编者按：Internet 时代对中文信息处理提出了更多、更新的需求，同时，致力于中文信息处理研究的队伍也在不断地壮大。在这支队伍中，既有在这个领域里长期辛勤耕耘的老兵，也有初出茅庐的新人。为了使研究者们得以在更高的起点上开展研究，我们特向该领域（或相关领域）的资深专家和学者约稿，这些稿件或是多年研究成果的厚实积累以及发轫于斯的深刻思考，或是具有前瞻性的前沿课题探索，或是相关研究工作系统而深入的综述。我们设立了一个约稿专栏，陆续刊登此类稿件，以飨读者。本期刊登其中的一篇，是赵军研究员的“命名实体识别、排歧和跨语言关联”，相信这篇论文对读者全面、深刻地了解乃至理解相关学术问题，一定会大有裨益。

文章编号：1003-0077(2009)02-0003-15

命名实体识别、排歧和跨语言关联

赵 军

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：命名实体是文本中承载信息的重要语言单位，命名实体的识别和分析在网络信息抽取、网络内容管理和知识工程等领域都占有非常重要的地位。有关命名实体的研究任务包括：实体识别、实体排歧、实体跨语言关联、实体属性抽取、实体关系检测等，该文重点介绍命名实体识别、排歧和跨语言关联等任务的研究现状，包括难点、评测、现有方法和技术水平，并对下一步需要重点解决的问题进行分析和讨论。该文认为，命名实体识别、排歧和跨语言关联目前的技术水平还远远不能满足大规模真实应用的需求，需要更加深入的研究。在研究方法上，要突破自然语言文本的限制，直接面向海量、冗余、异构、不规范、含有大量噪声的网页信息处理。

关键词：计算机应用；中文信息处理；命名实体识别；命名实体排歧；命名实体跨语言关联

中图分类号：TP391 **文献标识码**：A

A Survey on Named Entity Recognition, Disambiguation and Cross-Lingual Coreference Resolution

ZHAO Jun

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Named Entities are important meaningful units in texts. The recognition and analysis of named entities is of great significance in the field of Web information extraction, Web content management and knowledge engineering, etc. The research on named entities includes named entity recognition, disambiguation, coreference resolution, attribute extraction and relation detection, etc. Focusing on named entity recognition, disambiguation and cross-lingual coreference resolution, the paper gives a thorough survey on the state of the art of these tasks, including the challenges, methods, evaluations, performances and the problems to be solved. The paper suggests that, the performances of the current systems of named entity recognition, disambiguation and cross-lingual coreference resolution are far from the requirement of large-scale practical applications. In the view of methods and approaches, named entity recognition, disambiguation and cross-lingual conference resolution should be carried beyond the natural language texts and should be investigated directly among the large-scale, redundant, heterogeneous, ill-formed and noisy web pages.

Key words: computer application; Chinese information processing; named entity recognition; named entity disambiguation; named entity cross-lingual coreference resolution

收稿日期：2009-01-15 **定稿日期**：2009-02-10

基金项目：国家 863 计划资助项目(2006AA01Z144)；国家自然科学基金资助项目(60673042, 60875041)

作者简介：赵军(1966—)，男，研究员，博士生导师，主要研究方向为自然语言处理、信息抽取、知识工程。

1 引言

命名实体是文本中承载信息的重要语言单位。按照 Automatic Content Extraction (ACE)评测计划的定义,实体概念在文本中的引用(entity mention,也可称为指称项)可以有三种形式:命名性指称、名词性指称和代词性指称[1]。例如在句子"[[中国]乒乓球男队主教练][刘国梁]出席了会议,[他]指出了当前经济工作的重点。"中,实体概念"刘国梁"的指称项有三个,其中"中国乒乓球男队主教练"是名词性指称,"刘国梁"是命名性指称,"他"是代词性指称。本文主要讨论与命名性指称相关的研究。围绕命名实体有一系列的研究任务,例如:命名实体的识别、排歧、属性抽取、关系抽取等。其中,命名实体识别任务是识别出文本中实体概念的命名性指称项,并标明其类别(例如人名、地名、机构名、产品名等);命名实体排歧解决的是一个命名性指称项指称多个实体概念的问题以及多个命名性指称项指称同一个实体概念的问题。例如:"迈克尔·乔丹"可以是 NBA 篮球运动员,也可以加利福尼亚大学的统计学教授,需要在具体上下文环境中,把具有歧义的命名性指称项映射到它实际所指的实体概念上去。再例:"迈克尔·乔丹"、"飞人"和 Michael Jordan 都可能指 NBA 篮球运动员,实体排歧过程把它们映射到同一概念实体上去。其中,把"迈克尔·乔丹"和 Michael Jordan 映射到同一实体概念的任务涉及到两种语言间的对应关系,因此也称为命名实体的跨语言关联;命名实体的属性抽取,指的是从网页中抽取出特定实体概念的属性类别和值,例如:对于篮球运动员迈克尔·乔丹,抽取出其"出生日期"是"1963 年 2 月 17 日"、"出生地"是"纽约布鲁克林"、职业是"篮球运动员 后卫"等;实体关系检测指的是通过分析网页信息判断两个实体是否存在关系,存在什么类型的关系。命名实体的识别、排歧、属性抽取以及关系抽取等技术在网络信息抽取、网络内容管理和知识工程等领域中占有非常重要的地位。欧盟于 2008 年启动第七框架计划 (Seventh Framework Programme)项目 OKKAM (http://www.okkam.org/),旨在研发一系列与实体相关的技术,将现有互联网转化为以实体为核心的网络(Enabling the Web of Entities)。本文结合课题组的研究工作,对命名实体识别、排歧和跨语言关联的难点、方法、现有技术水平、存在的问题等进行系统的分析和评述,并对需要重点解决的问题进行分析和讨论。由于篇幅限制,有关实体属性抽取、实体关系检测以及在名词性指称项、代词性指称项层次上的实体识别和分析研究不在本文论述的范围。

2 命名实体识别

一般来说,命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)命名实体。其中时间、日期、货币和百分比的构成有比较明显的规律,识别起来相对容易,而人名、地名、机构名的用字灵活,识别的难度很大,因此命名实体识别通常指的是人名、地名和机构名的识别。

命名实体识别的过程通常包括两部分:①实体边界识别;②确定实体类别(人名、地名、机构名等)。英语中的命名实体具有比较明显的形式标志(即实体中的每个词的第一个字母要大写),所以实体边界识别相对容易,重点是确定实体类别。

2.1 命名实体识别和分类的难点

命名实体识别的主要难点在于:

(1) 命名实体形式多变:命名实体的内部结构很复杂,对中文命名实体来说,情况尤其如此:

人名:人名一般包含姓氏(由一到两个字组成)和名(由若干个字组成)两部分,其中姓氏的用字是有限制的,而名的用字很灵活。人名还有很多其他形式,可以使用名来指代一个人,可以使用字、号等其他命名,也可以使用姓加上前缀或后缀以及职务名来指代一个人。例如:"杜甫、杜子美、子美、杜工部"都是同一个人,"李杜"则是一个简称。

地名:通常的形式是若干个字组成地名,可能包括作为后缀的关键字。也存在一些简称来指称地理位置。例如:"湖北、湖北省、鄂"均是指同一个地方,又如:"广州、广州市、羊城"也是指同一个地方。

机构名:机构名可以包含命名性的成分、修饰性成分、表示地名的成分以及关键词成分等。例如:机构名"北京百富勤投资咨询公司"中,"北京"是表示地名的成分,"百富勤"是命名性成分,"投资咨询"是修饰性成分,"公司"是关键词成分。机构名内部还可以嵌套子机构名,例如:机构名"北京大学附属小学"中嵌套了另一个机构名"北京大学"。机构名中还有很多简称形式,例如:"中国奥委会"、"北师

大二附中"等。

(2) 命名实体的语言环境复杂

命名实体的使用环境也很复杂。命名实体是语言中非常普遍的现象，因此可以出现在各种语言环境中。同样的汉字序列在不同语境下，可能具有不同的实体类型，或者在某些条件下是实体，在另外的条件下就不是实体。例如：

人名："彩霞"在某些条件下指人名，而某些条件下就是一种自然现象；

地名："河南"在某些条件下是一个省名，在某些条件下是泛指；

机构名："新世界"在某些条件下指机构名，在某些条件下只是一个短语。

和英语相比，汉语命名实体识别任务要复杂得多，主要表现在：

(1) 汉语文本没有类似英语文本中空格之类的显式标示词边界的标示符，分词和命名实体识别互相影响。

(2) 英语的命名实体往往是首字母大写的，例如：Liu Chang Le is the founder of Phoenix TV"。而中文文本中没有这样的标示，例如："凤凰卫视的创始人是刘长乐"。

2.2 系统评测和技术水平

对命名实体识别进行评测的国际会议有 MUC (Message Understanding Conference)、SigHAN (The Special Interest Group for Chinese Information Processing of the Association for Computational Linguistics)、CoNLL (Conference on Computational Natural Language Learning)、IEER (Information Extraction-Entity Recognition Evaluation) 和 ACE (Automatic Content Extraction)。MUC 是由美国政府支持的一个专门致力于真实新闻文本理解的例会，在信息提取技术的评测方面起着重要作用。MUC-6 和 MUC-7 设立的命名实体识别专项评测大大推动了英语命名实体识别技术的发展。此外，MUC-6 和 MUC-7 还设立了多语言实体识别评测任务 MET(Multilingual Entity Task)，对日语、西班牙语、汉语等多种语言命名实体识别任务进行评测[2]。SigHAN 从 2003 年开始举办第一届中文分词评测 BAKEOFF (The First International Chinese Word Segmentation BAKEOFF)，2006 年和 2008 年举行的 BAKEOFF-3 和 BAKEOFF-4 设立了命名实体识别专项评测[3]。中国大陆在 2003 年和 2004 年举办的 863 计划"中文信息处理与智能人机接口技术评测"中设立了中文命名实体识别评测任务。以上评测对命名实体识别技术的发展起到了推动作用。

