DOI:10.13366/j.dik.2012.03.002

专题·细粒度信息检索与知识挖掘

Special Research on Information Retrieval and Knowledge Mining with Fine Granularity

编者按：随着用户信息需求的日益精确化，信息检索与挖掘研究呈现细粒度和语义（关联）化的发展趋势。如何实现细粒度信息的抽取、组织，进而实现有效的检索与挖掘，是值得深入研究的问题。2010年，武汉大学启动了“70后”学者学术发展计划。以此为契机，我与吴丹、刘萍、曾子明、安璐、吴佳鑫、邓胜利等7位青年教师在原信息检索研究小组的基础上，组建“信息检索方法与技术团队”，在学院的大力支持下，成功入选武汉大学首批“70后”学者学术团队。团队成立后，我们以“跨语言环境下的细粒度信息检索与可视化知识挖掘”为切入点，开展协作攻关研究。本组专题则是团队成员在此基础上的部分研究成果。

本专题包括4篇论文。陆伟、鞠源等的《产品命名实体特征选择与识别研究》，采用条件随机场（CRF）模型，对实体特征的选择和识别效果进行了探索；吴丹、何大庆等的《跨语言信息检索中的命名实体识别与翻译》，提出基于信息抽取的命名实体识别与翻译方法，并基于此实现了跨语言信息检索。这两篇论文侧重于细粒度信息的抽取识别与检索应用研究。刘萍、高慧琴等的《基于形式概念分析的情报学领域本体构建》，引入形式概念分析的理论方法，以情报学领域为例，探讨如何将形式概念分析应用于本体构建中；安璐、余传明的《基于自组织映射的期刊主题专业化与综合性分析》，采用自组织映射技术，探索了图书情报领域的主题专业化与综合性特征。这两篇论文主要侧重于关联分析与知识挖掘。

陆　伟
武汉大学信息管理学院

产品命名实体特征选择与识别研究

Research on Product Named Entity Feature Selection and Recognition

陆　伟　鞠　源　张晓娟　吴　丹
（武汉大学信息资源研究中心，武汉，430072）

［摘要］　随着互联网经济的飞速发展，信息抽取领域的产品命名实体识别在商务智能领域有着广泛的应用。本文采用条件随机场（CRF）模型，选取词汇、词法和词形上一系列的特征进行训练，通过交叉验证对识别效果进行评价，并通过识别效果指导特征的选取。实验中比较了两种标注方式（BRAND/TYPE和PROD），并取得了令人满意的识别效果。在与最大熵模型对比中，验证了CRF模型对于产品实体识别的优越性。

［关键词］　产品命名实体识别　CRF模型　交叉验证　最大熵模型
［中图分类号］　G350　［文献标识码］　A　［文章编号］　1003-2797（2012）03-0004-09

［Abstract］　With the booming of online business, the recognition of product entity has been widely applied in Business Intel-

［基金项目］　本文系武汉大学“70后”学者学术发展计划项目“跨语言环境下的细粒度信息检索与可视化研究”及国家自科基金项目“基于语言模型的通用实体检索建模及框架实现研究”（71173164/G031401）的成果之一。

［作者简介］　陆伟，男，博士，教授；鞠源，男，硕士，硕士研究生；张晓娟，女，硕士，博士研究生；吴丹，女，博士，副教授。

ligence. In our paper, we use the CRF model and select a series of lexical, syntactic and semantics features as feature space. In addition we use cross evaluation methods to evaluate the classifier's performance and also guide previous feature selection. In corpus construction step, we adopt two different labeling strategies, which we analyze in the evaluation section. The Experiment has achieved satisfactory result. By comparison with max entropy model, we further demonstrate the efficiency of CRF model we use in this experiment.

[Key words] Product entity recognition CRF model Cross validation Max entropy

1 引言

命名实体识别(识别文本中的人名、组织机构名、地名、时间等)是自然语言处理领域的重要任务,已有研究表明,命名实体的识别能够改进信息检索的效果(尤其体现在未登录词的识别上)。由于事件通常由人物、时间、地点等要素构成,因此命名实体识别也是事件抽取的重要内容。在实体识别的基础上分析实体的语义关系,又是本体构建及自动文摘生成的基础。

关于命名实体识别的研究,国外开始的比较早,Rau[1]早在1991年就公司名称的识别和抽取进行了研究;Bikel[2]等最早利用隐马尔科夫模型对英文地名、人名和机构名实体进行识别;Borthwick[3]提出基于最大熵模型针对英文和日文的命名实体识别方法;Ratinov[4]等使用未标注文本训练词类模型,有效的提高了命名实体的识别效率。在国内,中文命名实体识别也获得较为广泛的关注。中科院计算所张华平、刘群[5]等人提出基于层叠隐马尔科夫模型的中文命名实体识别,采用该方法设计的ICTCLAS系统在SIGHAN汉语分词竞赛中取得第一名的成绩;南京大学周俊生等人[6]利用层叠条件随机场模型对中文机构名进行自动识别,在测试中获得优于其他识别算法的效果和性能。

