中文分词与词性标注研究

梁喜涛,顾 磊
(南京邮电大学 计算机学院,江苏 南京 210023)

摘 要: 分词和词性标注是中文语言处理的重要技术,广泛应用于语义理解、机器翻译、信息检索等领域。在搜集整理当前分词和词性标注研究与应用成果的基础上,对中文分词和词性标注的基本方法进行了分类和探讨。首先在分词方面,对基于词典的和基于统计的方法进行了详细介绍,并且列了三届分词竞赛的结果;其次在词性标注方面,分别对基于规则的方法和基于统计的方法进行了阐述;接下来介绍了中文分词和词性标注一体化模型相关方法。此外还分析了各种分词和词性标注方法的优点和不足,在此基础上,为中文分词和词性标注的进一步发展提供了建议。

关键词: 中文分词;主动学习;词性标注;自然语言处理;一体化模型
中图分类号: TP311 **文献标识码:** A **文章编号:** 1673-629X(2015)02-0175-06
doi:10.3969/j.issn.1673-629X.2015.02.040

Study on Word Segmentation and Part-of-speech Tagging

LIANG Xi-tao, GU Lei
(College of Computer and Software, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Word segmentation and Part-Of-Speech (POS) tagging are the basic task of the CLP (Chinese Language Processing) and are widely applied in the semantic understanding, machine translation, information retrieval and other fields. In this paper, based on collecting current research and application results of word segmentation and part-of-speech tagging, analyze and classify the basic methods of Chinese Word Segmentation (CWS) and POS tagging. First in terms of word segmentation, dictionary-based segmentation method and statistics-based segmentation method were introduced in detail and some word segmentation results of the competition were also listed. Secondly in terms of POS tagging, rule-based method and statistics-based method were expounded. Next, the main methods of building the model for joint CWS and POS tagging were presented. In this paper, also analyze the advantages and disadvantages for methods of CWS and POS tagging, based on which suggestions for the further development are put forward.

Key words: Chinese word segmentation; active learning; POS tagging; CLP; joint model

0 引 言

现代汉语词法、句法分析是进行汉语语义理解、中英文机器翻译、中文信息检索等首要解决的问题。只有把分词和词性标注都处理好才能处理词法、句法问题[1]。在中文分词和词性标注的研究中,未登录词识别是棘手问题之一,它是影响分词效果的技术瓶颈。随着中文分词和词性标注竞赛的举行,许多实用的好方法层出不穷。其中基于字符的分词和词性标注方法在这几次中文分词竞赛中占据了主导地位。相对于基于词的分词方法,它能够取得令人满意的未登录词识别率,大大加速分词和词性标注进程[2]。

文章内容主要围绕中文分词和词性标注展开。首先分别对中文分词和词性标注的主流方法进行了简单介绍,并对各种方法的优缺点进行了简短总结;然后介绍了分词和词性标注一体化模型的工作原理及相关技术和方法;最后总结了分词和词性标注方法还面临的主要问题,并提出了今后值得研究的方向。

1 中文分词

中文分词是计算机进行汉语处理的基础。国内的中文分词研究始于20世纪80年代,CDWS(Chinese Distinguishing Word System)书面汉语自动分词系统是

收稿日期: 2014-03-07 **修回日期:** 2014-06-11 **网络出版时间:** 2014-12-27
基金项目: 国家自然科学基金资助项目(61302157);教育部人文社会科学研究青年基金(12YJC870008);江苏省教育高校哲学社会科学基金(2013SJB870004);江苏省社科研究文化精品课题(12SWC-030)
作者简介: 梁喜涛(1989-),男,硕士研究生,研究方向为中文信息处理;顾 磊,副教授,硕士生导师,研究方向为中文信息处理、机器学习。
网络出版地址: http://www.cnki.net/kcms/detail/61.1450.TP.20141227.1343.026.html

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

我国第一个实用的自动分词系统，这是汉语自动分词的第一次尝试，具有很大的理论和指导意义。中文分词主要方法有：基于词典的方法和基于统计的方法。另外对于获得大量已标注的样例困难或耗费时间的问题，本节还介绍了利用主动学习进行分词的思想。