MUC 的测试表明，许多英语命名实体系统已经具备了相当程度的大规模文本处理能力，其中 Language Technology Group Summary 开发的英语命名实体识别系统在 MUC-7 评测中取得第一名，其准确率和召回率分别达到 95% 和 92%[1,4]。和英语命名实体识别任务相比，汉语命名实体识别任务复杂得多，技术还不算成熟，还有很多问题没有得到很好解决。参加 MET-2[1] 评测的汉语命名实体识别系统对人名、地名、机构名识别的最优性能指标(准确率，召回率)只有(66%，92%)、(89%，91%)和(89%，88%)[1]。

BAKEOFF-3[3] 的命名实体识别评测分简体文本和繁体文本两类，简体语料由微软亚洲研究院(MSRC)和 LDC(Linguistic Data Consortium)提供，繁体语料由香港城市大学(CITYU)提供，评测分 Open Track 和 Closed Track 两种类型，表 1 是 BAKEOFF-3 列出了训练集和测试集规模。在 Open Track 中参评系统在使用评测方提供的训练语料外，还可以使用其他语言资源(例如：词典、人名表、地名表、词性标注工具、语料库等)；在 Closed Track 中，参评系统只能使用评测方提供的训练语料。共有 16 个系统参加了 BAKEOFF-3 的命名实体识别测试，表 2 给出了六个测试类别的测试性能最好的系统的指标。

表 1　BAKEOFF-3 命名实体识别评测数据集的指标

数据来源	简繁体类别	训练集规模	测试集规模
MSRC	简体	1.3M/63K(词/词次)	100K/13K(词/词次)
LDC	简体	632K(词)	61K(词)
CITYU	繁体	1.6M/76K(词/词次)	220K/23K(词/词次)

表2 BAKEOFF-3命名实体识别六个评测任务中性能最好的系统的测试指标

数据来源	测试类别	测试性能						
		P	R	F	ORG-F	LOC-F	PER-F	GPE-F
MSRC（简）	Closed	0.889 4	0.842 0	0.865 1	0.831 0	0.854 5	0.900 9	～
	Open	0.922 0	0.901 8	0.911 8	0.859 0	0.903 4	0.960 4	～
LDC（简）	Closed	0.802 6	0.726 5	0.762 7	0.658 5	0.304 6	0.788 4	0.820 4
	Open	0.761 6	0.662 1	0.708 4	0.520 9	0.285 7	0.742 2	0.793 0
CITYU（繁）	Closed	0.914 3	0.867 6	0.890 3	0.804 6	0.921 1	0.908 7	～
	Open	0.869 2	0.749 8	0.805 1	0.680 1	0.860 4	0.809 8	～

注：P,R,F分别表示精确率、召回率和F值，ORG-F、LOC-F、PER-F和GPE-F分别表示机构名、地名、人名和地理政治实体的F值。

2004年“863”举办的命名实体识别评测[5]也分简体文本和繁体文本两类，简体语料由山西大学（SXU）提供，繁体语料由香港城市大学（CITYU）提供。只有一种评测任务，评测组织单位给出标注规范、少量开发语料（帮助参评系统理解标注规范）和评测语料，不提供训练语料。表3给出了2004年“863”评测的任务和数据集的规模。参评系统可以使用其他语言资源（例如：词典、人名表、地名表、词性标注工具、语料库等）。共有八个系统参加了2004年“863”命名实体识别测试，表4给出了两个测试类别的测试性能最好的系统的指标。

表3 2004年“863”命名实体识别评测数据集的指标

数据来源	简繁体类别	训练集规模（词/词次）	测试集规模（字）
SXU	简体	NONE	约400K
CITYU	繁体	NONE	约400K

表4 2004年“863”命名实体识别两个评测任务性能最好的系统的测试指标

数据来源	测试类别	测试性能								
		ORG			LOC			PER		
		P	R	F	P	R	F	P	R	F
SXU（简）	开放	0.646 4	0.574 1	0.608 1	0.870 2	0.784 3	0.825 1	0.813 8	0.884 7	0.847 8
CITYU（繁）	开放	0.398 6	0.253 2	0.309 7	0.683 9	0.700 4	0.692 1	0.398 6	0.253 2	0.309 7

注：P,R,F分别表示精确率、召回率和F值，ORG、LOC、PER分别表示机构名、地名和人名。

分析以上评测结果可以看到：在BAKEOFF-3 MSRC语料和BAKEOFF-3 CITYU语料上的评测结果要好于BAKEOFF-3 LDC语料上的评测结果以及“863”语料上的评测结果。究其原因，可能涉及多个方面，但是其中一个很重要原因是：BAKEOFF-3 MSRC和CITYU评测提供了相当规模的训练集，而BAKEOFF-3 LDC只提供了小规模的训练集，而863评测根本不提供训练集。因为训练集和测试集在题材和体裁方面比较类似，可能使得各个系统在BAKEOFF-3 MSRC语料和BAKEOFF-3 CITYU语料上的评测性能较高，在真实的应用环境中，命名实体识别的性能会大打折扣。

2.3 现有的方法

有关命名实体识别已有大量研究，大致有两种方法。第一种就是人工组织规则方法[6-8]，这种方法代价昂贵，系统性能的好坏主要依赖于有经验的语言学家。当系统移植到新的领域或者语种时，规则需要大量人工修改，甚至需要重新总结和组织。第二种是机器学习的方法，包括：语言模型方法[9-10]、隐马尔可夫模型[11-13]、最大熵模型[4,14]、错误驱动的学习方法[15]、决策树方法[16]、DL-CoTrain和Co-Boost[17,18]等。

命名实体的内部构成和外部语言环境具有一些

特征[19-21]，例如：人名姓氏用字相对集中；地名前面通常有“去”、“在”等词语，并以“县”、“街”、“开发区”等词结尾；机构名通常以“部”、“公司”等词结尾等。实际上，无论何种方法，都在试图充分发现和利用实体所在的上下文特征和实体的内部特征，包括词形、词性和角色级特征等。考虑到每一类命名实体都具有不同的特征，因此，如果用统一的模型去识别人名、地名和机构名是不合适的。例如，人名识别适合于用基于字的模型描述其内部构成，而地名和机构名更适合用基于词的模型描述。另外，在人名中，不同国家的人名的用字、构成都不一样，例如：日本人名用字相对较广，具有相对明显的姓氏特征，但姓氏集合却很大（现版本共收集日本人名姓氏9 189 个），而且日本人名姓氏很多和地名重叠；苏俄人名常用“斯、基、娃”等汉字；欧美人名常用“朗、鲁、伦、曼”等汉字。吴友政等在 3.8 万个欧美人名、4.4 万个苏俄人名和 1.5 万个日本人名的实体名列表上，对外国人名用字进行了较深入的定量分析，表 5 给出了外国人名用字的统计信息[22]。可以看出，300 个高频欧美人名用字覆盖了 98.75%的欧美人名；300 个高频苏俄人名用字覆盖了 99.32%的苏俄人名；1 000 个高频日语人名用字覆盖 94.19%的日本人名；452 个苏俄人名用字与 1 000 个日本人名用字的相同用字只占 14.4%；647 个欧美人名用字与 1 000 个日本人名用字的相同用字只占 26.1%；300 个苏俄人名用字与 300 个欧美人名用字的相同用字虽然达到了 78%，但它们在各自人名中概率分布却存在较大的差别[22]。

表 5　外国人名用字分析

语　种	覆盖率	用字集合大小
欧美人名	98.75%(300)	647
苏俄人名	99.32%(300)	452
日本人名	94.19%(1 000)	1 693
苏俄人名—欧美人名	78%(300, 300)	
苏俄人名—日本人名	14.4%(452, 1 000)	
欧美人名—日本人名	26.10%(647, 1 000)	

以上量化分析表明，不同类型的外国人名用字存在较大差别，如果按照人名的用字和构成特点，把人名分成多个类别并分别利用不同模型进行识别，对于提高人名识别的正确率是非常有益的。为了准确刻画不同实体的内部特征，吴友政等使用多个细分类的实体模型来解决这个问题。把人名分为中国人名、日本人名、苏俄人名、欧美人名和简称人名；地名分为单字地名和多字地名；机构名分为简称机构名和全称机构名等。实验证明这种划分方法取得了良好的效果。

2.4　需要解决的问题

虽然在一些评测中，命名实体识别的性能达到较高水平，但是评测有很大的局限性，在真实应用环境中，命名实体识别的性能会大打折扣，命名实体识别问题还远远没有得到解决。

（1）系统的自适应能力不强

自适应能力不强是统计学习方法普遍存在的问题，主要表现为，一个系统在与训练集题材和体裁相似的测试集上表现好，但是在不相似的测试集上性能有明显下降。自适应问题也是命名实体识别面临的一个重要挑战。具有良好自适应能力的系统可以在具有一定差异的语料上运行并具有好的识别效果，但是目前命名实体识别系统的自适应能力还不强。

吴友政等以 1998 年 1 月～5 月的《人民日报》标注语料库为训练集，开发了汉语命名实体识别工具，在 1998 年 6 月的《人民日报》语料上的开发测试结果和 2004 年“863”测试的结果对比如表 6 所示，可以看出 2004 年“863”测试的性能明显较低[22]。性能下降的原因中，一个主要的原因是系统识别模型的自适应能力不足。2004 年“863”测试的语料主要来源于当年的“新浪”等网站，虽然和《人民日报》一样都是新闻语料，但是因为年代不一样，命名实体有很大的差异。BAKEOFF-3 LDC 语料的评测结果低于 BAKEOFF-3 MSRC 语料和 BAKEOFF-3 CITYU 语料的评测结果的一个重要原因也是现有系统的自适应能力不足造成。

