命名实体识别研究进展综述

孙 镇[1,3] 王惠临[2]
[1](北京大学信息管理系 北京 100871)
[2](中国科学技术信息研究所 北京 100038)
[3](全国组织机构代码管理中心 北京 100029)

【摘要】介绍命名实体识别的研究背景和意义,总结国内外命名实体识别研究历史,详细介绍目前主流的技术方法和评估方法,讨论命名实体识别技术的发展趋势。
【关键词】命名实体识别 信息抽取 信息检索 机器翻译 评估方法
【分类号】TP391

Overview on the Advance of the Research on Named Entity Recognition

Sun Zhen[1,3] Wang Huilin[2]
[1](Department of Information Management, Peking University, Beijing 100871, China)
[2](Institute of Scientific & Technical Information of China, Beijing 100038, China)
[3](National Administration for Code Allocation to Organizations, Beijing 100029, China)

【Abstract】The paper introduces the background and motivation of Named Entity Recognition, and summarizes the history development of Named Entity Recognition at home and abroad, as well as the related technology and evaluation method. Finally, it discusses the new development trends on Named Entity Recognition.
【Keywords】Named entity recognition Information extraction Information retrieval Machine translation Evaluation methods

1 引 言

命名实体识别(Named Entity Recognition,NER)的主要任务是识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类[1]。命名实体识别技术是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。从语言分析的全过程来看,命名实体识别属于词法分析中未登录词识别的范畴。命名实体是未登录词中数量最多、识别难度最大、对分词效果影响最大的问题。根据 SIGHAN(http://www.sighan.org/) Bakeoff 数据评测结果,未登录词造成的分词精度失落至少比歧义大 5 倍以上[2],可见命名实体地位的重要性。

针对"命名实体识别"和"Named Entity Recognition"两个关键词对中国知网(http://www.cnki.net)、Google 学术(http://scholar.google.com/)、ACL Anthology(http://www.aclweb.org/anthology/)、ICML 会议(http://machinelearning.org/icml.html)以及 IEEE(http://ieeexplore.ieee.org)等期刊论文数据库和会议论文集进行文献检索,时间覆盖范围从 1991 年至 2009 年,检索文献主要包括中文和英文文献。笔者主要回顾命名实体识别任务的历史发展,并分别针对命名实体识别研究内容及进展、系统评测组织及方法等几个方面进行研究,综述有关文献,最后

收稿日期:2010-03-24
收修改稿日期:2010-04-11

对未来发展趋势进行展望。

2　命名实体识别研究历史

国外对于英文命名实体识别的研究开始比较早。1991 年 Rau 在第 7 届 IEEE 人工智能应用会议上发表了“抽取和识别公司名称”的有关研究文章,首次描述了抽取和识别公司名称的系统,该系统主要采用启发式算法和手工编写规则的方法[3]。1996 年,命名实体评测作为信息抽取的一个子任务被引入 MUC－6[4],在其后的 MUC－7 的 MET－2[5]以及 IEER－99、CoNLL－2002、CoNLL－2003、IREX、LREC 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。

由于英文命名实体的识别中只需考虑词本身的特征而不涉及分词问题,因此实现难度相对较低。根据 MUC① 以及 ACE② 的评测结果,测试的准确率、召回率、F1 值目前大多可以达到 90% 左右。

由于中文内在的特殊性决定了在文本处理时首先必须进行词法分析,中文命名实体识别的难度要比英文的难度大。中文命名实体识别起步较晚,20 世纪 90 年代初期开始,国内一些学者对中文命名实体(如:地名、人名、组织机构名等)识别进行了一些研究。如:孙茂松等[6]在国内比较早开始进行中文人名识别,他们主要采用统计的方法计算姓氏和人名用字概率。张小衡等[7]对中文机构名称进行识别与分析,主要采用人工规则对高校名进行了实验研究。Intel 中国研究中心的 Zhang 等[8]在 ACL2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统,该系统利用基于记忆的学习(Memory Based Learning, MBL)算法获取规则,用以抽取命名实体及它们之间的关系。

