

命名实体识别研究发展综述

周玉新

内蒙古民族大学计算机科学与技术学院 内蒙古通辽 028000

摘要:随着互联网技术的飞速发展和极大普及,以及相关领域研究的不断深入,可用信息资源得到了极大丰富。人们迫切需要从海量的非结构化文本中获取有用的信息。在这一背景下,信息抽取技术应运而生。命名实体识别自诞生之日起,就被看作信息抽取系统的一个重要子任务,受到广大国内外学者的广泛关注。本文探讨了命名实体识别的基本概念和意义,并对现有的命名实体识别技术、特征、评估方法进行了总结。

关键词:命名实体识别;信息抽取;评估方法

目前广泛应用于自然语言处理领域的“命名实体”最初于 1996 年在第六届信息理解会议(MUC-6)上提出,那时,MUC 主要侧重于信息抽取任务。信息抽取是从给定文本中抽取诸如公司活动和国防活动等特定的信息,这些文本可以是结构化、半结构化或非结构化的数据。在进行信息抽取任务时,人们发现识别文本中某些具有特殊意义的实体,如包括人名、机构名称和地名在内的名称和包括具有特殊意义的时间、日期及百分数在内的数字是必不可少的。在文本中识别这些实体的任务被称为“命名实体识别”,普遍被认为是信息抽取的一个重要子任务,它的主要任务是抽取文本中的专有名词、生物物种和有意义的时间、日期和数量短语并进行分类。

早期的命名实体识别研究工作主要对文本中的“专有名词”进行识别,其中研究最多的三种“专有名词”是人名、地名和结构名称,这些名称可以被转换为更细化的类型,如地名可以被细化为:城市、州和国家。同样,人名可以细化为政治家和演员等。

近年来,除了识别一般文本中的专有名词外,一些生物医学语料库,如 GENIA 的出现引起了一些命名实体识别研究者对生物医学实体识别研究的兴趣,这些命名实体主要包括蛋白质、DNA、RNA 和细胞类型等。大多数生物医学领域的命名实体识别研究主要集中于对蛋白质的识别,也有一些关于药品和化学名称实体识别的研究。

最近的一些研究并不局限于抽取这些可能的实体类型,一些细化的类如博物馆、河流或机场等引起了一些研究人员的兴趣,并且还增加了一些范围更广的类,如产品和事件,以及物质、动物、种族或颜色等。随着命名实体识别研究范围的进一步扩大,针对不同的特定领域,越来越多的实体类型得到了广大实体识别研究工作者的关注。

1 学习方法

在信息抽取系统中,识别未知实体的能力是一个非常重要的部分,这样的能力取决于系统所使用的识别和分类规则,这些规则由与正例和负例相关的特有规则触发。早期的研究大多采用基于人工构造规则的方法,而现在大多使用监督的机器学习方法。

监督学习方法的思想是在大量标注的文档上学习命名实体正例和负例的特征并设计捕获给定类型本质的规则。而语料库的缺乏和构造这些资源的高昂成本导致了两种可替代的学习方法:半监督学习和无监督学习。

1.1 监督学习

目前,命名实体识别所使用的主流技术是监督学习。监督学习包括隐马尔科夫模型、决策树、最大熵模型、支持向量机、条件随机域等^[1],这些方法都是命名实体识别系统的变体,这些系统都是读取大量的标注语料,存储一系列实体,并且构造基于特征的判别规则。

通常提出的基本监督方法包括标注测试语料库的词,这些词在训练集中被注释为实体。系统的性能依赖于同时出现在训练语料库和测试语料库中的词所占的比例,通常称之为词汇转移。

1.2 半监督学习

由于可用标注语料库的匮乏以及大量未标注语料库的存在,研究人员提出了一种半监督学习方法,也称为弱监督学习。主要的半监督学习方法被称为“bootstrapping”,它只需要提供少量的标注数据,例如一组

种子用于开始的学习。然后,系统搜索包含这些已提供数据的句子并尝试发现出现在相似上下文中实体的其他实例。接着将学习过程应用于新发现的例子以发现新的相关上下文。通过重复这一过程收集大量命名实体和大量上下文信息。半监督方法只需要较少的已标注数据,从而在大量无标注数据的条件下获得可以与监督学习方法相媲美的性能。

1.3 无监督学习

由于现实中存在的大量无标注数据,在未进行标注的数据中,试图找到隐藏的实体,即无监督学习问题被提上日程。提供给系统的实例是无标记的,这区别于监督学习和半监督学习。典型的无监督学习方法是聚类^[2],我们可以尝试根据上下文的相似性从聚类组中收集命名实体。

2 命名实体识别的特征空间

特征是用来描述命名实体的各种属性,对不同的识别系统来说所采用的特征也不同。我们通常用特征向量来描述系统所使用的特征,特征向量描述是由一个或多个布尔型数据、数值数据和标量数据所表示的每个词的文本抽象。经常用于命名实体识别和分类的特征通常包括三种:词级特征、列表查找特征以及文档和语料特征。

3 评估指标

对命名实体识别系统的发展来说,对系统的全面评估是必不可少的,许多系统被要求根据它们标注文本的能力来对系统进行排序。目前,通常采用的评估指标主要有正确率、召回率和 F 值,它们的定义如下:

正确率 = 识别出的正确实体数 / 识别出的实体数

召回率 = 识别出的正确实体数 / 样本中的实体数。

两者的取值都在 0 和 1 之间,数值越接近 1,正确率或召回率就越高。正确率和召回率有时会出现矛盾的情况,这时需要综合考虑它们的加权调和平均值,也就是 F 值,其中最常用的 F1 值,当 F1 值较高时说明试验方法比较有效。F1 值定义如下:

$F1 \text{ 值} = (2 * \text{正确率} * \text{召回率}) / (\text{正确率} + \text{召回率})$ 。

4 结语

命名实体识别作为信息抽取的重要子任务,从提出伊始就得到了广大国内外学者的广泛重视,并且受到了各方面的持续关注,取得了巨大的进展。本文探讨了命名实体识别的基本概念和意义,并对现有的命名实体识别技术、特征、评估方法进行了总结。目前,对某些领域如新闻的命名实体识别研究已经相当成熟,如何将新闻领域中成熟的技术方法应用于一些新兴领域如生物医学等是未来命名实体识别系统发展的趋势。

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] 俞鸿彪,张华平,刘群,吕学强,施水才.基于层叠隐马尔科夫模型的中文命名实体识别[J].通信学报,2006(02).
- [3] 张祝玉,任飞亮,朱靖波.基于条件随机场的中文命名实体识别特征比较研究[C].第四届全国信息检索与内容安全学术会议论文集,2008.
- [4] 王丹,樊兴华.面向短文本的命名实体识别[J].计算机应用,2009,29(1).