中文微博命名实体识别

邱泉清 苗夺谦 张志飞
(同济大学计算机科学与技术系 上海 201804)
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 200092)

摘 要 微博这一媒体形式的迅速发展为命名实体识别提供了一个新的载体。根据微博文本的特点,提出针对中文微博的命名实体识别方法。首先,对微博文本做规范化处理,消除由于微博表达不规范造成的干扰;在建立中文人名库、常用地点库等知识库的基础上,选取适合微博的特征模板,使用条件随机场方法进行实体识别;同时,将正确的识别结果添加到知识库中以提升识别效果。在真实微博数据上的实验表明,该方法能够有效地完成中文微博的命名实体识别任务。

关键词 中文信息处理,微博,命名实体,条件随机场
中图法分类号 TP391 **文献标识码** A

Named Entity Recognition on Chinese Microblog

QIU Quan-qing MIAO Duo-qian ZHANG Zhi-fei
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)
(Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China)

Abstract The rapid development of microblog brings a new carrier for named entity recognition. The paper proposed an approach for named entity recognition on Chinese microblog according to the features of microblog. First of all, the paper normalized the text of the microblog and eliminated the interference caused by non-standard expression, then constructed several knowledge bases, such as Chinese person names, common place names and organization names, and devised feature templates for the recognition method based on conditional random fields. Meanwhile the correct recognition results were added to the knowledge bases to improve the performance of recognition. The experiment results show that our approach is effective to recognize named entities on Chinese microblog.

Keywords Chinese information processing, Microblog, Named entity, Conditional random fields

1 引言

命名实体识别任务主要是要识别出文本中出现的专有名称和有意义的数量短语并加以归类,其中,主要的识别对象是人名、地名和组织名[1]。命名实体识别是信息抽取、信息检索、机器翻译、文摘生成等技术的关键问题[2]。

目前,对非微博形式的命名实体识别研究已较为成熟,但是针对微博内容的研究则比较少。微博是近年来发展非常快且影响非常大的网络全民媒体形式。Twitter.com自从2006年在美国上线以来,受到很多网民的欢迎。国内的各大互联网服务商也相继推出各自的中文微博平台,如新浪微博(weibo.com)、腾讯微博(t.qq.com)等,同时也将这一媒体形式推广到移动终端上,仅仅新浪微博的注册用户量已突破3亿,每日微博量超过1亿。此外,微博的即时性很强,信息在微博上的传播速度很快,全民媒体在消息传递方面的功能不容忽视[3]。因此,对微博内容进行命名实体识别具有重要的实际意义。

微博的文本长度限制在每条140字以内,而每条微博的平均长度约为50字。相比其他形式的文本(如新闻文本),其具有文本短、口语化、网络化、表达不清晰等特点,这也给针对微博的命名实体识别造成了一些新的困难:

1)特殊的表达形式:微博中常用两个"#"符号间的内容表示主题,"[ ]"及其中的内容则常表示为表情(如[晕]、[赞]等);微博中常出现多种语言(日语、英语等)混合的现象等。

2)简称、代称:如翔爷(刘翔)、范爷(范冰冰)、给力芬(格里芬)等。

3)语意不完整:如"我中奖了"写成"我中了";另外文本由于长度短,缺少上下文环境。

因此,依据微博的这些特点,提出了适合微博的命名实体识别方法,对微博文本进行了规范化,并且建立了适合于微博的知识库,并通过实验证明了此方法对于微博命名实体识别的有效性。

到稿日期:2012-09-10 返修日期:2012-12-15 本文受国家自然科学基金项目(60970061,61075056,61103067),中央高校基本科研业务费专项资金资助项目资助。

邱泉清(1988-),男,硕士生,CCF学生会员,主要研究方向为模式识别、自然语言处理等,E-mail:qiuquanqing612@yahoo.cn;**苗夺谦**(1964-),男,教授,博士生导师,CCF高级会员,主要研究方向为粗糙集理论、粒计算、Web智能、模式识别等;**张志飞**(1986-),男,博士生,CCF学生会员,主要研究方向为自然语言处理、机器学习等。