国外对于命名识别研究主要包括:Bikel[9]等最早提出了基于隐马尔可夫模型的英文命名实体识别方法,其在 MUC－6 测试文本集的测试结果为:英文地名、机构名和人名的识别精度分别达到了 97%、94% 和 95%, 召回率分别达到了 95%、94% 和 94%;Liao 等[10]提出了基于条件随机场模型,采用半监督的学习算法进行命名实体识别; Ratinov 等[11]采用未标注文本训练词类模型(Word Class Model)的办法,可以有效地提高 NER 系统的识别效率,并针对 CoNLL－2003 的数据集开发出 F1 值达到 90.8% 的命名实体识别系统。中文名实体识别也获得了广泛关注。Tsai 等[12]提出基于最大熵的混合的方法; 冯元勇等[13]提出基于单字提示特征的中文命名实体识别快速算法;郑逢强等[14]将《知网》中的义原作为特征加入到最大熵模型中,以此来训练产生性能更好的模型。

2004 年举行的 863 命名实体识别评测,成绩最好的命名实体识别系统的准确率、召回率和 F1 值分别为 81.10%、83.69%、82.38%,其中人名、地名、组织机构名各项的 F1 值最高分别为 85.51%、82.51%、60.81%。可见中文命名实体识别评测结果比英文结果偏低,尤其中文机构名称的识别难度更大一些。

3　研究内容及发展

3.1　研究主体

命名实体是命名实体识别的研究主体,一般包括 3 大类(实体类、时间类和数字类)和 7 小类(人名、地名、机构名、时间、日期、货币和百分比)命名实体。实际研究中,命名实体的确切含义需要根据具体应用来确定,比如,可能需要把产品名称[15]、旅游景点名称[16]等作为命名实体。在面向生物命名实体信息提取时,还包括蛋白质、基因、核糖核酸、脱氧核糖核酸、细胞等特殊生物实体[17]。

由于数量、时间、日期、货币等实体识别通常可以采用模式匹配的方式获得较好的识别效果,相比之下人名、地名、机构名较复杂,因此近年来的研究主要以这几种实体为主。同时生物领域的实体识别也比较活跃。这些实体中以机构名和生物实体识别难度最大,普遍存在嵌套和缩写的识别问题。从研究的发展趋势上看,由原来的单独针对人名、地名等进行识别发展到开始采用统一的方法同时进行各类中文命名实体的识别,而且识别效果也得到了提高,其中部分研究成果发表在 ACL(http://www.aclweb.org/)年度会议以及 COLING(http://nlp.shef.ac.uk/iccl/)、SIGHAN 等国际会议上。

这种方法虽然考虑了人名、地名和机构名的共同特点,能够有效地解决多种命名实体间的歧义问题。但是,它不能充分分析不同命名实体间的差异性,制约了整体的识别性能。

① http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
② http://www.itl.nist.gov/iad/mig/tests/ace/

主体所属的领域[18]，包括新闻、生物医学等领域都有相关研究表明命名实体识别呈现弱领域相关性。不同领域具有类似特点，但是从某个领域移植到一个新领域，通常会遇到严重的性能下降问题，主要原因在于命名实体的标记定义不同、不同领域有着不同的形态语法特点。

命名实体识别对英语、中文、德语、日语、西班牙语、葡萄牙语等都有相应研究。初期研究主要以英文为主，1995 年以后开始出现对欧洲和少数亚洲语言的研究。随着多语言信息抽取任务的提出，对多语言（Multilingual）[19]、独立语言（Language Independent）的命名实体研究也开始增加，CoNLL－2002[20]、CoNLL－2003[21]都将独立语言的命名实体识别作为共享任务提出。

3.2 命名实体识别特点及难点

评判一个命名实体是否被正确识别包括两个方面：实体的边界是否正确；实体的类型是否标注正确。主要错误类型包括文本正确，类型可能错误；反之，文本边界错误，而其包含的主要实体词和词类标记可能正确。