在命名实体中,产品命名实体是比较新的一类实体,相对于其他实体有其特殊性,主要体现在命名方式灵活且缺乏特定的线索词(如区,先生,公司等)来做表征。产品命名实体的有效识别在商务智能领域有着重要的研究意义。目前国内外相关研究还不是很多,比较典型的有:赵军等人在文献[7]中将产品实体分为品牌名、型号名和产品名三种类型,利用层级隐马尔科夫模型对中文文本中产品名进行识别,但并未将HHMM与其他模型相对比,标注策略也较单一;北京大学的Wenyuan Yu[8]等利用知识库的方法,先识别出与产品名相关的实体和属性,再对各相关实体的语义角色予以识别;昆明理工大学张朝胜[9]等人利用条件随机场模型识别英文产品命名实体,但没有对特征的选择进行验证,也没有和其他模型进行比较;Nichalin Suakkaphong[10]的等人也采用了CRF模型,但研究重点在于利用自举(bootstrapping)的方法进行半监督机器学习,实现在少量标注的情况下对疾病命名实体进行准确识别。本文采用了比较常用的CRF模型,通过对识别效果的评价反过来指导特征选择,并针对前面的研究中单一的标注策略,采用两种标注方式进行标注,并对他们的识别效果进行比较。

TREC(文本检索测评会议)自2009年起在传统命名实体(人名、地名、机构名等)的基础上,新增产品实体检索的查找。本文即在该任务的背景下,采用CRF模型,对网页中的产品命名实体进行识别,并与最大熵模型进行对比。

2 CRF模型概述

CRF模型即条件随机场模型,是一种基于统计的序列标记识别和分割的无向图模型,它可以在给定需要标记的观察序列的条件下,计算整个标记序列的联合概率分析,而不是在给定当前状态条件下,定义下一个状态的状态分布。CRF模型是对隐马尔科夫模型和最大熵马尔可夫模型的改进,没有HMM产生式模型那样严格的独立性假设,因为可以容纳上下文信息,特征选择上比较灵活。同时CRF模型计算全局最优输出节点的条件概率,对所有特征进行全局归一化,得到全局最优解。它既保留了最大熵马尔可夫模

产品命名实体特征选择与识别研究

Product Named Entity Feature Selection and Recognition Research

陆 伟 鞠 源 张晓娟 吴 丹

型等条件概率框架的优点，同时又解决了(Label-bias)标记偏置的问题。

因此，CRF 模型非常适用于序列的标注，被广泛应用与分词、词性标注领域。命名实体识别实际上是序列标注的问题，所以本文采用 CRF 模型进行产品名实体的识别。对于给定词序列 $x=(x_1,x_2,\cdots,x_n)$和标注序列 $y=(y_1,y_2,\cdots,y_n)$，定义一个条件随机场模型如下：

$$p(y \mid x) = \frac{1}{Z_{(x)}}\exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x)\right)$$

其中，$Z_{(x)}$是归一化因子；表示给定词序列的长度；$f_k(y_{i-1}, y_i, x)$是特征函数，λ_k 是第 k 个特征函数的权重系数。CRF 模型能够通过训练规模较小的语料库，获得比较理想的准确率。但算法复杂度比较高，训练代价比较大。

CRF 模型最早是由卡耐基梅隆大学的 John Lafferty 等人在文献[11]中提出的，Lafferty 作为第一作者为 CRFs 写了 C++的工具包，第二作者 McCallum 实现 java 的 CRF 算法（在 mallet package 中）。另外在 SourceForge 上有 IIT Bombay 的 Sarawagi 贡献的 java 版 CRF 工具包。考虑到效率和识别效果等因素，本文最终选用 C++实现的 CRF++工具包。

3 实验流程与方法

实验过程如图 1 所示，包括①数据采集：选取国外一些正规的电子商务的网站，如微软的 Bing Shopping，PriceGrabber.Com，爬取有关产品介绍和用户评论的信息。②语料库标注：采用两种不同的方式对语料库进行标注。③特征的选取：实验中选取词本身、字形、词性等一系列特征，并通过识别效果评价对特征的选取进行验证。④基于一定的机器学习算法进行实体识别：定义特征模板，利用条件随机场模型，对产品命名实体进行标注。⑤效果评价：通过计算查全、查准率和 F 值评测识别效果，对特征的选取进行验证，并对两种标注方式进行分析。⑥对比和改进：用最大熵模型对产品实体进行识别，评测它的效率并将其与本实验的识别效果进行对比。

3.1 语料库构建

语料库的构建是机器学习过程中最为基础和重要的部分，产品语料的质量对产品实体识别的效果有很大影响。本语料库的构建一要保证语料库内容的规范性，必须是从正规网页中抓取的相关产品信息；同时也要保证标注形式的统一性，产品语料库的标注形式和策略必须事先商定好，并进行多轮标注，尽量减少人为主观因素的影响。