2 相关工作

对于命名实体识别的研究已经取得了不少的成果，该方法主要分为3类：基于规则的方法、基于统计的方法、规则与统计相结合的方法[2]。统计方法中常用的有最大熵模型[4]、隐马尔可夫模型[5]、条件随机场[6]等。

虽然对于一般文本的识别研究较为成熟，但针对微博尤其是中文微博的识别研究则较少。

2.1 非微博命名实体识别研究

周昆[7]通过构建规则库，采用规则匹配的方法识别命名实体。Della Pietra等人[8]最早将最大熵方法引入到自然语言处理中来建立语言模型。Chen等人[9]建立了条件随机场和最大熵模型，识别F值达到86.2%。

2.2 微博命名实体识别研究

Ritter等人[10]对Twitter文本设计新的词性标注和分块方法，然后对实体进行分类，提高了传统方法在微博命名实体识别的准确率。Ek等人[11]使用正则表达式结合分类器对瑞典语手机短信进行命名实体识别，并优化算法，提高其在手机上的运行效率。

3 微博命名实体识别方法

3.1 微博文本规范化

微博文本规范化的目的是消除干扰，降低噪声，主要包括：

1)清除无意义的符号。主要有以下几种：用于表示表情的“[ ]”符号及表情内容(如“[晕]”)；用于表示主题的一对“#”符号及其间的内容；网页链接(如“http://t.cn/zOl83Ai”)；重复的符号(如“好开心～～～～”)。

2)语言表达统一。建立繁简转换字库，将所有繁简夹杂的微博转换为中文简体。对于出现的日文和英文，由于情况较少且不影响整体识别，因此将其去除。

3)去除符号“@”及其后的名称。微博中用“@+用户名”表示链接某个用户(用户名既可以是真实人名也可以是非人名)，对于实体识别没有实际意义，因此去除该词语。

4)去除长度小于5个字符的微博。有些微博的长度过短，不包含命名实体，或者由于规范化处理1)和3)使得微博的长度小于5个字符，将这些微博去除。

3.2 条件随机场

条件随机场(Conditional Random Fields，CRFs)是一种无向图模型。它没有隐马尔可夫模型那样强的独立性假设，同时也克服了标记偏置问题。Lafferty等人[12]在2001年提出这种方法用于为切分和标记序列数据建立统计模型。

3.2.1 无向图模型

条件随机场最简单和普遍的结构是线性链结构，如图1所示。在图形模型中的各输出结点被连接成一条线性链的特殊情形下，CRFs假设在各个输出结点之间存在一阶马尔可夫独立性，二阶或更高阶的模型可类似扩展。

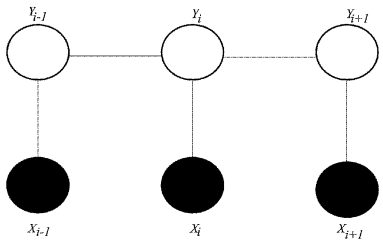

图1 线性链结构的条件随机场无向图模型

在给定观察序列$X=(X_1,X_2,\cdots,X_n)$的条件下，标记序列$Y=(Y_1,Y_2,\cdots,Y_n)$的条件概率分布$P(Y|X)$构成条件随机场。

设X和Y均为线性链表示的随机变量序列，则$P(Y|X)$称为线性链条件随机场。在X取值为x的条件下，Y取值为y的条件概率满足[13]：

$$P(y|x)\propto\exp\left(\sum_{i,k}\lambda_k f_k(y_{i-1},y_i,x)+\sum_{i,k}\mu_k g_k(y_i,x)\right) \quad (1)$$