英语中的命名实体具有比较明显的形态标志，如人名、地名等实体中的每个词的第一个字母要大写等，所以实体边界识别相对汉语来说比较容易，任务的重点是确定实体的类型。和英语相比，汉语命名实体识别任务更加复杂，由于分词等因素的影响难度较大，其难点主要表现在如下几个方面：

（1）命名实体类型多样，数量众多，不断有新的命名实体涌现，如新的人名、地名等，难以建立大而全的姓氏库、名字库、地址库等数据库。

（2）命名实体构成结构比较复杂，并且某些类型的命名实体词的长度没有一定的限制，不同的实体有不同的结构，比如组织名存在大量的嵌套、别名、缩略词等问题，没有严格的规律可以遵循；人名中也存在比较长的少数民族人名或翻译过来的外国人名，没有统一的构词规范。因此，对这类命名实体识别的召回率相对偏低。

（3）在不同领域、场景下，命名实体的外延有差异，存在分类模糊的问题。不同命名实体之间界限不清晰，人名也经常出现在地名和组织名称中，存在大量的交叉和互相包含现象，而且部分命名实体常常容易与普通词混淆，影响识别效率。在个体户等商户中，组织名称中也存在大量的人名、地名、数字的现象，要正确标注这些命名实体类型，常常要涉及上下文语义层面的分析，这些都给命名实体的识别带来困难。

（4）在不同的文化、领域、背景下，命名实体的外延有差异。对命名实体的定界和类型确定，目前还没有形成共同遵循的严格的命名规范。

（5）命名实体识别过程常常要与中文分词、浅层语法分析等过程相结合，分词、语法分析系统的可靠性也直接决定命名实体识别的有效性，使得中文命名实体识别更加困难。

4 主要技术方法

命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法等。

4.1 基于规则和词典的方法

基于规则的方法多采用语言学专家手工构造规则模板，选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词（如尾字）、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。

基于规则和词典的方法是命名实体识别中最早使用的方法，大多数参加 MUC－7 会议评测的系统，都是基于手写规则的方法。采取这种方法的代表性系统包括 GATE（http://gate.ac.uk/）项目中的 ANNIE 系统以及参加 MUC 评测的 FACILE 系统等。它们都是依赖于手工规则的系统，都使用命名实体库，而且对每一个规则都赋予权值。当遇到规则冲突的时候，选择权值最高的规则来判别命名实体的类型。王宁等[22]利用规则的方法进行金融领域的公司名识别，该系统对知识库的依赖性强，同时开放和封闭测试的结果也显示了规则方法的局限性。

一般而言，当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。基于规则的方法的另外一个缺点是代价太大，存在系统建设周期长、移植性差而且需要建立不同领域知识库作为辅助以提高系统识别能

力等问题。

4.2　基于统计的方法

基于统计的方法利用人工标注的语料进行训练，标注语料时不需要广博的语言学知识，并且可以在较短时间内完成。在 CoNLL－2003 会议上，所参赛的 16 个系统全部采用基于统计的方法，该方法成为目前研究的主流方法。这类系统在移植到新的领域时可以不做或少做改动，只要利用新语料进行一次训练即可。基于统计机器学习的方法主要包括：隐马尔可夫模型（Hidden Markov Model，HMM）、最大熵（Maximum Entropy，ME）、支持向量机（Support Vector Machine，SVM）、条件随机场（Conditional Random Fields，CRF）等。

在这 4 种学习方法中，最大熵模型结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，有时甚至导致训练代价难以承受，另外由于需要明确的归一化计算，导致开销比较大。而条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题。一般说来，最大熵和支持向量机在正确率上要比隐马尔可夫模型高一些，但是隐马尔可夫模型在训练和识别时的速度要快一些，主要是由于在利用 Viterbi 算法求解命名实体类别序列的效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，如短文本命名实体识别[23]。

