学校编码：10384　　分类号_____密级_____
学号：200328023　　UDC _____

厦门大学

硕 士 学 位 论 文

基于条件随机场的中文命名实体识别

Chinese Named Entity Recognition based on Conditional Random Fields

向 晓 雯

指导教师姓名： 史 晓 东 教授
专 业 名 称： 计算机应用技术
论文提交日期： 2 0 0 6 年 4 月
论文答辩时间： 2 0 0 6 年 6 月
学位授予日期： 2 0 0 6 年 月

答辩委员会主席：_____
评 阅 人：_____

2006 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘要

命名实体识别属于自然语言处理的基础研究领域，是信息抽取、信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础。因此，对命名实体识别的研究具有很大的实用意义。

本文针对现代汉语文本的特点，主要研究以人名、地名和组织名的识别为核心内容的中文命名实体识别问题，我们以一种较新型的统计模型--条件随机场为基本框架，设计并实现了一个中文命名实体识别系统。具体说来，本文的主要内容如下：

本文首先分析了命名实体识别的难点，人名、地名、组织名的相关语言学知识，并对现有的一些命名实体识别方法和中文命名实体识别系统进行了简要介绍。

接着，详细介绍了条件随机场的定义、模型结构、势函数、参数估计和训练方法、概率计算方法等。进一步地，将条件随机场模型应用于中文命名实体识别任务，提出了适合于各类中文命名实体的特征模板，并通过实验进行验证，确定了有效特征。

本文最后，实现了一个基于条件随机场的中文命名实体识别系统，系统采用了层叠结构，以模型训练模块和命名实体识别模块作为系统的核心组成部分，在低层条件随机场模型中进行人名、简单地名以及简单组织名的识别，低层的识别结果传递到高层模型，再进行复合地名与复合组织名的识别。实验结果表明，基于条件随机场的中文命名实体识别系统能够获得较为满意的效果，在对 2004 年 863 中文命名实体识别评测语料的开放测试中，系统识别的精确率、召回率和 F 值分别为 82.50%、76.04%和 79.14%。

关键词：命名实体；条件随机场；特征

Abstract

Named entity recognition is one of the fundamental problems in many natural language processing applications, such as information extraction, information retrieval, machine translation, shallow parsing and question answering system. The research of named entity recognition is of great worth.

According to the modern Chinese characteristics, this paper mainly researches Chinese named entity recognition including person names, location names and organization names. We design and implement a Chinese named entity recognition system based on conditional random fields.

This paper is organized as follows:

First, it introduces the difficulties of named entity recognition and the characteristics of person names, location names and organization names. It also compares various named entity recognition methods and some existing Chinese named entity recognition systems.

Then this paper introduces the definition of conditional random fields, the graph structure, potential functions, parameters estimation and probability computations. Regarding conditional random fields as the basic frames, this paper proposes different feature templates for different kinds of named entities.

Finally, it presents a cascaded Chinese named entity recognition system based on conditional random fields. In the system, person names, simple location names and simple organization names are recognized by the lower model at first, and then the result of the lower model is passed to the high model for recognizing the complex location names and organization names. The experimental results show that the system has achieved good performance. In the open test, the recall, precision and F-measure has reached 82.50%, 76.04% and 79.14%, respectively.

Key word: Named Entity; Conditional Random Fields; Feature

目录

Contents

第一章 引言

1.1 研究背景和意义

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网上出现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（Named Entity，NE）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。狭义的讲，可以把命名实体分为人名、地名、组织名等。广义的讲，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。

命名实体识别任务包括（1）发现命名实体，即判断一个文本串是否代表一个命名实体；（2）标注命名实体，即将发现的命名实体标注为某一种具体类型。

命名实体识别属于文本信息处理的基础研究领域，它的研究成果将直接影响到文本信息自动化处理的深层次研究，它是信息抽取、信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础。因此可以说，命名实体识别的研究具有较高的实用意义。

（1）信息抽取

在信息抽取研究中，人们需要从文本中自动抽取出特定的事实信息，形成结构化数据。例如，从一篇新闻报道中抽取出事件的详细情况，包括事件发生的时间、地点、参与人物等。命名实体识别是实现信息抽取的第一步，也是信息抽取中最有实用价值的一项关键技术。

（2）信息检索

在目前大规模知识库的情况下，信息检索过程对于准确率和相关度的要求要高于召回率，而提高准确率和改善相关度的一条重要途径就是以短语为索引词。索引的知识粒度越大，确定性越强，歧义性越小。有实验报告证明，命名实体的识别可以改善系统检索文档的相关度，并提高检索系统的召回率和准确率。