式中，f_k和g_k是特征函数，λ_k和μ_k是通过参数估计确定的参数。

3.2.2 特征模板

特征模板在命名实体识别时用来匹配信息构成具体特征，选择合适的特征模板显得尤为重要。针对微博的特点，选取了一组特征模板，如图2所示。

单个词：	Word(−2) Word(−1) Word(0) Word(1) Word(2)
单个词性：	POS(−2) POS(−1) POS(0) POS(1) POS(2)
相邻两个词：	Word(−1)/Word(0) Word(0)/Word(1)
相邻两个词性：	POS(−2)/POS(−1) POS(−1)/POS(0) POS(0)/POS(1) POS(1)/POS(2)
相邻三个词性：	POS(−2)/POS(−1)/POS(0) POS(−1)/POS(0)/POS(1) POS(0)/POS(1)/POS(2)

图2 条件随机场特征模板

Word(0)表示当前词；POS(0)表示当前词的词性；括号中的数字表示与当前词的距离，负数表示当前词左侧，正数则表示当前词右侧；多个特征表示组合模板。

3.3 中文命名实体识别知识库

中文命名实体特征很大一部分以隐含的语义特征形式存在。针对微博特征，建立外部语义知识库，主要包括：

1)人名指示词库、地名指示词库和机构名指示词库：命名实体指示词通常标志着该词的周围会出现相应类型的命名实体。

2)中国人名姓氏表：中国人名通常由姓氏+名构成，从维基百科中获得中国人名姓氏表。由于微博中常常讨论名人，因此也建立了娱乐明星、体育明星等名人表。

3)常用地名表：从维基百科的“中文地名列表”中收录了常见的中国地名。

4)常见组织名：抽取1998年1月份《人民日报》语料、新浪微博语料中出现的组织名，建立常见组织名列表。

4 实验

4.1 实验语料

在两个语料库上开展实验，分别是：

1)北京大学收集的《人民日报》语料库中1998年01月份的语料，该语料已有标注。

2)新浪微博收集的2012年7月的8000条微博。由于微博的总量比较大，且为了避免由于发布时间过近而造成选取的微博中某些实体重复出现的频率过高，而其他实体出现概率小的问题，故在7月的微博中，从时间上平均地选取了8000条。采用ICTCLAS系统进行分词及词性标注，然后人工标注语料实体信息。

此外，选取《人民日报》语料中1月1日至1月20日的语料、新浪微博语料中的3000条微博作为训练集；《人民日报》

语料中1月21日至1月31日的语料、新浪微博语料中的另外5000条微博作为测试集。

4.2 评价标准

采用的评价标准包括准确率 P(Precision)、召回率 R(Recall)和 F 值(F-measure)[1],具体定义如下:

$$P=\frac{\text{正确识别出的命名实体个数}}{\text{识别出的命名实体个数}}\times 100\%$$

$$R=\frac{\text{正确识别出的命名实体个数}}{\text{标准结果中的命名实体个数}}\times 100\%$$

$$F\text{值}=\frac{2\times P\times R}{P+R}\times 100\%$$

4.3 实验结果及分析

4.3.1 不同语料的命名实体识别效果

采用条件随机场方法对《人民日报》语料和新浪微博语料进行命名实体识别,比较在新闻语料及微博语料上的表现差异,如表1所列。

表1 不同语料的识别效果

实体类型	语料库	准确率(P)	召回率(R)	F值
人名	《人民日报》	99.75%	100.00%	99.87%
人名	新浪微博	62.09%	63.01%	62.55%
地名	《人民日报》	90.89%	94.36%	92.60%
地名	新浪微博	75.04%	82.23%	78.47%
组织名	《人民日报》	68.62%	70.97%	69.78%
组织名	新浪微博	42.50%	61.04%	50.11%

在该实验中,测试集和训练集各自使用相应语料库的数据。实验结果表明,条件随机场方法对于传统新闻语料的命名实体识别任务,表现相当出色,尤其对于人名识别,准确率和 F 值都接近100%。但是在对新浪微博语料的命名实体识别中,识别效果明显下降,这也说明微博的表达不规范等特点对命名实体识别带来了较大影响,需要进一步改进。