有关增强统计机器学习模型的自适应能力的研究已经有很多，但是性能还不满意，值得深入研究。与命名实体识别系统的自适应问题相关的研究工作之一见文献[23]，刘康等针对自然语言理解中的领域适应性问题，从领域概率分布的角度给出了一种基于混合模型的领域适应性学习方法，它可以把判别式模型和产生式模型集成到一个框架下。判别式模型比产生式模型有较好的分类效果，产生式模型的优势在于在训练过程中可以较容易地引入非标注样本从而具有较好的推广性能，而混合模型可以集中它们各自的优势。对于领域适应性问题，他们利用不同领域概率分布之间的差异性，调节训练集中

表6 吴友政等的系统在《人民日报》语料和"863"语料上的测试性能

类别	性能					
	精确率		召回率		F1值	
	RMRB	"863"	RMRB	"863"	RMRB	"863"
人名	0.940 6	0.813 8	0.952 1	0.884 7	0.946 3	0.847 8
地名	0.939 8	0.870 2	0.934 8	0.784 3	0.937 3	0.825 1
机构名	0.846 9	0.646 4	0.868 6	0.574 1	0.857 6	0.608 1

样本的权重，从而使得训练得到的分类器更加适应于目标领域。在命名实体识别、文本倾向性分类等两个自然语言处理任务的实验结果证明了该方法的有效性，相对于传统监督学习以及半监督学习的方法，该方法有较大的优势，同时也证明了混合模型的方法要优于单个模型的方法[23]。

(2) 面向网页信息的、多分类的命名实体识别技术

目前的命名实体识别研究和测试主要是面向较为规范的自然语言文本语料进行的，这远远不能满足实际应用的需求。信息提取、问答系统和 Semantic Web 等是命名实体识别的重要应用领域，这些应用都是面向海量网页信息。与规范的文本语料相比，网页信息不规范、存在很多噪音，有些根本就不构成自然语言句子，因此通常的命名实体识别模型所依赖的上下文特征发生了明显变化，使得识别性能剧烈下降。

另一方面，从应用需求看，目前三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)的命名实体类别是远远不够的，例如：产品名(例：摩托罗拉 V8088 折叠手机)[24]、事件名(例：第 6 届苏迪曼杯羽毛球混合团体赛)、手术名(例如：胆结石腹腔镜手术)等在商务、新闻和医药领域都非常重要。在命名实体细分类方面已经有一些工作。自动内容抽取评测 ACE-2007(Automatic Content Extraction)[1]把实体分为七大类(Person, Location, Organization, Geo-Political Entity, Facility, Vehicle 和 Weapon)、四十五小类；Satoshiba Sekine 等设计了四级共 200 个类别[25]。但是不同应用需要不同命名实体分类体系，不可能固定一种统一的分类体系。

网络应用中命名实体上下文特征明显变化使得识别性能明显下降，而需要的命名实体类型更多、更细，而且有些实体类别是未知、或者是随时间演化的，这就需要我们研究开放域的命名实体识别和分类方法，这种方法不能限定领域、不能限定语料类别。高性能的命名实体识别器需要大规模的训练数据支持，而在开放域命名实体识别和分类任务中，我们不可能为每种类型网页语料、每种实体类别体系都建立训练集，如何自动地生成相应的训练集是需要重点研究的问题之一。在海量冗余网络数据中，有些命名实体的出现有明显的模式，识别出这些命名实体从而建立一个命名实体列表，再利用这些命名实体去发现更多的模式，从而识别出更多的命名实体，迭代地自动生成命名实体标注语料库是一种值得深入研究的方法[26]。

(3) 简称机构名的识别

命名实体特别是机构名，在实际应用中经常以缩略形式出现，本文把缩略形式的机构名称为机构名简称。机构名简称的构成方式可以分为两种，一种是不省略关键词的简称(例如："美国福特公司"简称为"福特公司"，"福建省绿得罐头饮料有限公司"简称为"绿得公司")，另一种是省略了关键词的简称(例如："上海华联超市股份有限公司"简称为"上海华联"，"武汉钢铁集团公司"简称为"武钢"，"东风汽车电子仪表股份有限公司"简称为"东风电仪")。第一种类型机构名简称同机构名全称的识别过程一样，而第二种类型机构名简称的识别是有困难的。对于第二种类型机构名简称，大多数情况下其全称在简称上下文中出现，这时可以在识别出全称的基础上，利用简称字符串和全称的字对应关系识别出简称[22]；也有一些常用的机构名，其简称出现的上下文中不出现全称，这时可以基于一个常用机构名列表，利用简称和全称的字对应关系识别出简称[27]。也有一些工作不只是研究机构名简称识别，而是研究所有缩略语识别问题。孙栩等认为，"缩略语内部存在概念结构，在形成缩略语时，人们试图通过字符来最终表达每一个概念，虽然新缩略语一直被制造，其概念结构可能并未改变"。在这个思路下，他把在缩略语中出现的汉字聚为 30 类(每

个类对应一个概念），然后利用统计语言模型判断一个未登录词串是否构成缩略语。这种方法也可以用于机构名简称的预测[28]。尽管已经有这样一些方法，但是机构名简称的识别仍然是一个难点，值得深入研究。

3　命名实体排歧

命名实体的歧义指的是一个命名实体指称项可对应到多个命名实体概念，例如：给定命名实体指称项“华盛顿”，它既可以指华盛顿州，也可以指美国的第一任总统。在这种情况下，需要确定一个指称项真正指向的实体概念是什么，这就是命名实体排歧。

3.1　命名实体排歧的难点

命名实体排歧的任务可由如下定义：给定一个描述所有实体概念集合 $E = \{e_1, e_2, \cdots, e_m\}$，在一个数据集合 D（文档、网页、论文等）中，所有实体概念都以指称项形式存在，这些指称项可以是命名性、代词性和名词性的，其中命名性指称项还可以是全称、简称、别称以及不同语言的称呼形式。所有指称项构成集合 $R = \{r_1, r_2, \cdots, r_n\}$，命名实体消歧的任务就是基于上下文环境和世界知识将这个集合中的所有元素划分到其对应的实体概念上去。在给定实体概念集合的情况下，命名实体消歧的任务在于构建分类算法。但是，很多情况下，实体概念集合是未知的，这时除了构建分类算法之外，还需要首先检测出底层的实体概念集合。

随着目前信息规模的快速增长，命名实体的歧义性问题越来越严重，特别是在 Web 的环境下，命名实体的歧义显得尤其明显。除了全称、简称、别称、不同语言的称呼形式外，命名实体的歧义还可以由以下几方面原因引起：拼写错误、转喻的使用、属性替代等。

3.2　系统评测和技术水平

目前，在命名实体消歧的领域，评测主要有 UNED 组织的 Web People Search Evaluation (WePS)[29]，目前已经进行到第二届。第一届 (WePS1) 作为 SEMEVAL2007 的子任务举行，目前进行的第二届 (WePS2) 将作为 WWW 的一个 workshop 举行。WePS 的任务主要集中在 Web 环境中的人名消歧任务：给定一个包含某个有歧义人名的网页集合，按照网页中人名指称项所指向的人物概念来对网页进行聚类，并抽取网页中关于某个任务概念的特定属性来辅助进行人名消歧。

WePS 发布的评测语料由三部分构成：第一部分是有关人名的网页集合（Yahoo 搜索引擎的前 150 个网页），目前有共计 109 个人名对应的近 10 000 个网页；第二个部分是每个人名的聚类结果；第三个部分是给定网页中对应人物概念的特定属性标注，共有别名、出生日期、工作等 18 个属性。所有语料由人工标注。

在 WePS-1 任务中，共有 29 家单位参与了评测，其中 16 个单位最终提交了结果。评测采用 Purity（相当于 precision）、Inverse Purity（相当于 recall）及以上两个评测指标的调和值 F 值来衡量系统的效果。下表为 16 家参与单位的最终评测结果[29]。

表 7　WePS-1 测试结果

		Macro-averaged Score			
		F-measures			
rank	team-id	α=,s	α=,2	pur	Inv_Pur
1	CU_COMSEM	,78	,83	,72	,88
2	IRST_BP	,75	,77	,75	,80
3	PSNUS	,75	,78	,73	,82
4	UVA	,67	,62	,81	,60
5	SHEF	,66	,73	,60	,82
6	FICO	,64	,76	,53	,90
7	UNN	,62	,67	,60	,73
8	ONE-IN-ONE	,61	,52	1,00	,47
9	AUG	,60	,73	,50	,88
10	AWAT-IV	,58	,64	,55	,71
11	UA-ZSA	,58	,60	,58	,64
12	TITPI	,57	,71	,45	,89
13	JHU1-13	,53	,65	,45	,82
14	DFK12	,50	,63	,39	,83
15	WIT	,49	,66	,36	,93
16	UC3M_13	,48	,66	,35	,95
17	UBC-AS	,40	,55	,30	,91
18	ALL-IN-ONE	,40	,58	,29	1,00

在采用的方法方面，在提交了结果的 16 个系统中，15 个系统采用了基于文本向量空间的聚类或

pair-wise classification 方法[29]，另外一个系统采用了基于社会网络的聚类方法。在基于文本向量空间的方法中，主要差异在如何选取特征上。参与评测的系统主要使用人名周围出现的词以及命名实体来构建人名的特征向量。在划分的方法上面，大部分单位使用了 agglomerative clustering 方法[29]或 pair-wise classification 方法[29]，并基于经验来选取相似度或类别数目作为划分的终止条件。