基于统计的方法对特征选取的要求较高，需要从文本中选择对该项任务有影响的各种特征，并将这些特征加入到特征向量中。依据特定命名实体识别所面临的主要困难和所表现出的特性，考虑选择能有效反映该类实体特性的特征集合。主要做法是通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。有关特征可以分为具体的单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征等。张祝玉等[24]针对条件随机场的特征选取与组合进行了比较研究，通过实验比较得出在训练时应优先选择贡献度大的特征，同时还表明使用组合特征可以提升系统的性能。

基于统计的方法对语料库的依赖也比较大，而可以用来建设和评估命名实体识别系统的大规模通用语料库又比较少。SIGHAN Bakeoff 08 测评中，中文命名实体识别使用的语料主要包括：香港城市大学语料库（1 772 202 字，训练集）、微软亚洲研究院语料库（1 089 050字，训练集）、北京大学语料库（1 833 177 字，训练集）[25]。这些语料库比较小、应用不广泛，无法应用于大规模的 NER 系统。因此，目前的问题是如何最大限度地使用这些有限的语料库。

针对外部资源的使用，借助于 Wikipedia、HowNet 等知识库的方法可以较好地解决新词识别等问题。Kazama 等[26]采用 Wikipedia 抽取分类标识和条件随机场模型进行命名实体识别，通过对 CoNLL－2003 数据集进行测试，证明该方法能够提高系统的精确度，但在消除歧义方面仍有不足。Cucerzan[27]采用 Wikipedia 作为知识库对大规模命名实体识别语义消歧进行研究，提高了 Wikipedia 抽取的上下文与文档上下文信息的一致性，取得较好的语义消歧效果。

4.3　混合方法

自然语言处理并不完全是一个随机过程，单独使用基于统计的方法使状态搜索空间非常庞大，必须借助规则知识提前进行过滤修剪处理。目前几乎没有单纯使用统计模型而不使用规则知识的命名实体识别系统，在很多情况下是使用混合方法：

（1）统计学习方法之间或内部层叠融合，如俞鸿魁等[28]采用层叠隐马尔可夫模型对中文进行分词。

（2）规则、词典和机器学习方法之间的融合，其核心是融合方法技术。在基于统计的学习方法中引入部分规则，将机器学习和人工知识结合起来。

（3）将各类模型、算法结合起来，将前一级模型的结果作为下一级的训练数据，并用这些训练数据对模型进行训练，得到下一级模型。这种方法在具体实现过程中需要考虑怎样高效地将两种方法结合起来，采用什么样的融合技术。由于命名实体识别在很大程度上依赖于分类技术，在分类方面可以采用的融合技术主要包括如 Voting，XVoting，GradingVal，Grading 等。

Lin 等[29]将最大熵方法与基于词典匹配和规则模式的后处理相结合，前一阶段运用 ME 方法识别文本中的生物实体，第一阶段机器学习方法可能产生一定程度的边界识别错误和语义分类错误，通过第二阶段基于词典和规则模式匹配的后处理，修正实体边界并改进实体语义分类结果，提高了系统的准确率与召回率。

5 评测组织及方法

命名实体识别的进步得益于有关评测,其方法主要采用语言学专家的标注识别结果与机器自动识别的结果进行比较来进行评测。由于边界特征、类型属性选择以及性能要求等因素的影响,对一个实体识别正确性的定义是不简单的,不同的系统侧重点不同,目前主流的评价方法以有关评测会议标准为主。

目前,比较有影响力的评测会议主要有信息理解研讨会(Message Understanding Conference,MUC)、多语种实体评价任务(Multilingual Entity Task Evaluation,MET)、自动内容抽取(Automatic Content Extraction,ACE)、文本理解会议(Document Understanding Conference,DUC)、SIGHAN 的 Bakeoff 评测等,对信息抽取技术的发展起到了很大的推动作用。

