文章编号:1007-1423(2016)03-0024-03 DOI:10.3969/j.issn.1007-1423.2016.03.006

命名实体识别综述

陈基

(四川大学计算机学院,成都 610065)

摘要:
互联网的普及和发展,信息资源得到极大的丰富,同时也造成信息过载的问题。人们迫切需要快速准确地获取信息的技术方法,信息抽取技术就应运而生。命名实体识别作为信息抽取的一个子任务被提出,受到国内外学者的重视,并进行一系列研究。探讨命名实体的概念和意义,对现有的命名实体识别研究进行总结归纳。

关键词:
命名实体;条件随机场;信息抽取;评价指标

0 引言

MUC-6 第一次提出命名实体识别[1],现在在自然语言处理中已经被广泛使用。信息抽取就是从非结构化的文本中(例如,新闻)抽取结构化的数据和特定的关系。在定义任务的过程中人们注意到识别信息单元的名称,像人名、机构名、地名、时间等是必不可少的。识别上面所说的实体的名称,就叫做命名实体识别[1]。命名实体识别是信息抽取的子任务,识别的好坏直接关系到抽取的好坏。

早期的命名实体识别工作,主要识别一般的"专有名词"[2],包括三类名词:人名、地名、机构名。这也是 MUC-6 最早定义的任务要识别的名词。随着研究的进行,人们对这些名词进行更细致的划分。对于地名,可以进行细分为:国家名、省/州、城市名、街道名等[3]。类似的人名可以细分为:政客、演员等[4]。除了识别一般的专有名词,人们也开始关注对于特定领域的命名实体识别。在生物医学领域,对于基因名、蛋白质名的识别已经有许多工作在开展,也取得了不错的效果[5]。针对社交媒体文本中存在大量的电影、歌曲等,识别电影名、歌曲名、邮件地址等实体[6]。随着研究范围的扩大,针对不同的特定问题特定领域,越来越多的实体类型被提出。

1 技术方法

早期的研究大多数通过人工构造规则的方法,现在多采用监督学习的方法,自动构造规则或者进行序列标注。监督学习的从发点是从标注好的文档的正负例里面学习特征,通过自动学习到的这些特征来识别命名实体。1.1 部分对序列标注方法进行更细致的介绍。序列标注的方法的主要缺点是需要大量标注好的语料。当没办法获取大规模的语料或者代价比较昂贵时,人们提出其他的解决方法,包括:半监督和无监督方法。这两部分内容在 1.2 和 1.3 中介绍。

1.1 有监督方法

有监督学习方法将命名实体识别看做序列标注问题。序列标注模型包括:隐马尔科夫模型 Hidden Markov Models (HMM)[7],最大熵马尔科夫模型 Maximum Entropy Markov Models (MEMM) 和条件随机场 Conditional Random Fields (CRF)[8]等。这些模型都是基于大量的标注语料,定义一系列实体,通过学习得到基于特征的判别规则。隐马尔科夫模型描述了一个含有隐含未知参数的马尔可夫过程,针对命名实体识别这里的未知参数为实体类型。

隐马尔科夫模型考虑了上下文信息,测试时求得的解是全局最优的解,得到最优的马尔科夫链,这是传

统分类算法做不到的。隐马尔科夫模型缺点是假设可观测变量之间独立,而且限制观测变量是词语本身,限制了特征的选择。例如像字数、DF 词频、位置等对实体类型很有预示作用的特征都无法很方便地使用。

最大熵马尔科夫模型只计算给定可观测变量下隐藏变量的概率,将模型由隐马尔科夫的生成模型变成判别模型,克服了隐马尔科夫的模型的缺点,可以方便使用各种特征。不过也带来新的问题——标记偏置问题。

条件随机场模型将最大熵马尔科夫模型里面的条件概率转化为特征函数的形式,分解为两部分:转移特征和状态特征。通过训练得到不同特征的权值,测试的时候一般采用维特比(Viterbi)算法进行求解。条件随机场模型克服最大熵马尔科夫模型的标记偏置问题,不过也带类训练速度偏慢的问题。

在这些模型基础上,国内外学者针对不同的问题还提出许多改进的版本:层叠隐马尔科夫模型[9]、层叠条件随机场[10]等。

1.2 半监督方法

半监督也叫弱监督,主要的技术叫拔靴法(Bootstrapping),只提供很少的标注数据,例如一些种子用于开始的学习。例如识别疾病名的系统,需要用户提供一些样例。然后系统就会搜索包含这些实体的句子,辨别它们的上下文环境。接着系统就会寻找其他跟之前样例有相识的上下文的疾病名。学习的过程就是不断地循环这个过程,发现新的上下文,发现新的疾病名,产生大量的基疾病名和上下文。辨别上下文环境的方法包括:M. Collins 和 Singer 采用模板的方式[11]、A. Cucchiarelli 和 Velardi 采用句法分析树[12]等。半监督的方法可以在很少量的标注数据和大量无标注的数据条件下,取得比较好的效果。