3.3 现有的方法

通常，实体消歧可以利用两方面的知识：一个是上下文信息，如实体周围出现的词语；另一个是世界知识，如实体的分类体系、实体的关联等等。根据实体概念集合是否已知，命名实体消歧的任务又有所不同。在给定实体概念集合的情况下，命名实体消歧的任务在于构建分类算法；在实体概念集合未知的情况下，除了构建分类算法之外，还需要首先检测出底层的实体概念集合，也就是判断两个实体指称项是否指向相同的实体概念。在大多数的情况下，底层实体概念集合是未知的，以下主要介绍底层实体概念集合未知的情况下命名实体消歧的主要方法。

第一种方法是基于文本向量空间的聚类方法[30-38]。给定一个实体指称项 r 及其上下文，r 的歧义在一定程度上可被上下文所消解，两个指向相同实体概念的指称项的上下文在一定程度上具有相似性。一个实体指称项可以表示成在其上下文中出现的词和命名实体的文本向量。基于实体的文本向量表示，两个实体指称项的距离被表示成两个文本向量的距离。最终，基于实体指称项之间的距离，使用聚类方法来确定哪些实体指称项最终指向相同的实体概念。

第二种方法是基于社会网络的方法[39-43]。这种方法认为，实体指称项的意义被其相关联的实体所决定。该方法的第一步是构建社会网络，即通过实体之间存在的关系将关联的实体指称项相互链接，这样，实体指称就表示成社会网络中的一个节点。然后，通过一定方法(如 Random Walking 来计算两个实体指称项在社会网络之间的距离，如果距离低于一个阈值，则认为两个实体指称项指向同一个实体概念[40,42]。

第三种方法是基于分类的方法[44]。该方法认为，类别信息是标识一个实体概念的显著特征，如果把一个实体指称项将其与类别信息关联起来，就可以在某种程度上减少该实体指称项的歧义。如果能够选择一个合适的分类体系，歧义将会减少到一个极少的程度。对于一个实体指称项，H. P. Han 等首先通过网络挖掘验证的方法，确定可能与该指称项相关的若干个类别作为该指称项可能指向的实体概念集合；然后使用与这些类别相关的网页作为训练集来构建一个对该实体进行排歧的分类器；最终通过将实体指称项分派到指定的类别中来消除歧义[44]。

第四种方法是基于生成模型的方法[38]。该方法通过计算一个实体概念 e 到一个实体指称项 r 的生成概率，并选择以最大化概率生成实体指称项的实体概念 $e = \arg_e \max p(r \mid e)$ 作为实体指称项所指向的实体概念来消歧。

3.4 存在的问题

尽管目前存在着各种不同的命名实体消歧的方法，但是，命名实体消歧还存在着很多问题，其中，最突出的两个问题是：

(1) 命名实体排歧算法在大规模语料上应用的问题

基于聚类的命名实体排歧方法的时间复杂度与文档集合的大小成平方的关系，所以，当文档集合达到海量规模时，聚类时间需求急剧上升。这导致在大规模语料上聚类算法的时间复杂度不能达到实际应用的需求。在海量语料的情况下，通常也不可能以人工方法维护一个包含所有实体概念的实体数据库，这使得给定目标实体概念集合的实体排歧方法无法使用。要在海量语料中应用基于目标实体概念集合的分类方法来解决命名实体排歧问题，需要能够自动地构建和维护一个目标实体概念库，这方面的工作需要加强。

(2) 算法的自适应性问题

基于聚类的命名实体排歧方法的性能取决于两方面：实体指称项之间的相似度度量算法以及聚类算法的终止条件。目前的相似度度量通常与语料类型密切相关，如在文本语料中两个实体的相似度是根据它们的上下文相似性决定的，而在网络数据中两个实体的相似度是根据它们在社会网络中的链接强度决定，这就导致实体排歧算法在新的应用环境中无法使用。另一方面，目前的聚类终止条件更多地取决于经验选择，不同数据集上的终止条件可能有很大的不同。以上两方面的原因导致基于聚类的实体排歧方法在应用到新的领域时性能往往有明显

下降，常常需要重新定义新的相似度度量、训练参数和调节终止条件。

而目标实体概念集合已知的实体排歧方法需要给定目标实体概念及其特征，如果更换领域则需要知道新领域的实体概念集合，这也使得给定目标实体概念集合的实体排歧方法难以应用到新的领域。

4　命名实体的跨语言关联

实体概念在文本中的指称项有三种形式：命名性指称、名词性指称和代词性指称，实体关联的任务是把指称同一实体概念的所有指称链接起来[45]。命名性指称项的关联可以有多种类型，例如：全称、简称、别称、多语言命名性指称的关联等。本文重点讨论汉英双语命名性指称项的关联方法。

4.1　命名实体跨语言关联的难点

有关命名实体的跨语言关联，最直接的方法就是实体翻译。针对命名实体的翻译/音译已经有一些研究[46-51]，但是都不成熟。有关人名，中文人名和英文人名对应一般在发音(phonetic)上相似，比如：迈克尔·欧文/Michael Owen。对于中文人名，一般它的英文翻译就直接是其拼音序列，比如：胡锦涛/Hu Jintao，当然也有例外；对于地名来说，部分中英文地名对应在发音上存在相似性，如：渥太华/Ottawa；部分在意义上存在相似性，如：中非共和国/Central African Republic；还有一部分既包含音译又包含意译，如：小落矶山/Little Rocky Mountains；对于机构名，大多数中英文机构名对应的组成部件是对应的意译，比如：海关总署/General Administration of Customs，也有一部分中英文机构名除了存在意义上的翻译以外，其组成部件还包含发音相似的部分，比如：康奈尔大学/Cornell University。

(1) 音译的难点

人名一般是音译，地名和机构名中的名字部分也需要音译。近年来，音译研究受到越来越多的关注，特别是当音译涉及的两种语言的字符集差异比较大的情况(例如：英文和中文这两种语言)。机器自动音译通常分为两个方向：正向音译和反向音译。给一个双语人名对(s,t)，其中s代表源语言人名，t代表目标语言人名，正向音译是把s翻译成t，反向音译是把t翻译成s。例如："Clinton→克林顿"是正向翻译，"克林顿→Clinton"是反向音译。反向音译是人名音译的难点，主要困难在于：①传统的统计翻译模型根据语言模型来选择最可能的翻译结果，这种方法在意译任务中是有效的，但是在音译任务中效果不明显[52]。②反向音译比正向音译要更难，其中一个原因是：在正向音译过程中，源语言人名中的很多不发音音节信息已经丢失了，反向音译过程要恢复出这些丢失的音节非常困难。例如：当"Campbell"正向音译为"坎贝尔"时，"p"的发音信息已经丢失了，反向音译要恢复出"p"来很困难。

(2) 地名翻译的难点

地名翻译的难点，除了上面提到的命名部分需要音译的问题之外，还有命名部分和关键词部分的划分问题。例如，待翻译的中文地名是"维多利亚瀑布"，如何确定哪些是命名部分需要音译，哪些是关键词部分需要意译。地名的翻译同样存在习惯性用法的问题。总的来看，因为地名还是有限的，构造也比较简单，因此相对于人名翻译和机构名翻译，地名翻译要简单一些。

(3) 机构名翻译的难点

机构名的翻译难度更大，包含音译问题、音译部分和意译部分的区分问题、习惯性用法问题，并且它的结构更加复杂。

给定一个中文的机构名，在翻译的时候，首先需要确定其中的每个中文词或者短语应该选择哪个对应的翻译(一个中文词或者短语往往可以对应多个翻译)，而且需要确定这些翻译词语在构成目标语言机构名时应该摆放的位置。传统的机器翻译系统不是专门针对机构名翻译而设计的，没有充分利用机构名在结构上的特性，因此对机构名翻译的效果不好。

另外，机构名翻译也有译文是否符合习惯的问题。比如：汉语机构名"中国银行"只能翻译成英文"Bank of China"，而非"China Bank"。虽然"China Bank"也是意义上正确的翻译，但是由于"Bank of China"是官方翻译，并且已经成了习惯，所以"中国银行"对应的正确英文翻译只能是"Bank of China"。

4.2　系统评测和性能

与命名实体翻译相关的评测是与 ACE-2007 一同举办的 Entity Translation (ET)评测[53]。参评系统的输入是源语言文本(汉语或者阿拉伯语)，输出是该文本中出现的命名实体的英文列表。ET 需要翻译的实体类型同 ACE 实体检测和跟踪任务

(Entity Detection and Tracking)定义的实体类别相同,即人名、地名、机构名、地理政治实体、设施、运载工具和武器。按照 ACE 定义,实体概念的指称项有三种形式:命名性指称、名词性指称和代词性指称。ET 任务要求对输入文本中以上七种实体类别的三种指称形式进行翻译,但是不要求指出各个指称项的具体出现位置。ET 的测试语料来源于 LDC,分新闻和博客两种类型,分两个测试集,22 500 词的源语言文本和 15 000 词的源语言文本。评测指标涉及实体指称项的召回率以及翻译准确率。ET 评测结果不公开。

4.3 现有的方法

有关双语命名实体的关联,目前有三种模式:①给定源语言实体名,用机器翻译方法直接进行翻译,得到目标语言实体名;②给定源语言实体名,用网页挖掘辅助翻译的方法找出目标语言的实体对应;③源语言实体名未知,批量地从语料库或网页中抽取实体翻译对应。

(1) 给定源语言实体名,用机器翻译方法直接进行翻译

邹波等针对英汉人名翻译问题,系统地比较了以下 5 种音译模型在进行英汉人名音译时的性能,包括:基于记忆的学习方法、最大熵模型、条件随机场模型、基于短语的统计翻译模型和基于双语 N-gram 的统计翻译模型。实验表明:①以上 5 种模型的翻译性能都不高;②5 种模型的差距不明显,而且它们的翻译结果重合度很高;③正确结果大多数会出现在结果列表中,但是很多出现在靠后的位置。这可能预示着单纯用统计方法进行英汉人名音译在方法上是不足够的,需要求助于别的手段获取更好的音译结果[51]。