图 1 产品命名实体识别流程

3.1.1 数据采集

笔者爬取 Bing Shopping 和 Price Grabber 网站上关于 Cell phone，Car Accessories，Camera，TV，Audio Electronic，Computer，handbags，Toys，Watches 等九大产品类目下的相关产品的描述信息和评论（reviews）。存储为如下的 xml 格式：

```
<?xml version="1.0" encoding="UTF-8" ?>
- <corpus>
  - <record id="1">
      <category>electronics</category>
      <entry>LG 42LG70 - 42 LCD TV</entry>
      <abstext>The LG70 series are the ideal HDTVs for home theater enthusiastsIf they prove one
        thing, it's that nobody demands from LG products than LG itself</abstext>
    </record>
  - <record id="2">
      <category>electronics</category>
      <entry>Garmin n vi 1390T - GPS receiver</entry>
      <abstext>nuvi 1390T packs big features into a slim designThe ultra-thin navigator includes lane
        assist with junction view, free traffic alerts, hands-free calling and ecoRoute to.</abstext>
```

图 2 产品记录存储格式

每个产品记录由 category（种类）、entry（产品条目）、abstext（产品描述文本）三个字段构成，共抓取 8000 条记录。构成语料库的每个语句中都要涉及产品名，这样才有足够的区分度将产品名和其他词语区分开来。

Bing shopping 上采集的记录大多属于电子商务

产品，其命名规则大致符合一般的产品规范。而针对TREC中还要识别的医药类产品，后期将从Wikipedia中搜集一些医药产品的描述，对数据集中进行扩展。

3.1.2 语料库标注

关于产品实体的标注方式，张朝胜等人[9]采用PROD-O的形式，即是产品实体的统一标注为PROD，不是产品的标为O；刘非凡、赵军等人[7]在研究中将产品实体细分为品牌实体BRAND和型号实体TYPE进行标注；Nichalin Suakkaphong[10]等采用BIO（B-begin标注实体的开始部分，I-in在实体内部的部分，O-非实体）对Disease Named Entity进行标注，均获得较高的识别效果。本文同时采用PROD-O和BRAND/TYPE-O两种不同的形式对语料库进行标注，并在实验效果评价中对两种形式的识别结果进行分析。

为保证标注的统一和规范，减少人为主观因素的影响，本实验对语料库进行两轮标注。

（1）第一轮标注。明确标注策略，产品名的界限是一个需要界定清楚的问题，本文在对语料标注时采用了如下规范：①对代表产品类型的单词（如LCD，GPS，MP3等），统一不标注为产品名。②对于一些存在歧义的词，如Apple等，参照Hints（即上面xml文件中的entry字段）来标注；对于同一单词，代表产品名还是公司名（如Microsoft，Motorola等）根据上下文来区分。③在一些product review中，前面若提到Nokia2330这一产品，后面则可能会直接用2330来代替该产品，这里我们会将2330也标注为产品名。

设计在线标注平台CrfMarker，组织多人去进行标注。CrfMarker将产品描述文本拆成一个个词，让用户去标注属于产品名的单词。每条记录描述文本（abstext）上面都会显示该产品的条目（entry）字段，作为标注提示（hints）。如图3所示：

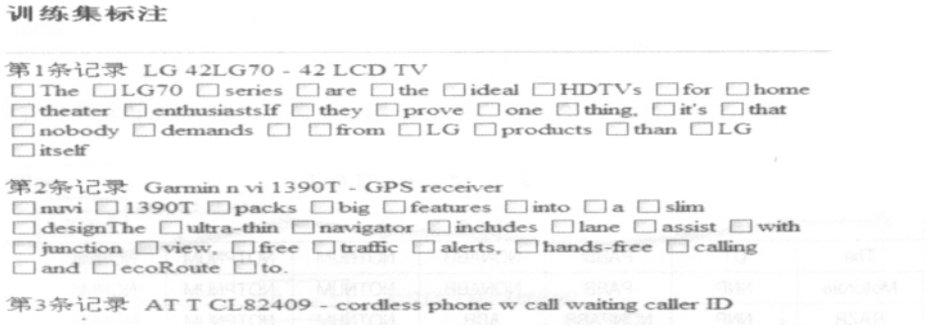

图3 在线标注平台

在具体标注过程中，将标注人员随机分为7组，每组标注2～3类产品的实体。这样可以保证每类产品实体都会被均匀标注到；对于用户的标注，后台会记录下来的，我们取每个词被标注的频率最高的标签（Label）作为最终标注结果。如LG70有3人标注为BRAND，2人标注为TYPE，1人标注为O，则赋予PROD标签。这样可以在最大程度上减少标注过程中的主观误差。

（2）第二轮标注。为克服“大众的偏见”，第二轮标注由笔者等进行集中核查。在此轮中并不直接修改标注结果，而是在第一轮标注结果的后面加上修改，删除等Operator符号，然后和大家讨论确定最终的标注结果。其界面如图4所示：

图4 第二轮标注界面

3.2 特征选择及特征模板的确定

3.2.1 特征选择

特征选择，是机器学习中一个重要的步骤，产品实体特征的选取首先要对英文产品命名实体的特点进行分析，考虑产品实体的特殊性，选取出的特征要有较大的区分度。特征及其组合对于分类和识别的效果有直接影响。特征选择常用的方法有信息增益和开方检验等。