4.3.2 规范化对命名实体识别的影响

对规范化后的微博语料进行训练和测试,观察规范化后的命名实体识别效果,如表2所列。

表2 规范化后的识别效果

实体类型	是否规范化	准确率(P)	召回率(R)	F值
人名	未规范化	62.09%	63.01%	62.55%
人名	规范化	74.77%	76.38%	75.57%
地名	未规范化	75.04%	82.23%	78.47%
地名	规范化	85.93%	87.39%	86.65%
组织名	未规范化	42.50%	61.04%	50.11%
组织名	规范化	47.80%	74.72%	58.30%

根据表2,从人名、地名和组织名在3个评价标准上的识别效果来看,规范化后的表现都得到了提高,其中对人名准确率的提高最显著。由此可以说明规范化对微博命名实体识别有很大帮助,有效提高了识别效率。

4.3.3 知识库对命名实体识别的影响

在微博文本规范化的基础上,引入建立的知识库,进一步验证知识库对提高命名实体识别效率的有效性。

从表3中可以看出,加入外部语义知识特征后,人名、地名和组织名的识别准确率、召回率和 F 值均有提升。主要原因在于外部语义知识能够描述中文词语间的隐含关系,这对于人名、地名和组织名的识别均有很大帮助,而且部分在微博中常出现的热词也能够帮助提高识别效果。人名、地名、组织名的 F 值分别提升了5.84%、2.39%和7.05%。

表3 知识库的有效性

实体类型	是否使用知识库	准确率(P)	召回率(R)	F值
人名	未使用	74.77%	76.38%	75.57%
人名	使用	85.37%	78.05%	81.54%
地名	未使用	85.93%	87.39%	86.65%
地名	使用	88.43%	90.27%	89.34%
组织名	未使用	47.80%	74.72%	58.30%
组织名	使用	51.46%	82.41%	63.35%

结束语 本文提出使用条件随机场方法对微博文本进行命名实体识别,根据微博特点,对微博内容规范化,并且建立适合于微博的特征模板以及知识库。通过若干实验说明了直接使用传统方法对于微博命名实体识别的缺陷,并且验证了规范化及知识库对于微博命名实体识别的有效性,有效提高了识别效果。但是由于微博的特殊性,识别结果并未达到对类似《人民日报》语料库的识别效果,因此,如何进一步提高其识别效果是下一步的工作重点。

参 考 文 献

[1] 命名实体评测大纲[C/OL]. 863命名实体识别评测组,2004. http://www.863data.com.cn

[2] 张晓艳,王挺,陈火旺. 命名实体识别研究[J]. 计算机科学,2005,32(4):44-48

[3] 郑斐然,苗夺谦,张志飞,等. 一种中文微博新闻话题检测的方法[J]. 计算机科学,2012,39(1):138-141

[4] 杨华. 基于最大熵模型的中文命名实体方法研究[D]. 长沙:国防科学技术大学,2008

[5] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87-94

[6] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报,2006,34(5):804-808

[7] 周昆. 基于规则的命名实体识别研究[D]. 合肥:合肥工业大学,2010

[8] Della Pietra S,Della Pietra V,Mercer R L,et al. Adaptive language modeling using minimum discriminant estimation[C]// Acoustics,Speech,and Signal Processing,ICASSP-92. USA,1992:633-636

[9] Chen A,Peng F,Shan R,et al. Chinese named entity recognition with conditional probabilistic models[C]// Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. Australia,2006:173-176

[10] Ritter A, Clark S, Mausam, et al. Named entity recognition in tweets:an experimental study[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. USA,2011:1524-1534

[11] Ek T,Kirkegaard C,Jonsson H,et al. Named entity recognition for short text messages[J]. Procedia-Social and Behavioral Sciences,2011,27:178-187

[12] Lafferty J,Mccallum A,Pereira F. Conditional random fields:probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 8th International Conference of Machine Learning. USA,2001:282-289

[13] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012:194-196