陈钰枫等系统研究了汉语机构名的组成特点,采用一种"语块"单位作为机构名的组成部分,然后根据汉英机构名的对齐特点定义了三类语块,并归纳了这三类语块在翻译时的排序规律,在此基础上提出了一种基于结构的汉语机构名英译方法。首先对汉语机构名进行语块自动切分,然后采用同步上下文无关文法实现语块的翻译和调序。实验证明,这种方法对于汉英机构名翻译是有效的,可以达到较高的正确率。另一方面,实体翻译模型的性能在很大程度上依赖于实体对齐的效果。陈钰枫等研究发现,双语实体的识别错误极大地限制了对齐性能,而普遍采用的对齐特征无法有效地克服实体识别错误带来的影响。为了解决这个问题,他们提出了双语实体识别与对齐相结合的实现方法,并提出基于翻译比率和类别约束的双语实体对齐方法。实验证明这种对齐方法提高了汉英实体的对齐性能[54,67]。

(2) 给定源语言实体名,用网页挖掘辅助翻译方法找出目标语言的实体对应

在机器自动音译不能得到好的翻译效果的情况下,研究人员尝试使用网络挖掘辅助翻译的方法从网页中找出源语言实体的目标语言对应。

在音译方面,J. H. Wang、P. J. Cheng、M. Nagata 和 Y. Zhang 等分别利用挖掘混合语言网页辅助翻译的方法进行音译[55-58]。L. Jiang 研究如何利用网络挖掘提升英汉人名音译的效果[59-60]。首先使用级联隐马尔可夫模型将英文人名的字母序列转化为汉语拼音序列;第二步,利用英文人名作为查询词搜索中文网页,并从中抽取出与英文人名具有较强统计关联度的 N-gram 作为翻译候选;第三步,利用第一步的音译结果作为锚点字对第二步网络挖掘得到的翻译候选进行过滤;最后,基于发音相似度、上下文特征、共现强度等特征,利用最大熵模型选出最优翻译对应。实验结果显示,比起单纯的音译方法和网络挖掘方法,音译和网络挖掘方法的综合使用能显著提升英汉人名音译的性能。F. Yang 等针对汉英人名反向音译任务进行研究,提出挖掘英语单语网页辅助统计音译的方法,重点解决统计音译方法的"正向音译过程中源语言人名的不发音音节的丢失问题"和"统计翻译方法很难获得不符合一般翻译规律的翻译结果的问题"。F. Yang 的方法由两个阶段组成。第一个阶段中,统计音译结果被划分成音节,然后将这些音节组成查询,利用基于音节的搜索过程从一个大规模 Web 词典中搜索与音译候选相似的单词,从而提高召回率;第二个阶段中,将矫正过的音译候选作为查询在 Web 中提取其上下文信息和点击率信息,然后利用 AdaBoost 判断其是否是一个正确的音译。这个阶段可以调整每个音译候选的得分,使之更合理,从而提高音译的精确率。实验结果显示,这种方法可以将 top-100 召回率从 41.73%提高到 59.28%,top-5 精确率可以从 19.69% 提高到 52.19%[61]。

在机构名翻译方面,杨帆采用一种利用启发式网络挖掘和不对称对齐技术的汉英机构名翻译方法,对于输入的汉语结构名,分以下 4 个步骤从混合语言网页中寻找其英文翻译。首先是机构名的分词:在对汉语机构名进行组块划分(地名组块、命名

组块、修饰成分组块和关键词组块)的基础上，对汉语机构名进行分词。这种基于组块划分的汉语机构名分词方法可以在组块划分阶段划分出大多数未登录词(OOV)，使得分词效果得到提高；第二步是包含汉英实体翻译对应的混合语言网页的搜索：根据机构名的分词结果和自动翻译结果，综合考虑一个词语在汉语机构名中的信息量以及它的翻译可信度来确定它是否进入查询，从而搜索到既包含输入的汉语机构名、又包含它的英语翻译对应的网页；第三步是英文实体翻译对应的抽取：利用一种非对称的对齐方法——扩展的KM算法，直接对源语言机构名和目标语言英文词串进行对齐，从而从英文词串中抽取出与源语言机构名对齐效果最好的词串片段作为输出。与传统的对齐方法相比，这种方法可以避免命名实体识别的很多错误，从而将汉英机构名翻译的L1值从18.35%提高到48.79%(top 1)和53.61%(top 5)[62]。

还有一些研究不只局限于实体对应的抽取，而是利用网页挖掘辅助翻译的方法找出固定短语的翻译对应[63-65]，这些方法也可以借鉴。

(3) 源语言实体名未知，批量地从语料库或网页中抽取实体翻译对应

根据语料种类的不同可以分为：基于平行语料的抽取、基于可比语料的抽取和基于混合语料的抽取等。

• 基于平行语料的实体翻译对应抽取

基于句子级平行语料的实体翻译对应抽取的工作有不少[64,66,72]。他们的方法通过融合多种特征，包括音译特征、意译特征、命名实体标注特征或者一些平行语料库特有的特征(如位移、同现特征)，将中英文句对中的实体进行对齐。这三篇文章所介绍的方法都取得了良好的效果。F. Huang的研究结果显示，精确度达到73.8%，召回率达到90.5%，F值为81.3%[64,72]。D. Feng的形容结果显示，精确度63.6%，召回率82.4%，F值71.8%[66]。

平行语料为实体翻译对应的抽取提供了很多有用信息，然而该方法具有很大的局限性。因为平行语料难以获得，需要人工参与，很费时，大规模平行语料不易建立，因此从平行语料中大规模地抽取命名实体对应并不现实。

• 基于可比语料的抽取

从可比语料中获取双语实体对应，是近年来兴起的一种方法。所谓可比语料，是指相互不为翻译、篇章之间不对齐、但是主题相同的双语文档集合，比如从不同的新闻机构获得的语言不同但是时间段相同的新闻就可以构成可比语料[68]。从可比语料中抽取命名实体对的最大优点就是这种语料库比较容易获得。

R. Sproat等以英文为源语言、中文为目标语言，以发音相似为主要特征，从可比语料库中抽取英中命名实体对应，测试结果表明：MRR可以达到0.3[64]。相关的工作还有文献[70]，他们从可比语料中抽取新词的翻译。

从可比语料中抽取双语实体对应，其最大的优点是这种语料容易获得，但是这种方法同样存在很多问题。①与双语对齐的语料库相比，可比语料中对双语实体对应的约束更少，搜索的空间更大，而可利用的特征要少很多，因此抽取难度更大，噪音很多，精确率不高；②双语平行语料中，所有源语命名实体在目标语言文本中都有其翻译对应；而在可比语料中，这种对应关系无法保证，存在大量的没有翻译对应的实体。这种现象严重影响了召回率。

• 基于混合语言语料的抽取

一些网页中存在的混合语言文本也是双语实体对应抽取的信息来源。D. K. Lin等利用词对齐方法从海量网页挖掘以括号标识的汉语固定短语及其英语翻译对应(其中包括很多实体)。首先，通过括号模式从网页中提取的局部可比文本"……招聘中国银行(Bank of China)宣讲会……"，其中"中国银行"及其翻译对应"Bank of China"同时出现；然后采用对齐的方法，获得"中国银行"和"Bank of China"的对齐概率，通过滤掉对齐概率低的对应，提高了翻译对应获取的精确率[71]。D. K. Lin的方法通过一种可靠的模式，找到一个文本片段，在该片段中，英文实体及其中文翻译对应同现的概率很高。该方法的出发点是寻找汉英实体对应同现的文本，从而可以将后续的对齐工作局限在该局部文本中，简化了计算量。但是，由于这种方法要求汉英短语翻译对应在上下文中以括号方式"同现"，因此会丢失掉很多以其他模式出现的短语翻译对应(这些短语翻译对应均在网页中出现过，但没有位置上的显著联系)[71]。

虽然基于混合语言语料的双语实体对应抽取方法准确性较好，而且可以抽取出那些没有翻译线索的双语实体对应来，但是，混合语言网页资源是有限的，只能解决一部分问题。

4.4 需要解决的问题

尽管目前存在着各种不同的方法，但是，命名实

体跨语言关联还存在着很多问题，其中，需要重点研究的问题之一是如何从海量、冗余、异构、不规范、含有大量噪声的网页中抽取出可靠的翻译对应。研究显示，单纯的机器翻译方法对于实体翻译来说是很难的。另一方面，各种实体及其翻译对应又以各种形式出现在网络上，我们需要做的就是找到一种链接方法，把它们关联起来。这涉及到两个方面的问题：首先是在海量网络资源上挖掘更多的信息源，包括：教科书、学术论文以及 Deep Web 等等，扩大双语实体对应的信息来源。另一方面，从海量、冗余、异构、不规范、含有大量噪声的网页中抽取出可靠的翻译对应，需要自然语言处理、信息检索和网络挖掘技术的密切配合，包括：命名实体的分析、翻译、查询构造、网页搜索、模式匹配、双语对齐等多种技术。

5 下一步的工作

命名实体是文本中承载信息的重要语言单位，命名实体的识别和分析研究在网络信息抽取、网络内容管理和知识工程等领域占有非常重要的地位。本文重点介绍了命名实体识别、排歧和跨语言关联等任务的研究现状，包括难点、评测、现有的方法、技术水平，并对下一步需要重点解决的问题进行和分析和讨论。目前的命名实体识别、排歧和跨语言关联的技术水平还远远不能满足大规模真实应用的需求，还需要更加深入地研究。从研究方法上来讲，命名实体识别、排歧和跨语言关联的研究要突破自然语言处理领域的限制，面向真实的互联网应用，研究面向海量、冗余、异构、不规范、含有大量噪声的网页的实体识别、排歧和跨语言关联技术。除了本文重点讨论的实体识别、排歧和跨语言关联技术外，实体属性抽取、实体关系检测等技术的研究也同样具有重要的学术意义和广泛的应用价值，需要深入研究。

致谢：感谢与作者合作的研究生们对本文的贡献，特别是吴友政（命名实体识别和分类）、刘非凡（产品名识别、实体同指消解）、杨帆（实体的跨语言关联）、陆敏（实体的跨语言关联）、邹波（实体的音译）、韩先培（实体排歧、实体属性抽取）、刘康（实体属性抽取、命名实体识别系统的自适应）等。感谢国家语言资源监测与研究中心的老师和同学对本文研究工作的支持。

参考文献：

[1] NIST. The ACE 2007 (ACE07) Evaluation Plan: Evaluation of the Detection and Recognition of ACE Entities, Values, Temporal Expressions, Relations, and Events [EB/OL]. [2007]. http://www.nist.gov/speech/tests/ace/2007/doc/ace07-evalplan.v1.3a.pdf.