在一般命名实体识别的研究中，所选取的特征往往可以归纳为词汇特征、词法、句法特征和语义特征这三大类[10]。

(1)词汇特征。即词形上的特征，包括大小写、数字、标点符号、前后缀等单词形态上的特征。很多产品实体名称中含有以上一个或几个特征，比如 Sony Errsion(首字母大写)、Nokia N97(大写，含数字)、BMW(全部大写)、P&G(标点符号特征)，这些均可作为产品实体识别的重要特征。

(2)词法、句法特征。词法、句法特征会考虑词在所在句子结构中的成份和上下文信息，比较灵活。一般在 NER 的识别中通常都会用到词性 POS 特征。不过词性特征很灵活，词性标注的过程本身也会产生不确定性，不容易掌控。对识别效率是否改进，在后面的实验效果评测中会分析。

(3)语法特征。考虑词的意义以及所属的既定语义类别，定义这类特征往往需要预先构建好的字典或相应的本体。由于在产品命名实体领域还没有相关的产品列表和本体，需要手工构建而考虑人力有限，所以并没有采用这一类的特征。

综合上述各种特征和前人研究中应用的方法，本文定义如下 7 维的特征，即：①词本身的特征 Word；②词性特征 POS；③含有大写字母特征，是则标注 PABB，否则 NONPABB；④缩略词特征，是缩写则标注为 ABB，否则 NONABB；⑤是否是纯数字，是则标为 NUM，否则 NOTNUM；⑥是否含有数字 PNUM，否则标为 NOTPNUM；⑦是否含有连接符，如“－”，“&”等，是则标记为 MK，否则为 NONMK。

这些特征均是由计算机自动生成的，其中词性特征是由斯坦福词性标注工具 Stanford pos tagger[12] 生成，其它均通过自己编写的程序进行自动判断生成。结合前面人工的标注，语料库标注后的格式如下表 1。

表 1 语料库标注格式

特征维							标注
The	DT	PABB	NONABB	NOTNUM	NOTPNUM	NONMK	O
Motorola	NNP	PABB	NONABB	NOTNUM	NOTPNUM	NONMK	BRAND
RAZR	NNP	NONPABB	ABB	NOTNUM	NOTPNUM	NONMK	TYPE
V3i	NNP	PABB	NONABB	NOTNUM	PNUM	NONMK	TYPE
is	VBZ	NONPABB	NONABB	NOTNUM	NOTPNUM	NONMK	O
fully	RB	NONPABB	NONABB	NOTNUM	NOTPNUM	NONMK	O
loaded	VBN	NONPABB	NONABB	NOTNUM	NOTPNUM	NONMK	O

共有 8 列，0～7 列分别对应上述 7 个特征，列 1 为词性特征，其他为词形方面的特征，最后一列是赋予的标注。

3.2.2 确定特征模板

CRF＋＋工具包为机器学习过程中的模型(model)训练提供了特征模板。通过控制特征模板的内容，可以灵活选取特征，也可以对特征进行自由的组合。使模型训练时不仅考虑上下文因素，还可以综合考虑多种特征组合。

特征模板格式[13]：U02：%x[0,0] 其中 U02 是特征项编号，%x[i,j]可以精确到第 i 行的第 j 列特征。若将表 1 中的第一行作为当前行，那么%x[0,0]则表示 The 这个单词本身，%x[0,1]则表示当前行的词性特征，%x[0,0]/%x[0,1] 表示综合词和词性的复合特征。

特征模板的确定需要进行很多次实验，对比每个特征项的添加以及各种特征组合对于实验结果的影响，选取使识别效果达到最佳的模板。实验表明当滑

动窗口大小设置为 2 时，即考虑每一维特征上的上、下两个位置范围内的取值情况，再对不同特征项进行复合时识别效果最佳。

3.3 实验测评与分析

实验中对产品命名实体识别的效果测评通过计算实体识别的 P 值（准确率 Precision Rate）、R 值（召回率 recall rate）和 F 值（F value）来衡量的。其中，其计算公式分别为公式（1）、（2）、（3）：

$$P = \frac{truePos}{totalmarkedPos} \tag{1}$$

$$R = \frac{truePos}{totalPos} \tag{2}$$

$$F = \frac{2 * prec * recall}{prec + recall} \tag{3}$$

其中 truePos 是被标注正确的命名实体个数，totalmarkedPos 是实验模型识别出来的产品实体个数，totalPos 是测试集中标注的实体总数。F 值是对 P 值和 R 值的综合评价指标。实验中，分别对不同的实体标签计算他们的 P 值、R 值和 F 值，从而得出一系列评价指标。

3.3.1 两种标注方式的效果评价

对于 PROD /O 和 BRAND/TYPE/O 这两种标注方式，实验中采用交叉验证的方法对其识别效果进行评价，从语料库中抽取 1200 多个标注好的句子，随机按（7∶3）的比例分成两份，70% 那份做训练集训练 CRF 模型，另一份做测试集，计算 P、R、F 的值。这样重复 N 次，并保证每次的用于训练的集合和用于测试的集合都不相同，以消除误差。

