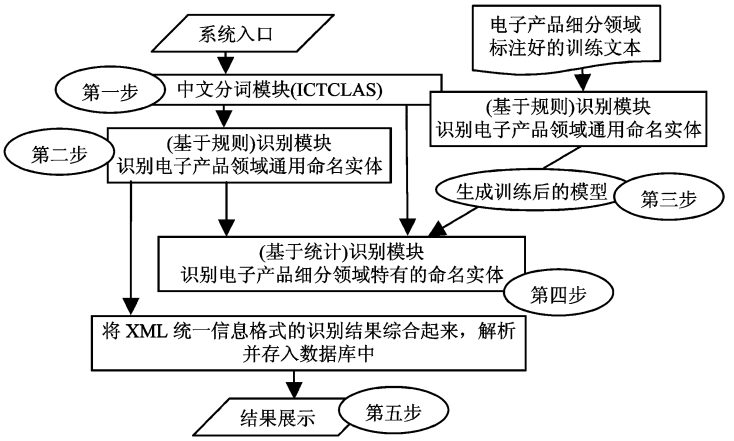

图 1 系统架构

实体的开始）

盖 /nx NM （NM代表的意义是手机类型命名实体的中部）

手机 /n O

……

这种格式的文本处理后同样以XML格式存放。

表1 标注流程

分词后的文本	标注后的文本	规则识别后修正的标注文本
三星/nz	三星/nz BS	三星/nz BS
GT-i7500/nx	GT-i7500 /nx TS	GT-i7500/nx TS
尺寸/n	尺寸/n O	尺寸/n O
为/p	为/p O	为/p O
115/m	115/m O	115×56×119mm/BE O
×/w	×/w O	……
56/m	56/m O	
×/w	×/w O	
11/m	11/m O	
9mm/nx	9mm/nx O	
……	……	

第五步，对两种方式识别的结果（XML识别后文本）进行综合，并对其解析，将结果存入数据库中，并在系统中以可视化、人性化的方式展现。

5 系统输出

我们选择手机领域作为电子产品细分领域。因为现在没有公共标准的电子产品领域语料库，所以本实验使用爬虫软件在IT新闻类网站爬取大量的文本，选取其中80%作为统计识别模块的训练文本，将训练模型输入系统，剩余的20%作为测试文本输入。最终识别的命名实体类型主要有手机品牌、型号、手机外观（直板、滑盖、旋屏等）、机身大小等20类。评价指标选用准确率（P）、召回率（R）和F值。具体定义如下：

P=系统标注正确的NE总数/系统标出的NE总数；

R=系统标注正确的NE总数/测试集中出现的NE总数；

$F=2PR/(P+R)$。

实验设计两种形式：第一种，使用普通的规则和统计模块的串行结构，即在训练模型和进入统计模块前没有使用规则模块的识别结果修正文本标注；第二种，使用叠加模型加入规则识别模块。对比这两种实验结果，验证本文所采取的层叠模型结合两种方法所带来的识别效果的提升。

实验结果对比如表2所示。

表2 实验结果对比

	普通串行结构	层叠模型有效结合
准确率（P）	96.045	97.159
召回率（R）	76.749	77.201
F值	85.319	86.038

从实验结果来看，由于本文进行了GATE分词模块的替换、识别规则的优化以及CRF ++工具包的有效特征模板的选定等工作，命名实体识别的准确率和召回率都有所提高，满足了实际项目的需求。

其次，由于层叠模型的结合只能改善特定的由于分词等错误引起的识别错误和遗漏，而本实验采用的训练以及测试文本数量有限，因此出现层叠模型能改善的语料情况不多，从结果数据上看，提升的效果并不十分明显。但通过上述3个结果的对比，反映出这两种方法经过使用本文设计的层叠模型，能在一定程度上改善识别效果，并且随着训练文本和测试文本的增加，效果会更加明显。

6 结论和展望

本文以作者实习单位的项目为背景，采用层叠模型将规则方法和统计方法相结合对电子产品领域命名实体识别进行了高效的实现。系统中替换了GATE开源框架的中文分词模块，添加了对电子产品领域命名实体识别的规则和规则库语料，并通过实验比较选取了CRF++工具包识别效果最优的特征模板和标注方式，通过层叠模型将GATE开源框架和CRF++工具包集成到自己开发的系统中，进一步提高了命名实体识别的效果，最终展现了标准统一的操作界面和流程，并为今后的扩展应用提供了方便的接口，基本满足了项目的初步需求。

总结中发现，实验中有个别不经常出现的实体类型因为规则编写不到位而无法识别，因此下一步将继续优化识别规则。在统计识别模块的训练文本标注工作中，虽然工作量已经减少很多，但还是耗时耗力，因此下一步将围绕利用机器标注大多数语料而展开工作。

参 考 文 献

[1] 季妲，罗振声．基于统计和规则的中文姓名自动辨识[J]．语言文字应用，2001 (1)：14-18．

[2] Fu G H，Luke K K．Chinese named entity recognition using lexicalized HMMs[J]．ACM SIGKDD Explorations，2005，7 (1)：19-25．

[3] The University of Sheffield．Developing Language Processing Components with GATE Version 5 (a User Guide)[OL]．[2008-08-01]．http://gate.ac.uk/sale/tao/index.html．

[4] 张华平，刘群．ictclas4j 中文分词系统[OL]．[2008-09-01]．http://code.google.com/p/ictclas4j/

[5] 张晓艳，王挺，陈火旺．命名实体识别研究[J]．计算机科学，2005(32)：44-48．

[6] Burr Settles．Biomedical named entity recognition using conditional random fields and rich feature sets[C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications．Switzerland，Geneva，2004：104-107．

[7] 向晓雯．基于条件随机场的中文命名实体识别[D]．厦门：厦门大学，2006．

（责任编辑　许增棋）