MUC 会议是一个比较有影响力的评测会议,从 MUC-3 开始引入正式的评测标准,其中借用了信息检索领域采用的一些概念,如召回率和准确率等。在 MUC-5 会议上,组织者尝试采用平均填充错误率(Error Per Response Fill, ERR)作为主要评价指标。MUC-6 的评测更为细致,强调系统的可移植性以及对文本的深层理解能力。总体来说,在 MUC 中,主要根据两个评价指标衡量信息抽取系统的性能:召回率和准确率。召回率(REC)等于系统正确抽取的结果占所有可能正确结果的比例;准确率(PRE)等于系统正确抽取的结果占所有抽取结果的比例。为了综合评价系统的性能,通常还计算召回率和准确率的加权几何平均值,即 F 指数,计算公式如下:

$$\text{F-Measure} = \frac{(\text{beta}^2 + 1.0) \times \text{PRE} \times \text{REC}}{(\text{beta}^2 \times \text{PRE}) + \text{REC}}$$

其中,beta 是召回率和准确率的相对权重。beta 等于 1 时,二者同样重要;beta 大于 1 时,准确率更重要一些;beta 小于 1 时,召回率更重要一些。在 MUC 系列会议中,beta 取值一般为 1、0.5、2。

在 MUC-7 之后,MUC 被由 NIST 主导的 ACE 评测所取代。与 MUC 相比,ACE 评测不针对某个具体的领域或场景,采用基于漏报(标准答案中有而系统输出中没有)和误报(标准答案中没有而系统输出中有)为基础的一套评价体系。ACE 的评测原则采用对每篇文档单独评测,文档的评测是相互独立的,ACE08[30] 中引入了跨文档的评测任务。ACE 的评测采用系统输出与参考答案之间的全局最优匹配原则,当系统输出的结果与参考答案不匹配时(如发生错误警告、实体遗漏和类型错误等情况),评测系统将根据具体情况倒扣一定分数作为惩罚。

6 结 语

命名实体识别作为信息抽取、问答系统、机器翻译等任务中的基础工作,近年来在多媒体索引、半监督和无监督的学习、复杂语言环境和机器翻译等方面取得大量新的研究成果。未来的研究也将围绕这些方面展开,尤其是对多媒体信息处理,如从多媒体信息中抽取命名实体以及大规模文本处理和同时处理多种类型实体技术的使用。随着半监督的学习和无监督的学习方法不断被引入到这个领域,采用未标注语料集等方法将逐步解决语料库不足的问题。在复杂语言现象(如借喻等)研究以及命名实体识别系统与机器翻译的互提高方面,也有广阔的发展空间。命名实体识别将在更加开放的领域中,综合各方面的发展成果,为信息处理的深层次发展奠定更坚实的基础。

参考文献:

[1] Chinchor N. MUC-7 Named Entity Task Definition[C]. In: *Proceedings of the 7th Message Understanding Conference*, Virginia. 1998.

[2] Sproat R, Emerson T. The First International Chinese Word Segmentation Bakeoff[C]. In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan. 2003: 133-143.

[3] Rau L F. Extracting Company Names from Text[C]. In: *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*. 1991: 29-32.

[4] Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]. In: *Proceedings of the 16th International Conference on Computational Linguistics*. 1996.

[5] Chinchor N A. Overview of MUC-7/MET-2[C]. In: *Proceedings of the 7th Message Understanding Conference*. 1998.

[6] 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*,1995,9(2):16-27.

[7] 张小衡,王玲玲. 中文机构名称的识别与分析[J]. *中文信息学报*,1997, 11(4):21-32.

[8] Zhang Y, Zhou J F. A Trainable Method for Extracting Chinese Entity Names and Their Relations[C]. In: *Proceedings of the 2nd*

Chinese Language Processing Workshop, HongKong. 2000:66 – 76.

[9] Bikel D M, Schwarta R, Weischedel R M. An Algorithm that Learns What's in a Name[J]. *Machine Learning Journal Special Issue on Natural Language Learning*, 1999, 34(1 – 3):211 – 231.

[10] Liao W, Veeramachaneni S. A Simple Semi – supervised Algorithm for Named Entity Recognition[C]. In: *Proceedings of the NAACL HLT* 2009 *Workshop on Semi – supervised Learning for Natural Language Processing*. 2009:58 – 65.