本文取 N＝10，分别计算出每次两种标注方式中各实体标签的 P 值、R 值和 F 值，为了方便进行比较，我们直接取每个实体的 F 值，对于 PROD/O 方式，识别效果可以通过 PROD 标签的 F 值直接体现，即：

$$F_{(PROD|O)} = f_{(PROD)} \tag{4}$$

对于 BRAND/TYPE/O 标注方式，他的识别效果可以通过综合品牌实体和型号实体的识别效果来反映：

$$F_{(BRA/TYPE)} = \alpha * f_{(BRA)} + \beta * f_{(TYPE)} \tag{5}$$

公式（5）中 α，β 代表品牌标签和型号标签重要性的权值，不同的应用目的决定 α 和 β 的值不同，在不能确定的情况下，实验中取 α＝β＝0.5，根据 10 次交叉验证的结果，我们绘制出他们的插值曲线，如图 5 所示。

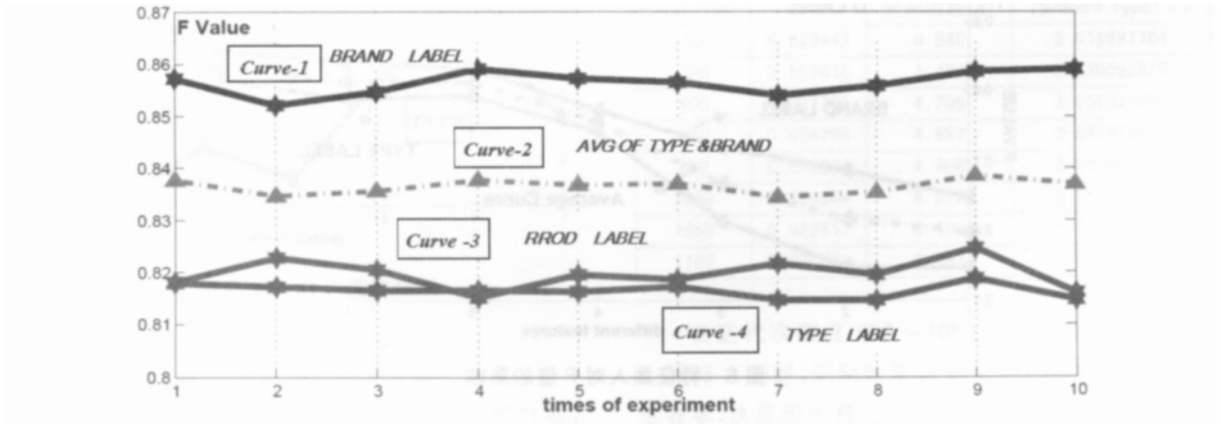

图 5 交叉验证插值曲线

图中曲线 4 代表型号（TYPE）实体的 f 值随实验次数的变化，曲线 1 代表品牌实体类别的 f 值，曲线 3 是产品 PROD 标签的插值线，可以看出 PROD 标签的识别效果略高于 TYPE 标签而远低于 BRAND 标签。为评价两种标注方式，我们取 BRAND 和 TYPE 实体 F 值的平均值，如中间的虚线（Curve 2）所示，明显高于 PROD 标签的标注效果，可见 BRA/TYPE 方式的整体识别效果是优于 PROD/O 方式的。

对 10 次实验结果各指标取平均值，我们可以得出两种标注方式最终的实体识别效果，见表 2。

表 2 PROD/O 与 BRAND/TYPE/O 识别效果比较

标注方式	标签	P 值(%)	R 值(%)	F 值(%)
PROD/O	PROD	89.91	75.26	81.97
	O	98.04	99.33	98.68
BRA/TYPE/O	BRAND	93.08	79.29	85.64
	TYPE	85.65	78.01	81.65
	O	98.54	99.43	98.98

从表中我们可以看出,第一种标注方式中 PROD 标签的 F 值只是略高于第二种标注方式中 TYPE 标签的值(召回率 R 的值还低于 TYPE),但与标签 BRAND 的识别效率相差很大。因此在后续的评价分析中,我们都是用 BRAND/TYPE/O 的标注方式。

3.3.2 特征选择的效果分析

本文选取的特征有词本身 Word 特征、词性特征(POS)、大小写特征(PABB)、缩略词特征(ABB)、纯数字特征(NUM)、含有数字特征(PNUM)和连字符特征(MK)等词形、词法方面的 7 维特征。

对于 7 维特征对于识别效果的影响,主观上的感觉是正面的,并没有通过实验去验证,例如词性特征,它不像其他特征那样相对确定,即便是规范的句子输入后,Stanford pos tagger 的标注结果也存在不确定性。尽管某些特征对于实体识别的益处是显然易见的,但我们能否依赖,还需要通过实验进行验证。