（3）机器翻译

在机器翻译领域，常常需要进行专有名词如人名、地名、组织名等的双语精确翻译，此时文本中存在的大量专有名词无法由人工来校对翻译。因此，准确而高效的自动抽取和识别出文本中的命名实体，对于提高双语翻译的准确率和实用性都具有重要的意义。

（4）组块分析

在组块分析过程中通过命名实体识别把一些重要的命名实体，例如将分词后被切碎的人名、地名、组织名等，合成为一个完整的命名实体，就可大大减少组块分析的错误率与复杂度。

（5）问答系统

一个问答系统不可能穷举用户可能提出的各种问题，例如，一篇文档中包含有“今天是星期天”的信息。当用户提出一个问题:“今天是星期几？”时，系统要能够根据问题，从这篇文档中提取出足够的信息，分析这些信息，然后做出回答。要做到这点，基础工作就是这个问答系统能够识别出这篇文档中的各类命名实体。在上例的问题中问的是时间，因此系统就应能识别出该类命名实体。

1.2 国内外命名实体识别研究现状

近年来，国内外对命名实体识别的研究逐步升温。命名实体识别系统的研究与评测也受到了很多会议的关注。

（1）信息理解系列会议

信息理解系列会议（Message Understanding Conferences，MUC）曾推动了上个世纪九十年代自然语言处理领域信息抽取研究的蓬勃发展。1995年9月举行的MUC-6会议首次出现了术语“命名实体”，并引入了英文命名实体识别的评测任务。在其后的MUC-7的MET-2[①]以及IEER-99[②]、CoNLL-2002、CoNLL-2003[③]等一系列国际会议中，命名实体识别都被作为其中的一项指定任务。

[①] MET-2: Second Multilingual Entity Evaluation Task，1998. 测试的语言包括中文、日文和西班牙语。

[②] IEER99: the 1999 Information Extraction-Entity Recognition Evaluation. 测试的语言包括英文和中文。

[③] CoNLL: Conferences on Natural Language Learning. CoNLL-2002 评测语言包括西班牙语和荷兰语，CoNLL-2003 评测语言包括英语和德语。

（2）自动内容抽取评测会议

2000年12月由美国国家标准技术研究所组织的自动内容抽取（Automatic Content Extraction，ACE）评测会议将实体识别作为它评测的两大任务之一。最近一次的ACE评测于2005年11月举行[①]，评测语种包括英文、中文和阿拉伯文，识别的实体共7类，包括人物（Person）、地理政治实体（Geo-Political Entity）、地名（Location）、组织（Organization）、武器（Weapon）、交通工具（Vehicle）、设施（Facility）等，另外还包括了对时间（Time）和数值（Value）的识别。命名实体可以看作是ACE识别的实体的子集。ACE识别的实体更像是名词短语，可以嵌套，类别也更多，同时还需要确定实体间的共指关系，因此难度较大。

（3）863评测会议

在国内，863计划中文信息处理与智能人机交互技术评测，于2003年首次将中文命名实体识别作为其分词标注评测的子任务，在2004年更将其作为一个独立的评测项目[②]。2004年的命名实体任务由三个子任务组成：命名实体、时间表达式、数字表达式，其中命名实体又分为人名、地名和组织名三类。

目前，在英文命名实体识别方面人们已经进行了大量的工作并取得了比较满意的效果。Bikel D.等提出的基于隐马尔可夫模型的英文命名实体识别方法，在MUC-6评测中，对英文地名、组织名和人名识别的准确率分别达到了97%，94%和95%，召回率分别达到了95%，94%和94%；在MUC-7评测中，表现最好的命名实体识别系统达到了95%的准确率和92%的召回率；在CoNLL-2003的命名实体识别评测中，成绩最高的命名实体识别系统的准确率、召回率和F值分别为88.99%、88.54%和88.76%。

相对来说，中文命名实体识别的研究要比英文困难许多。这主要表现在两个方面：一方面，中文文本中没有空格标志词语边界，即汉语在形式上，并没有“词”这个概念，因此常常要先对其进行词法分析—分词；另一方面，中文文本中没有明显的特征来表征一个命名实体，例如中文不像英文那样人名、国家名等专有名词的首字母均大写，并且中文词存在大量的兼类现象，例如“张”这个词可能表示一个常见中文姓氏，也可能表示一个常见量词。此外，对于中文命名实体的定

[①] http://www.nist.gov/speech/tests/ace/ace05/

[②] 2004 年 863 命名实体识别技术评测包括对简体中文文本和繁体（港澳台地区）中文文本两项评测。http://www.863data.org.cn/