1.1 基于词典的分词方法

基于词典的分词方法又叫机械分词方法，它按照一定的策略把待切分的字符串与分词词典中的词进行比对，如果在词典中找到待切分的字符串则匹配成功。基于词典的分词方法流程如图 1[3] 所示。基于词典的分词方法有正向匹配、逆向匹配、最长匹配和最短匹配。在实际的应用过程中，研究人员在此基础上还提出了一些新的方法。2007 年，张海营[4] 提出一种新的分词词典，通过为词典建立首字 Hash 表和词索引表使分词词典支持全二分最大匹配算法，降低了时间复杂度。2011 年，Mai F J 等[5] 提出了一种基于双向匹配和特征选择的分词算法，对于双向匹配算法不能够处理的歧义问题，再利用特征选择算法来解决。2012 年，曹月雷等[6] 提出了词典与后缀数组相结合的分词方法，利用后缀数组快速准确抽取文档中的中、高频词，利用词典切分其他词汇。

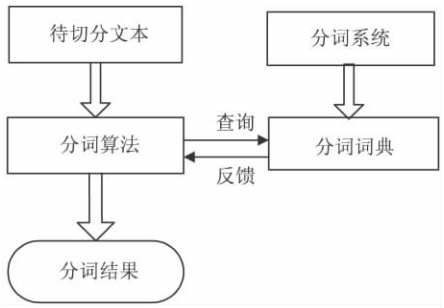

图 1 基于词典的分词方法流程图

机械分词方法实现比较简单，但效率和准确性容易受词典容量的约束。另外词典的结构还直接影响词典的查询速度[7]。因此在实际应用中，研究人员会对词典结构进行改造，利用 Hash 表、索引表和后缀数组等结构，采用更高效的查询算法来加快词典的查询速度。再或者把基于词典的分词方法和统计的分词方法结合起来，把机械分词方法当成一种初级分词手段，对待测试文本进行简单、快速处理，再利用统计其他语言信息的方法来处理歧义和未登录词等问题，从而提高切分的准确率。

1.2 基于统计的分词方法

基于统计的分词方法是根据相邻字的紧密结合程度来进行分词。这个方法需要计算训练语料中相邻字的紧密结合程度，当紧密程度高于某一阈值时，可认为相邻的字组成了一个词组。基于该方法的主要模型有 n - gram 模型和最大熵模型。

n - gram 模型认为第 i 个词的出现只与前面的 i - 1 个词有关，整句的概率就是各个词出现概率的乘积。最大熵模型是在给定的训练样本上挖掘潜在的约束条件，通过设置约束条件调节模型对未知数据的适应度和对已知数据的拟合程度[8]。统计模型通过待切分文档中组合词的词频以及基本词之间结合的稳定性来发现汉字串的结合规律。它可能会切分出一些不是词汇，但经常一同出现的汉字，如：是的、我的等。另外该方法还需要大规模标注语料支持，当分词领域变换后，必须提供相应领域的训练语料来训练模型。所以在使用过程中不断有学者对传统的统计模型进行改进或者在此基础上加入其他的方法来提高分词的准确率。2006 年，Wang Xinhao 等[9] 提出了利用最大熵和 n - gram 模型的方法，利用最大熵模型把分词转换成分类问题，把待切分文本中的每个字符都贴上位置信息，利用 n - gram 模型来表示文本字符间的关系，弥补最大熵模型的不足。2010 年，Zhang Liyan 等[10] 提出了一种基于最大熵模型的分词算法，利用词性标注和语料中词出现的概率来确立基于互信息的最大熵模型进行分词，最后用一个二进制模型得出当训练语料增大到一定值时分词准确率不会随着训练语料的扩大而提高的结论。2012 年，Zhang Meishan[11] 提出了统计和词典相结合的领域自适应分词方法，使用通用词典训练统计模型；当分词领域改变时，只需要在原有词典基础上再添加相应领域的词典，可大大增强分词的领域适应性。