同样采用交叉验证取平均值的方法,每新增一个特征,都对其识别效果进行的测评,以检验所选的特征是否对产品实体的识别有帮助。实验结果如表 3 所示。

表 3 特征选择效果综合测评

ID	Features	Times (s)	BRAND (%)	TYPE (%)	O (%)
1	Word	5.32	64.71	54.59	97.60
2	Word+NUM	9.1	68.63	54.71	97.79
3	Word+NUM+ABB	11.5	71.88	61.33	98.02
4	Word+NUM+ABB+PNUM	12.37	77.39	74.12	98.44
5	Word + NUM + ABB + PNUM+PABB	12.46	83.41	79.22	98.77
6	Word + NUM + ABB + PNUM+PABB+MK	13.66	84.19	79.05	98.84
7	ALL Features(+POS)	16.47	85.91	81.65	98.98

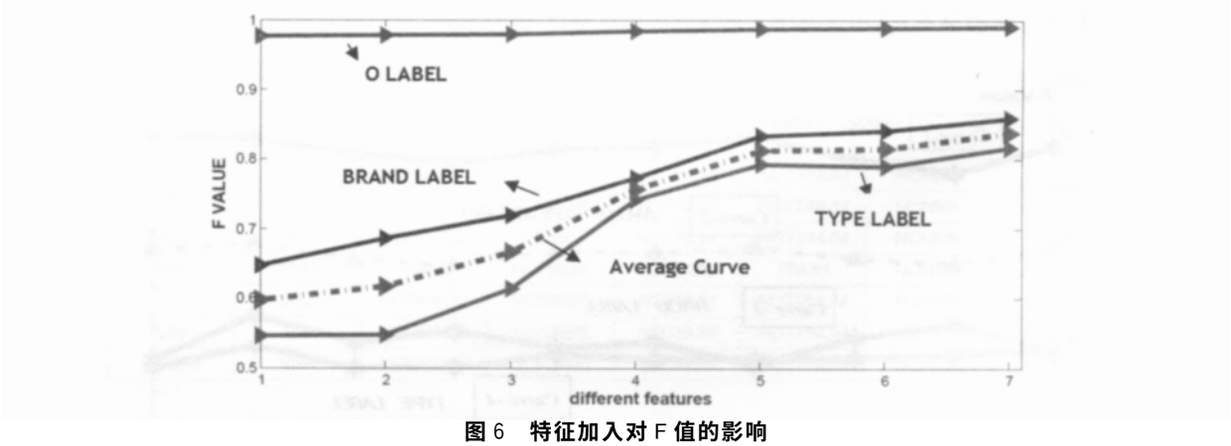

图 6 特征加入对 F 值的影响

表 3 中列出的是在不同特征加入时,BRAND 和 TYPE 实体识别的 F 值的变化,同时统计了训练这些特征的时间开销,可以看出训练代价随着特征的加入呈线性增长。从图 6 中,可以直观的看出当上述特征加入时,对产品实体识别的影响是积极的。进而验证了所选择的特征的正确性。中间的虚线表示 BRA 和 TYPE 标签 F 值的平均值变化,从曲线上升的趋势可以看出横轴上 4,5 对应的特征组合对模型识别效率的影响较大,结合表 3 的特征组合可知 PABB(含有大写字母)和 PNUM(含有数字)这两个特征相对比较重要,而纯数字特征 NUM 和 MK 则起一种辅助作用,最后加入的词性特征也使识别效果得到改善。