[2] Nancy A. Chinchor. Overview of MUC-7/MET-2 [C]//Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, 1998.

[3] Gina-Anne Levow. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition [C]//Proceedings of the Fifth SigHAN Workshop on Chinese Language Processing, Sydney: Association for Computational Linguistics, 2006:108-117.

[4] A. Mikheev, C. Grover, Moens M. Description of the LTG System Used for MUC-7 [C]//Proceedings of 7th Message Understanding Conference (MUC-7), Fairfax, Virginia, 1998.

[5] 863计划中文信息处理与智能人机接口技术评测组. 2004年度863计划中文信息处理与智能人机交互技术评测：命名实体评测结果报告 [R]. 北京：863计划中文信息处理与智能人机接口技术评测组，2004.

[6] Ralph Grishman, Beth Sundheim. Design of the MUC-6 evaluation [C]//Proceedings of 6th Message Understanding Conference, Columbia, MD, 1995.

[7] G. R. Krupka, K. Hausman. IsoQuest. Inc.: Description of the NetOwl TM Extractor System as Used for MUC-7 [C]//Proceedings of the 7th Message Understanding Conference. (MUC-7), Fairfax, Virginia, 1998.

[8] W. J. Black, F. Rinaldi, D. Mowart. FACILE: Description of the NE System Used for MUC-7 [C]//Proceedings of the 7th Message Understanding Conference. (MUC-7), Fairfax, Virginia, 1998.

[9] Youzheng Wu, Jun Zhao, Bo Xu, et al. Chinese Named Entity Recognition Model Based on Multiple Features [C]//Proceedings of Human Language Technology Conference & Conference on Empirical Methods in NLP (HLT/EMNLP), Vancouver, B. C., Canada: Association for Computational Linguistics, 2005: 427-434.

[10] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge [C]//Proceedings of the Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition: Combining Statistical and Symbolic Models, Sappora, Japan: Association for Computational Linguistics, 2003:

65-72.

[11] Daniel M. Bikel, Scott Miller, Richard Schwartz, et al. Nymble: a High-Performance Learning Name-finder [C]//Proceedings of Fifth Conference on Applied Natural Language Processing, New York, NY: Association for Computational Linguistics, 1997: 194-201.

[12] Jian Sun, Jianfeng Gao, Lei Zhang, et al. Chinese Named Entity Identification Using Class-based Language Model [C]//Proceedings of the 19th international conference on Computational linguistics (COLING 2002), Taipei: Association for Computational Linguistics, 2002: 1-7.

[13] Huaping Zhang, Qun Liu, Hongkui Yu, et al. Chinese Named Entity Recognition Using Role Model [J]. Special issue "Word Formation and Chinese Language processing" of the International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(2): 29-60.

[14] A. Borthwick. A Maximum Entropy Approach to Named Entity Recognition [D]. New York: New York University. 1999.

[15] J. Aberdeen, J. Burger, D. Day, et al. MITRE: Description of the Alembic system used for MUC-6 [C]//Proceedings of the 6th Message Understanding Conference (MUC-6), Columbia, Maryland, Association for Computational Linguistics, 1995: 141-155.

[16] S. Sekine, R. Grishman, H. Shinou. A decision tree method for finding and classifying names in Japanese texts [C]//Proceedings of the Sixth Workshop on Very Large Corpora, Canada: Association for Computational Linguistics, 1998: 171-178.

[17] Michael Collins, Yoram Singer. Unsupervised models for named entity classification [C]//Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, USA: Association for Computational Linguistics, 1999: 100-110

[18] Michael Collins. Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron [C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia: Association for Computational Linguistics, 2002: 489-496.

[19] 孙茂松，张维杰. 英语姓名译名的自动识别 [C]//陈力为. 计算语言学研究与应用. 北京：北京语言学院出版社，1993：144-149.

[20] 孙茂松，黄昌宁，高海燕，等. 中文姓名的自动辨识 [J]. 中文信息学报，1994，9(2)：16-27.

[21] 陈慧. 基于 DCC 动态流通语料库的中文组织名考察与研究[D]. 北京：北京语言大学，2008.

[22] 吴友政. 汉语问答系统关键技术研究 [D]，北京：中国科学院自动化研究所，2006.

[23] 刘康，赵军. 基于"产生/判别"混合模型的分类器领域适应性问题研究 [C]. 全国模式识别学术会议论文集，北京：中国自动化学会，中国科学院自动化研究所，2008：7-12(最佳学生论文).

[24] ZHAO Jun, LIU Feifan, Product Named Entity Recognition in Chinese Texts [J]. International Journal of Language Resource and Evaluation (LRE), 2008, 42(2): 132-152.

[25] Satoshi Sekine, Kiyoshi Sudo, Chikashi Nobata, Extended Named Entity Hierarchy [C]//Proceedings of The Third International Conference on Language Resources and Evaluation, Spain, 2002:1818-1824

[26] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, et al. Web-Scale Named Entity Recognition [C]// James G. Shanahan, et al. (Eds.) Proceedings of ACM 17th Conference on Information and Knowledge Management, California, 2008: 123-132.

[27] Fan YANG, Jun ZHAO. CRFs-Based Named Entity Recognition Incorporated with Heuristic Entity List Searching [C]//Proceedings of the Sixth SigHAN Workshop on Chinese Language Processing, Hyderabad, India, 2008: 171-174.

[28] 孙栩. 基于机器学习的汉语缩略语识别与预测 [D]. 北京：北京大学，2007.

[29] Javier Artiles, Julio Gonzalo, Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task [C]// Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague: Association for Computational Linguistics, 2007: 64-69.

[30] Amit Bagga, Breck Baldwin. Entity-Based Cross-document coreferencing using the vector space model [C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics , Canada: Association for Computational Linguistics, 1998: 79-85.

[31] Enrique Amigo, Julio Gonzalo, Javier Artiles, et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints [J]. Information Retrieval, DOI 10. 1007/S10791-008-9066-8, 2008.

[32] Michael Ben Fleischman, Eduard Hovy. Multi-Document Person Name Resolution [C]//Proceedings of the Workshop on Reference Resolution and its Applications (Held in cooperation with ACL-2004), Spain: Association for Computational Linguistics , 2004: 1-8.

[33] Nina Wacholder, Yael Ravin, Miscook Choi. Disambiguation of Proper Names in Text [C]//Proceedings of the fifth conference on Applied natural language processing, Washington: Association for Computational Linguistics, 1997: 202-208.

[34] Silviu Cucerzan. Large-Scale Named Entity Disambig-

uation Based on Wikipedia Data [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Prague,, Czech Republic: Association for Computational Linguistics, 2007: 708-716.

[35] Ted Pedersen, Amruta Purandare, Anagha Kulkarni. Name discrimination by clustering similar contexts [C]//Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2005), Mexico, 2005: 226-237.

[36] Ted Pedersen and Anagha Kulkarni. An Unsupervised language independent method of name discrimination using second order co-Occurence Features [C]//Proceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2006), Mexico, 2006: 208-222.

[37] Ted Pedersen and Anagha Kulkarni. Unsupervised Discrimination of Person Names in web contexts [C]//Proceedings of Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2007), Mexico, 2007: 299-310.

[38] Xin Li, Paul Morie, Dan Roth. Semantic Integration in Text: From Ambiguous Names to Identifiable Entities [J]. AI Magazine, 2005, 26(1): 45-58.

[39] Bradley Malin, Edorado Airoldi, Kathleen M. Carley. A Network Analysis Model for Disambiguation of Names in Lists [J] . A Network Analysis Model for Disambiguation of Names in Lists, 2005, 11(2): 119-139.

[40] Einat Minkov, William W. Cohen, Andrew Y. Ng. Contextual Search and Name Disambiguation in Email Using Graphs [C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-2006), Washington, USA, 2006: 27-34.

[41] Josepth Hassell. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text [C]//Proceedings of 5th International Semantic Web Conference (ISWC-2006), Athens, USA, 2006: 44-57.

[42] Ron Bekkerman, Andrew McCallum. Disambiguating Web Appearances of People in a Social Network [C]//Proceedings of the 14th international conference on World Wide Web (WWW-2005), Japan. 2005: 463-470.

[43] Zhaoqi Chen, Dmitri V. Kalashnikov, Sharad Mehrotra. Adaptive graphical approach to entity resolution [C]//Proceedings of ACM IEEE Joint Conference on Digital Libraries, Canada, 2007: 204-213.

[44] Xianpei Han, Jun Zhao, Person Name Disambiguation Based on Web-Based Person Mining and Categorization, Submitted to Second Web People Search Evaluation Workshop in conjunction with WWW2009, Spain, 2009.

[45] ZHAO Jun, LIU Feifan, Linguistic Theory Based Contextual Evidence Mining for Statistical Chinese Co-reference Resolution [J]. Journal of Computer Science and Technology (JCST), 2007, 22(4): 608-617.