[11] Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition[C]. In: *Proceedings of the* 13*th Conference on Computational Natural Language Learning*. 2009:147 – 155.

[12] Tsai T, Wu S, Lee C, et al. Mencius: A Chinese Named Entity Recognizer Using the Maximum Entropy – based Hybrid Model[J]. *International Journal of Computational Linguistics & Chinese Language Processing*, 2004, 9(1):65 – 81.

[13] 冯元勇,孙乐,李文波,等. 基于单字提示特征的中文命名实体识别快速算法[J]. *中文信息学报*, 2008, 22(1):105 – 110.

[14] 郑逢强, 林磊,刘秉权,等.《知网》在命名实体识别中的应用研究[J]. *中文信息学报*, 2008, 22(5):97 – 101.

[15] 刘非凡,赵军,吕碧波,等. 面向商务信息抽取的产品命名实体识别研究[J]. *中文信息学报*, 2006, 20(1):7 – 13.

[16] 薛征山,郭剑毅,余正涛,等. 基于 HMM 的中文旅游景点的识别[J]. *昆明理工大学学报:理工版*, 2009, 34(6):44 – 48.

[17] 邱莎. 基于统计的生物命名实体识别研究[D]. 成都:四川大学, 2006.

[18] 徐薇,付滨,刘柳,等. 中文命名实体识别系统的领域扩展[C]. 见:*第9 届全国计算语言学学术会议论文集*. 2007.

[19] Poibeau T. The Multilingual Named Entity Recognition Framework [C]. In: *Proceedings of the* 10*th Conference on European Chapter of the Association for Computational Linguistics*. 2003:155 – 158.

[20] Sang T K. Introduction to the CoNLL – 2002 Shared Task: Language – Independent Named Entity Recognition[C]. In: *Proceedings of the* 6*th Conference on Natural Language Learning*, Taipei, Taiwan. Morristown, NJ, USA: Association for Computational Linguistics, 2002:1 – 4.

[21] Sang T K, Meulder F D. Introduction to the CoNLL – 2003 Shared Task: Language – Independent Named Entity Recognition[C]. In: *Proceedings of the* 7*th Conference on Natural Language Learning at HLT – NAACL*, Edmonton, Canada. Morristown, NJ, USA: Association for Computational Linguistics, 2003:142 – 147.

[22] 王宁,葛瑞芳,苑春法,等. 中文金融新闻中公司名的识别[J]. *中文信息学报*, 2002, 16(2):1 – 6.

[23] 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):143 – 145.

[24] 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]. 见:*第4 届全国信息检索与内容安全学术会议论文集*. 2008.

[25] 第一届中国中文信息学会汉语处理评测(CIPS – CLPE)暨第四届国际中文自然语言处理 Bakeoff[EB/OL]. [2010 – 01 – 11]. http://www.china – language.gov.cn/bakeoff08/.

[26] Kazama J, Torisawa K. Exploiting Wikipedia as External Knowledge for Named Entity Recognition[C]. In: *Proceedings of the* 2007 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007:698 – 707.

[27] Cucerzan S. Large – Scale Named Entity Disambiguation Based on Wikipedia Data[C]. In: *Proceedings of Empirical Methods in Natural Language Processing*, Prague, Czech Republic. 2007:708 – 716.

[28] 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 – 93.

[29] Lin Y, Tsai T, Chou W, et al. A Maximum Entropy Approach to Biomedical Named Entity Recognition[C]. In: *Proceedings of the* 4*th ACM SIGKDD Workshop on Data Mining in Bioinformatics*. 2004.

[30] Automatic Content Extraction 2008 Evaluation Plan(ACE08)[EB/OL]. (2008 – 05 – 30). [2010 – 01 – 11]. http://nist.gov/speech/tests/ace/2008/doc/ace08 – evalplan.v1.2d.pdf.

(作者 E – mail:sunzhenyh@yahoo.com.cn)