图 7、8 分别刻画了有新的特征不断加入时召回率和准确率的变化情况。从图中可知(中间的虚线代

表 BRA/TYPE 综合指标的变化情况），特征的加入对主要有利于提高实体识别的召回率，而准确率呈现出不稳定的状态，受到多种因素的影响。

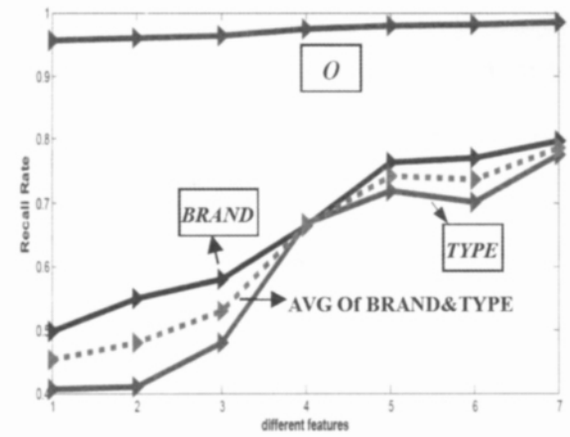

图 7　特征加入对 R 值的影响

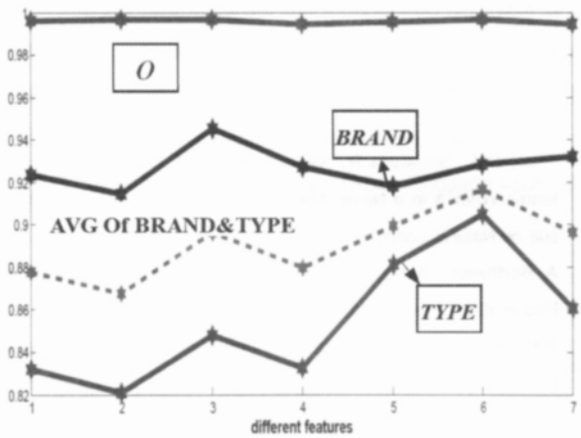

图 8　特征加入对 P 值的影响

4　模型比较

为了验证 CRF 相对于产品命名实体识别的性能，本文将最大熵模型进行产品实体识别的实验，并将识别效果与 CRF 模型的效果进行比较。最大熵模型[14]是传统的分类模型，经常被用于实体识别、语义角色标注和机器翻译等自然语言处理和模式识别相关领域，模型训练时将输入的训练集中每一行（特征及其标注）看成一个事件（Event）。例如 Sunny Sad Humid Outdoor，Sunny Sad Humid 特征看成事件的上下文 context，Outdoor 为事件的结果，每个事件构成一个约束条件（constraints）。目标函数是特征分布的熵的大小，模型训练的过程就是通过 GIS 迭代求出满足约束条件的目标函数值最大时，各个特征对于结果标签的权重 λ。

爱丁堡大学张乐博士[15]写的 C++的 maxent 工具包和 SourceForge 上 opennlp maxent JAVA Project[16]都是较常用的工具。本文同时使用两种工具包进行实验，并对他们的性能进行比较和优化。实验中还是使用上述 7 个维度的特征作为特征空间进行训练。对于 java 的工具包，输入的训练集可保持语料库中的标注格式不变，迭代时默认迭代次数是 100，跳出迭代的阈值默认为 0.00005，即两次 log 似然值的差值小于0.00005 或者不再变化时跳出迭代，经测试要达到默认阈值要迭代 20000 次，而实际上是不需要这么多次，这里存在一种过拟合的问题。所以我们采取不断修改迭代次数的方法来调试识别的效果。调试中为了放大不同参数变化对于识别效果的影响，我们用 $F=2*(f_{BRA}+f_{TYPE})$ 作为综合指标，实验结果如表 4 所示。

表 4　基于 Maxent 产品识别效果

迭代次数	阈值	时间消耗（s）	（Brand+Type）*2
50	6.829443	0.640	3.019891164
200	0.507616	1.482	3.030092879
800	0.037255	4.705	3.036323202
830	0.034796	4.857	3.047839578
850	0.033294	4.966	3.050803358
1000	0.024644	5.817	3.050803358
1050	0.022517	6.109	3.050803358
1100	0.020662	6.197	3.047645327

在迭代次数在 850～1050 次，阈值在 0.0225～0.0333 之间时，识别效果达到最佳，再往后反而会降低效率，这是因为模型训练过于严格，产生过拟合的现象使查全率降低，由于语料库特征中不存在数据稀疏的问题，所以取消了平滑参数，此时 BRAND 和 TYPE 实体的 F 值分别为 84.33%，68.21%，耗时 5.5s。对于 C++的工具包，输入训练的格式只要简单将最后一列标注的实体标签放在最前面，仍通过改变迭代次数方式进行调试，记录下 F 值的变化。如图 9 所示，当迭代次数为 200 时，识别效率达到最大，

BRAND 和 TYPE 的 F 值分别为 84.41% 和 73.07%，平均耗时 1.04s，可见 C++ 的工具包在识别效率和性能上比 java 的要高一些。但是和实验中基于 CRF 模型的识别效率（BRAND：85.91% TYPE：81.65%）相比要低，从而验证了实验模型的优越性。

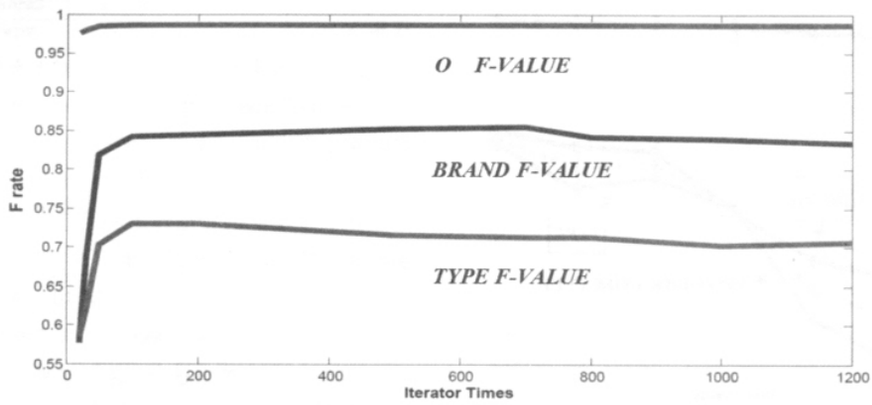

图 9 F 值随迭代次数的变化

5 总结

本文对英文产品语料库采用 PROD/O 和 BRAND/TYPE 两种方式进行标注，选择相应的特征和特征模板，利用条件随即场模型实现了英文产品命名实体的识别，并对识别效果进行评价，对两种标注方式进行分析。通过对识别效果验证特征的选择并不断修正特征模板，最终达到品牌实体 F 值 85.91%，型号实体 F 值 81.65% 的识别效果。最后与最大熵模型的性能效率进行比较，验证了实验模型在产品命名实体识别上的优越性。