[46] Y. Al-Onaizan and K. Knight. Named Entity Translation [C]//Proceedings of the second international conference on Human Language Technology Research (HLT-2002), San Diego, CA, 2002: 122-124.

[47] Kevin Knight, Jonathan Graehl: Machine Transliteration [J]. Computational Linguistics, 1998 24(4): 599-612.

[48] S. Stalls and K. Knight. Translating Names and Technical Terms in Arabic Text [C]//Proceedings of the COLING/ACL Workshop on Computational approaches to Semitic Languages, Canada, 1998.

[49] H. Meng, W. K. Lo, B. Chen, et al. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval [C]//Proceedings of the Automatic Speech Recognition and Understanding Workshop, Trento, Italy, 2001.

[50] 陆敏. 汉英实体翻译与实体对抽取技术研究 [D]. 北京: 中国科学院自动化研究所, 2007.

[51] 邹波, 英汉人名音译方法研究 [D]. 北京: 中国科学院自动化研究所, 2008.

[52] Wei Gao. Phoneme-based Statistical Transliteration of Foreign Name for OOV Problem [D]. Hong Kong: The Chinese University of Hong Kong. 2004

[53] NIST, The Evaluation Plan for the ACE 2007: Pilot Evaluation of Entity Translation [EB/OL]. [2007]. Available at http://nist.gov/speech/tests/ace/2007/doc/ET07-evalplan-v1.8.pdf

[54] 陈钰枫. 汉英命名实体翻译及对齐方法研究, 北京: 中国科学院自动化研究所, 2008.

[55] Jenq-Haur Wang, Jei-Wen Teng, Pu-Jen Cheng, et al. Translating unknown cross-lingual queries in digital libraries using a web-based approach [C]//Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, Tuscon, AZ, USA , 2004: 108-116.

[56] Pu-Jen Cheng, Wen-Hsiang Lu, Jer-Wen Teng, et al. Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora [C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL-04), Spain, 2004.

[57] Masaaki Nagata, Teruka Saito, and Kenji Suzuki. Using the Web as a Bilingual Dictionary [C]//Proceedings of ACL 2001 Workshop on Data-driven Methods in Machine Translation, France, 2001: 1-8.

[58] Ying Zhang, Fei Huang, Stephan Vogel. Mining translations of OOV terms from the web through cross-lingual query expansion [C]//Proceedings of the 28th International ACM SIGIR, Brazil, 2005: 669-670.

[59] Jiang, L., Zhou, M., Chien, L., et al. Named Entity Translation with Web Mining and Transliteration [C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), Menlo Park, CA: International. Joint. Conferences. on. Artificial. Intelligence and AAAI Press, 2007:1629-1634.

[60] 蒋龙,周明,简立峰.利用音译和网络挖掘翻译命名实体 [J].中文信息学报,2007, 21(1): 23-29.

[61] Fan Yang, Jun Zhao, Bo Zou, et al. Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages [C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH: Association for Computational Linguistics, 2008: 541-549

[62] 杨帆,赵军. 基于启发式网络挖掘和非对称对齐的汉英机构名翻译方法 [R],北京：中国科学院自动化研究所模式识别国家重点实验室,2008.

[63] Yunbo Cao, Hang Li, Base Noun Phrase Translation Using Web Data and the EM Algorithm [C]//Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Taipei, 2002: 1-7.

[64] Fei Huang, Stephan Vogel, Alex Waibel, Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization [C]//Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual and Mixed-language Named Entity Recognition, Japan: Association for Computational Linguistics, ,2003:9-16.

[65] Gaolin Fang, Hao Yu, Fumihito Nishino. Chinese-English Term Translation Mining Based on Semantic Prediction [C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Australia: Association for Computational Linguistics, 2006: 199--206.

[66] Dong-Hui Feng, Ya-Juan Lv, Ming Zhou, A New Approach for English-Chinese Named Entity Alignment [C]//Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain: Association for Computational Linguistics, 2004: 372-379.

[67] Yufeng Chen, Chengqing Zong, A Structural-Based Model for Chinese Organization Name Translation [J]. ACM Transactions on Asian Language Information Processing (ACM TALIP), 2008, 7(1):1-30.

[68] Pascale Fung, Percy Cheung. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM [C]//Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain: Association for Computational Linguistics, 2004: 57-63.

[69] Richard Sproat, Tao Tao, Chengxiang Zhai. Australia, 2006 Named Entity Transliteration with Comparable Corpora [C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Australia: Association for Computational Linguistics, 2006: 73-80.

[70] Li Shao, Hwee Tou Ng, Mining New Word Translation from Comparable Corpora [C]//The 20th International Conference on Computational Linguistics, Switzerland, 2004: 618-624.

[71] Dekang Lin, Shaojun Zhao, Benjamin Van Durme, et al, Mining Parenthetical Translations from the Web by Word Alignment [C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008), Columbus, OH: Association for Computational Linguistics,2008: 994-1002.

[72] Fei Huang, Ying Zhang, Setphan Vogel. Mining Key Phrase Translations from Web Corpora [C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), Vancouvor, Canada: Association for Computational Linguistics, 2005: 483-490.

命名实体识别、排歧和跨语言关联

作者： 赵军， ZHAO Jun
作者单位： 中国科学院,自动化研究所,模式识别国家重点实验室,北京,100190
刊名： 中文信息学报 ISTIC PKU
英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING
年，卷(期)： 2009，23(2)
引用次数： 0次

参考文献(72条)

1. The ACE 2007 (ACE07) Evaluation Plan:Evaluation of the Detection and Recognition of ACE Entities,Values,Temporal Expressions,Relations,and Events 2007

2. Overview of MUC-7/MET-2 1998

3. The Third International Chinese Language Processing Bakeoff:Word Segmentation and Named Entity Recognition 2006

4. Description of the LTG System Used for MUC-7 1998

5. 2004年度863计划中文信息处理与智能人机交互技术评测:命名实体评测结果报告 2004

6. Design of the MUC6 evaluation 1995

7. Description of the NetOwl TM Extractor System as Used for MUC-7 1998

8. FACILE:Description of the NE System Used for MUC-7 1998

9. Chinese Named Entity Recognition Model Based on Multiple Features 2005

10. Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge 2003

11. Nymble:a High-Performance Learning Namefinder 1997

12. Chinese Named Entity Identification Using Class-based Language Model 2002

13. Chinese Named Entity Recognition Using Role Model 2003(2)

14. A Maximum Entropy Approach to Named Entity Recognition 1999

15. MITRE:Description of the Alembic system used for MUC-6 1995

16. A decision tree method for finding and classifying names in Japanese texts 1998

17. Unsupervised models for named entity classification 1999

18. Ranking Algorithms for Named Entity Extraction:Boosting and the Voted Perceptron 2002

19. 英语姓名译名的自动识别 1993

20. 中文姓名的自动辨识 1994(2)

21. 基于DCC动态流通语料库的中文组织名考察与研究 2008

22. 汉语问答系统关键技术研究 2006

23. 基于"产生/判别"混合模型的分类器领域适应性问题研究 2008

24. Product Named Entity Recognition in Chinese Texts 2008(2)

25. Extended Named Entity Hierarchy 2002

26. Web-Scale Named Entity Recognition 2008

27. CRFs-Based Named Entity Recognition Incorporated with Heuristic Entity List Searching 2008

28. 基于机器学习的汉语缩略语识别与预测 2007

29. The SemEval-2007 WePS Evaluation:Establishing a benchmark for the web People Search Task 2007

30. Entity-Based Cross-document coreferencing using the vector space model 1998
31. A comparison of extrinsic clustering evaluation metrics based on formal constraints 2008
32. Multi-Document Person Name Resolution 2004
33. Disambiguation of Proper Names in Text 1997
34. Large-Scale Named Entity Disambiguation Based on Wikipedia Data 2007
35. Name discrimination by clustering similar contexts 2005
36. An Unsupervised language independent method of name discrimination using second order co-Occurrence Features 2006
37. Unsupervised Discrimination of Person Names in web contexts 2007
38. Semantic Integration in Text:From Ambiguous Names to Identifiable Entities 2005(1)
39. A Network Analysis Model for Disambiguation of Names in Lists.A Network Analysis Model for Disambiguation of Names in Lists 2005(2)
40. Contextual Search and Name Disambiguation in Email Using Graphs 2006
41. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text 2006
42. Disambiguating Web Appearances of People in a Social Network 2005
43. Adaptive graphical approach to entity resolution 2007
44. Disambiguation Based on Web-Based Person Mining and Categorization 2009
45. Jun Zhao. Fei-Fan Liu Linguistic Theory Based Contextual Evidence Mining for Statistical Chinese Co-Reference Resolution[期刊论文]-计算机科学技术学报（英文版） 2007(4)
46. Named Entity Translation 2002
47. Machine Transliteration 1998(4)
48. Translating Names and Technical Terms in Arabic Text 1998
49. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval 2001
50. 汉英实体翻译与实体对抽取技术研究 2007
51. 英汉人名音译方法研究 2008
52. Phoneme-based Statistical Transliteration of Foreign Name for OOV Problem 2004
53. The Evaluation Plan for the ACE 2007:Pilot Evaluation of Entity Translation 2007
54. 汉英命名实体翻译及对齐方法研究 2008
55. Translating unknown cross-lingual queries in digital libraries using a web-based approach 2004
56. Creating Muhilingual Translation Lexicons with Regional Variations Using Web Corpora 2004
57. Using the Web as a Bilingual Dictionary 2001
58. Mining translations of OOV terms from the web through cross-lingual query expansion 2005
59. Named Entity Translation with Web Mining and Transliteration 2007
60. 蒋龙. 周明. 简立峰 利用音译和网络挖掘翻译命名实体[期刊论文]-中文信息学报 2007(1)
61. Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages 2008
62. 基于启发式网络挖掘和非对称对齐的汉英机构名翻译方法 2008
63. Base Noun Phrase Translation Using Web Data and the EM Algorithm 2002

64. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization 2003

65. ChineseEnglish Term Translation Mining Based on Semantic Prediction 2006

66. A New Approach for English-Chinese Named Entity Alignment 2004

67. A Structural-Based Model for Chinese Organization Name Translation 2008(1)

68. Mining Very-Non-Parallel Corpora:Parallel Sentence and Lexicon Extraction via Bootstrapping and EM 2004

69. Australia,2006 Named Entity Transliteration with Comparable Corpora 2006

70. Mining New Word Translation from Comparable Corpora 2004

71. Mining Parenthetical Translations from the Web by Word Alignment 2008

72. Mining Key Phrase Translations from Web Corpora 2005

相似文献(10条)

1. 期刊论文 李宁 XML--中文信息处理的变革之路 -中文信息学报2003,17(2)

本文从中文信息面临的问题出发,阐述了中文信息处理走Internet开放变革之路的必要性.文中还介绍了Intemet上已经开展的与中文信息处理相关的部分工作,重点论述了XML在中文信息处理方面的优势,指出以XML为基础的Web服务是分布式环境中文信息处理技术的发展方向.作者为此提出了一个中文信息处理服务体系框架的构想.