1.3 无监督方法

无监督学习最典型的方法是聚类。比如,通过相似的上下文将不同的命名实体聚到一起。当然还有其他的无监督方法,包括:基于外部资源(wordNet)[13],当针对某个特定的领域的标注语料没有时候,可以采用外部资源比如 wordNet 进行迁移学习。首先,通过词在大规模语料中的共现,对 wordNet 里面的同义词分配一个实体类型。然后对于给定的文档中一个词,通过比较一定窗口的上下文,给它分配一个实体类型。基于点互信息[14],将点互信息做为特征对给定的词进行分类,判断输入哪个类型。还有基于词汇模板[15]等。

2 特征

特征是在算法假设下描述词的各种属性。例如一个布尔型的特征,如果当前单词是大写则为真,否则为假。特征一般用特征向量表示,一个维度代表一个特征取值可以是布尔型、数值型等,整个向量就表示词在假设条件下所有属性。特征一般分为三类:词级别特征,包括词本身是否大小写、前后文的词、词性等;字典级别特征,判断当前词是否属于某个字典,如地名字典,姓名字典等;全局特征。

3 评价指标

命名实体识别一般采用这几个评价指标:精确率(Precision)、召回率(Recall)和F 值。

表 1

	识别为正例	识别为负例
实际为正例	TP	FN
实际为负例	FP	TN

精确率 p 和召回率 r 定义如下:

$$P=\frac{TP}{TP+FP},\ r=\frac{TP}{TP+FN}$$

F 值是精确率和召回率的调和平均值。

$$F=\frac{2}{\frac{1}{P}+\frac{1}{r}}$$

4 结语

命名实体识别作为信息抽取的子任务,从一开始提出就得到国内外学者的重视,并成为研究热点,取得众多进展。本文主要从三类技术方法:监督学习、无监督学习、半监督学习,介绍了相关的研究工作。一般文本的命名实体识别已经相当成熟,目前大部分命名实体识别研究,侧重于对特点领域的命名实体,例如生物医学、社交媒体。

参考文献：

[1]Grishman, Ralph; Sundheim, B. 1996. Message Understanding Conference - 6: A Brief History. In Proc. International Conference on Computational Linguistics.

[2]hielen, Christine. 1995. An Approach to Proper Name Tagging for German. In Proc. Conference of European Chapter of the Association for Computational Linguistics. SIGDAT.

[3]Lee, Seungwoo; Geunbae Lee, G. 2005. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping. In Proc. International Joint Conference on Natural Language Processing.

[4]Fleischman, Michael; Hovy. E. 2002. Fine Grained Classification of Named Entities. In Proc. Conference on Computational Linguistics.

[5]Settles, Burr. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In Proc. Conference on Computational Linguistics. Joint Workshop on Natural Language Processing in Biomedicine and Its Applications.

[6]X Liu, M Zhou - Information Processing & Management, 2013

[7]Bikel, Daniel M.; Miller, S.; Schwartz, R.; Weischedel, R. 1997. Nymble: a High-Performance Learning Name-finder. In Proc. Conference on Applied Natural Language Processing.

[8]McCallum, Andrew; Li, W. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. In Proc. Conference on Computational Natural Language Learning.

[9]刘杰. 基于统计的中文机构名实体识别的研究[J]. 佳木斯大学学报(自然科学版),2010(03)

[10]俞鸿魁,张华平,刘群,吕学强,施水才. 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(02)

[11]Collins, Michael; Singer, Y. 1999. Unsupervised Models for Named Entity Classification. In Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[12]Cucchiarelli, Alessandro; Velardi, P. 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. Computational Linguistics 27:1.123-131, Cambridge: MIT Press.

[13]Alfonseca, Enrique; Manandhar, S. 2002. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In Proc. International Conference on General WordNet.

[14]Etzioni, Oren; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; Yates, A. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence 165.91-134

作者简介：

陈基(1990-),男,福建福州人,研究生硕士,研究方向为数据挖掘

收稿日期:2015-12-15　　修稿日期:2015-12-30

Survey of Named Entity Recognition

CHEN Ji

(College of Computer Science, Sichuan University, Chengdu, Chengdu 610065)

Abstract:

With the growing popularity and development of the Internet, information resources have been greatly enriched, but also result in information overload problem. For people's need of technical method that can find out information fast and accurately, information extraction technology is brought into being. Information extraction is presented as a subtask; named entity recognition is attached great importance. A series of studies are doing by scholars. Discusses the concept and significance of named entity, and gives a summary to named entity recognition.

Keywords:

Named Entity Recognition; Conditional Random Fields; Information Extraction; Evaluation Index