针对查全率较低的情况，系统拟采用半监督[17]的机器学习方法进行扩展训练，即将尚未标注的句子分成 N 份，从已经标注好的训练集中抽出部分作为测试集（如取 200 个文本句子），余下部分作为训练集以训练模型，进而用训练好的模型取代人工自动标注 N 份未标注的语料，并将识别概率差异较明显的结果（可自己设定阈值）作为标注好的放入训练集中去继续训练，并用新训练出的模型去识别固定测试集合，如果相对于原来的模型有所改善，则加入训练集进行下一轮迭代，否则舍去，以此来取代人工标注的方式，提高查全率，改善识别效果。

参考文献

1 Rau L F. Extracting Company Names from Text. In：Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications. 1991：29-32

2 Bikel DM，Schwarta R，Weischedel R M. An Algorithm that learns What's in a Name. Machine Learning Journal Special Issue on Natural Language Learning，1999，34(1-3)：211-231

3 A. Borthwick. A Maximum Ent ropy Approach to Named Entity Recognition. New York：NewYork University. 1999.

4 Ratinov L，Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In：Proceedings of the 13th Conference on Computational Natural Language Learning. 2009：147-155

5 俞鸿魁等. 基于层叠隐马尔可夫模型的中文命名实体识别. 通信学报，2006(2)

6 周俊生等. 基于层叠条件随机场模型的中文机构名自动识别. 电子学报，2006(5)

7 刘非凡等. 面向商务信息抽取的产品命名实体识别研究. 中文信息学报，2005(1)

8 Yu W，et al. PKUNEI - A Knowledge - Based Approach for Chinese Product Named Entity Semantic Identification，Computer Processing of Oriental Languages：Language Technology for the Knowledge-based Economy，2009，Springer Berlin / Heidelberg. p.297

（下转第 34 页）

标签	期刊名称	影响因子	标签	期刊名称	影响因子
5	Collection Building	—	35	Library Hi Tech	0.344
6	College & Research Libraries News	—	36	Library Resources & Technical Services	0.698
7	College & Research Libraries	0.781	37	Library Review	—
8	Drug Information Journal	0.504	38	Library Technology Reports	—
9	DTTP, Documents to the People	—	39	Library Trends	0.239
10	EContent	0.271	40	New Library World	—
11	EDUCAUSE Review	—	41	OCLC Systems and Services	—
12	European Journal of Information Systems	1.202	42	Online Information Review	1.103
13	IEEE NETWORK	3.068	43	Online	0.352
14	IEEE Transactions on Engineering Management	1.156	44	Performance Measurement and Metrics	—
15	IEEE Transactions on Professional Communication	0.609	45	Portal:Libraries and the Academy	1.146
16	IEEE Transactions on Software Engineering	3.569	46	Publishing Research Quarterly	—
17	Information Outlook	—	47	Reference & User Services Quarterly	0.339
18	Information Systems Management	1.242	48	Reference Services Review	—
19	Information Technology & People	—	49	School Librarian's Workshop	—
20	Information Technology and Libraries	0.703	50	Social Science Computer Review	0.714
21	Information Visualization	—	51	Social Science Information	0.341
22	International Journal of Human-Computer Interaction	—	52	Telecommunications Policy	1.244
23	Journal of Academic Librarianship	0.667	53	The Electronic Library	0.393
24	Journal of Educational Multimedia and Hypermedia	—	54	The Library Quarterly	0.364
25	Journal of Information Science	1.648	55	The Library	—
26	Journal of Information Systems Education	—	56	The Papers of the Bibliographical Society of America	—
27	Journal of Information Technology	1.966	57	The Serials Librarian	—
28	Journal of Operations Management	2.420	58	Universal Access in the Information Society	—
29	Journal of the American Society for Information Science and Technology	1.954	59	Young Adult Library Services	—

（**收稿日期**:2012-02-22）

（**上接第 12 页**）

9 **张朝胜等. 基于条件随机场的英文产品命名实体识别. 计算机工程与科学**, 2010(6)

10 Suakkaphong, N., Z. Zhang and H. C. Chen, Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields. Journal of the American Society for Information Science and Technology. 2011,62(4):727-737

11 J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Intl. Conf. on Machine Learning, 2001.

12 Stanford Log-linear Part-Of-Speech Tagger. [2011-05-08]. http://nlp.stanford.edu/software/tagger.shtml

13 **廖先桃**.CRF **理论、工具包的使用及在** NE **上的应用**.[2008-04-02]. http://it.hit.edu.cn/phpwebsite

14 A. McCallum, D. Freitag, and F. Pereira. Maximum Entropy Markov Models for InformationExtraction and Segmentation. In Intl. Conf. on Machine Learning, 2000.

15 Maximum Entropy Modeling Toolkit for Python and C + +. [2006-12-05]. http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

16 The Open NLP Maximum Entropy Package. [2010-12-05]. http://sourceforge.net/projects/maxent

17 Niu, C., et al., A bootstrapping approach to named entity classification using successive learners. 41st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference,2003:335-342

（**收稿日期**:2012-02-22）