2. 期刊论文 张玮.孙乐.冯元勇.李文波.黄瑞红.ZHANG Wei.SUN Le.FENG Yuan-yong.LI Wen-bo.HUANG Rui-hong 词汇搭配和用户模型在拼音输入法中的应用 -中文信息学报2007,21(4)

中文输入法是中文信息处理的难题之一.随着互联网上中文用户的不断增加,中文输入法的重要性也变得日益突出.本文在对句子中长距离词汇依赖现象观察的基础上,抽取出语料库中的词汇搭配来获取长距离特征,并以此构建基于词汇搭配关系的拼音输入法系统;同时将词汇搭配的思想应用到拼音输入法的用户模型中,从而使我们的输入法系统能够辅助用户更加有效的输入.实验表明基于词汇搭配关系的改进方法对提高输入法的准确率有积极的作用.

3. 期刊论文 唐慧丰.谭松波.程学旗.TANG Hui-feng.TAN Song-bo.CHENG Xue-qi 基于监督学习的中文情感分类技术比较研究 -中文信息学报2007,21(6)

情感分类是一项具有较大实用价值的分类技术,它可以在一定程度上解决网络评论信息杂乱的现象,方便用户准确定位所需信息.目前针对中文情感分类的研究相对较少,其中各种有监督学习方法的分类效果以及文本特征表示方法和特征选择机制等因素对分类性能的影响更是亟待研究的问题.本文以n-gram以及名词、动词、形容词、副词作为不同的文本表示特征,以互信息、信息增益、CHI统计量和文档频率作为不同的特征选择方法,以中心向量法、KNN、Winnow、Naive Bayes和SVM作为不同的文本分类方法,在不同的特征数量和不同规模的训练集情况下,分别进行了中文情感分类实验,并对实验结果进行了比较,对比结果表明:采用BiGrams特征表示方法、信息增益特征选择方法和SVM分类方法,在足够大训练集和选择适当数量特征的情况下,情感分类能取得较好的效果.

4. 会议论文 吕肖庆.北京大学文字信息处理技术国家重点实验室.尹江红.北京大学文字信息处理技术国家重点实验室.汪宾成.北京大学文字信息处理技术国家重点实验室.张建国.北京大学文字信息处理技术国家重点实验室.赵学亮.北京大学文字信息处理技术国家重点实验室.任力.北京大学文字信息处理技术国家重点实验室 超大字库及其相关技术的研制 2002

自计算机发明以来,汉字集合的选择、组织形式、特别是汉字编码问题,曾长期困扰着中文信息科技的发展.直到八十年代初,正式确立的中文简体字国家标准(GB2312)后,中文信息才有了统一的交换平台,应用软件也才得以蓬勃发展.但是,由于硬、软件环境的限制,以及编码工作本身的复杂性,虽然继GB2312标准之后推出了扩展标准(GBK)和GB18030,但常规的应用软件只能处理2万多个汉字,即使普通人在使用电脑时都可能遇到一些生僻字无法处理,比如人名、地名以及一些专用名词,而面对浩如烟海的中华古籍,2万余字的处理能力远远不够,长久以来一直让编纂辞书、整理古籍的专业人士扼腕痛惜.为了让源远流长的华夏文明能够凭借先进的计算机技术发扬光大,我们研制了超大字库及其相关的应用技术,不仅彻底解决了大量汉字的编码、显示问题,还经过长期积累,探索出超大字库录入的全新方法,并配备了排版、检索等工具.在中文信息处理方面为专业出版单位开拓了更为广阔的应用领域.近年来,该成果已不仅成功应用于古籍、辞书的编排与印刷,而且在医疗保险、户籍管理和历史档案的检索等方面,建立起了高水信的应用系统.

5. 期刊论文 李茹.王文晶.梁吉业.宋小香.刘海静.由丽萍.LI Ru.WANG Wen-jing.LIANG Ji-ye.SONG Xiao-xiang.LIU Hai-jing.YOU Li-ping 基于汉语框架网的旅游信息问答系统设计 -中文信息学报2009,23(2)

该文借助汉语框架网(Chinese FrameNet,简称CFN)在语义表达方面的独特优势,探讨用本体描述语言建立面向特定领域的汉语框架语义知识库,并且以旅游交通领域中问答系统设计为例分析方法的有效性.方法中首先利用TREC分类与本体分类相结合的方式为查询问句分类,然后提出基于CFN的问句分析策略,通过CFN语义分析得到问句中三元组:语义谓词、语义主体和语义客体,在问句分析的基础上从旅游本体知识库中对答案进行抽取并对答案处理,同时用本体编辑工具Protégé编码,实验证实方法是有效的.

6. 期刊论文 江敏.肖诗斌.王弘蔚.施水才.JIANG min.XIAO Shi-bin.WANG Hong-wei.SHI Shui-cai 一种改进的基于《知网》的词语语义相似度计算 -中文信息学报2008,22(5)

中科院刘群的基于<知网>的词语相似度计算是当前比较有代表性的计算词语相似度的方法之一.在测试中我们发现对一些存在对义或反义的词语与同义、近义词语一样具有较高的相似度,一些明显相似的词反而相似度较低,如"美丽"与"贼眉鼠眼"的相似度为0.814 815,与"优雅"的相似度为0.788 360,"深红"与"粉红"的相似度仅为0.074 074,这将不利于进行词语的极性识别.基于文本情感色彩分析的需要,把词语相似度的取值范围规定为[-1,+1],在刘群论文的基础上,进一步考虑了义原的深度信息,并利用<知网>义原间的反义、对义关系和义原的定义信息来计算词语的相似度.在词语极性识别实验中,得到了较好的实验结果:P值为99.07%,R值为99.11%.

7.期刊论文 乐明.YUE Ming 汉语篇章修辞结构的标注研究 -中文信息学报2008,22(4)

汉语篇章修辞结构标注项目CJPL采用大陆主要媒体的财经评论文章为语料,依据修辞结构理论(Rhetorical Structure Theory,RST),定义了以标点符号为边界的篇章修辞分析基本单元和47种区分核心性单元的汉语修辞关系集,并草拟了近60页的篇章结构标注工作守则.这一工作目前完成了对97篇财经评论文章的修辞结构标注,在较大规模数据的基础上检验了修辞结构理论及其形式化方法在汉语篇章分析中的可移用性.树库所带有的修辞关系信息以及三类篇章提示标记的篇章用法特征,可以为篇章层级的中文信息处理提供一些浅层语言形式标记的数据.

8.期刊论文 冯元勇.孙乐.李文波.张大鲲.FENG Yuan-yong.SUN Le.LI Wen-bo.ZHANG Da-kun 基于单字提示特征的中文命名实体识别快速算法 -中文信息学报2008,22(1)

近年来条件随机场(CRF)模型在自然语言处理中的应用越来越广泛.标准的线性链(Linear-chain)模型一般采用L-BFGS参数估计方法,收敛速度慢.本文在分析模型复杂度的基础上提出了一种改进的快速CRF算法.该算法通过引入小规模单字特征降低特征的规模,并通过在推理过程中引入任务相关的人工知识压缩Viterbi和Baum-Welch格搜索空间,提高了训练的速度.在中文863命名实体识别评测语料和SIGHAN06语料集上进行的实验表明,该算法在不影响中文命名实体识别精度的同时,有效地降低了模型的训练代价.

9.期刊论文 冯元勇.孙乐.董静.李文波.FENG Yuan-yong.SUN Le.DONG Jing.LI Wen-bo 基于分类信心重排序的中文共指消解研究 -中文信息学报2007,21(6)

共指消解是自然语言处理的核心问题之一.本文针对分步消解中分类器全局信息的不足,依据分类信心对全体提及配对进行排序,优先根据可靠的分类结果对提及进行聚集或分离.实验表明,该算法在多个学习框架下显著地改善了系统的整体性能.

10.期刊论文 董静.孙乐.冯元勇.黄瑞红.DONG Jing.SUN Le.FENG Yuan-yong.HUANG Rui-hong 中文实体关系抽取中的特征选择研究 -中文信息学报2007,21(4)

命名实体关系抽取是信息抽取研究领域中的重要研究课题之一.通过分析,本文提出将中文实体关系划分为:包含实体关系与非包含实体关系.针对同一种句法特征在识别它们时性能的明显差异,本文对这两种关系采用了不同的句法特征集,并提出了一些适合各自特点的新的句法特征.在CRF模型框架下,以ACE2007的语料作为实验数据,结果表明本文的划分方法和新特征有效的提高了汉语实体关系抽取任务的性能.