该方法可以识别出频率较高的未登录词，还很好解决了新词出现的问题，这是机械分词无法做到的。但获得训练语料通常都需要大量的人力和物力，而且分词精度与训练文本的选择也有关系。所以一般把词典和统计的分词方法结合起来，利用这两种算法的长处，来减少统计信息的计算步骤。利用统计方法对词典切分不正确的词消除歧义，对词典不含有的新词进行识别，准确率也比单一使用词典和统计模型要高。

1.3 基于主动学习的分词方法

在中文分词领域，国内外对主动学习进行分词的研究还比较少，该方法还是一种比较新的方法。它一般分为两部分：学习部分和选择部分。学习部分就是一个基本的分词器，用训练集合 L 来训练分词器提高性能；选择部分通过使用样本选择算法从未标记集合 U 中选择样本由专家进行标注后加入到 L 中来继续训练分词器，从而提高分词器的准确率[12]。一般主动学习分词方法流程如图 2 所示。

目前国外学者已经将主动学习应用到诸多自然语言处理相关的任务中，比如信息抽取、文本分类和基于短语结构的句法分析等。国内清华大学、北京理工大学、中科大、上海交大等将主动学习应用到文本分类和

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

组织机构名识别中,并取得了一定效果。2009 年,朱红斌等[13]提出了一种基于主动学习支持向量机的文本分类方法,先采用向量空间模型对文本特征进行提取,再使用互信息对文本特征进行降维,最后使用主动学习算法对支持向量机进行训练,使用训练后的分类器对进行新的文本分类。2010 年,陈锦禾等[14]提出了一种基于信息熵的主动学习方法,通过对未标注文档熵值的计算,结合二阶段学习策略,选取最有可能解决问题的样本进行标注,获得新的参数来重新训练分类器。2012 年,Li Shoushan 等[15]首次成功把主动学习应用到中文分词领域,通过仅仅注释不确定的字符边界,使人工标记的代价大大降低; 另外还引入了差异性测量标准,以此来避免重复注释问题。

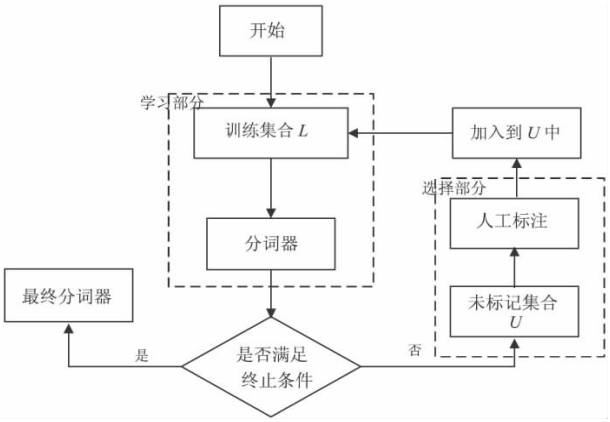

图 2　主动学习方法流程图

这些实例为进行主动学习分词提供了很好的范例,也证明了主动学习方法是切实可行的。主动学习过程中,一种常用的样例选择方法是倾向选择当前分词器最无法确定其类别的样例,以此来减少训练的数量和人工标注成本。与传统的监督方法相比,主动学习能够很好地处理训练数据集较大的情况,现已成为模式识别、机器学习和数据挖掘领域的研究热点。其中委员会查询、边缘查询、后验概率查询是该方法的典型代表,近年来有效地推动了主动学习的迅速发展。对于像微博这种数据规模较小、信息含量较少的短文本来说,主动学习能够有效利用文本信息,在有限的时间和资源的前提下进行分词。但在具体应用过程中要注意同应用领域的先验知识相结合,研究更加高效的主动学习选择策略,以减少标注样例的代价。

1.4　中文分词竞赛

2003 年、2005 年和 2006 年分别举办了中文分词竞赛,提出利用学习算法和启发式算法来进行分词的思想,还提出用基于语料库分词标准表述替代分词词典和分词手册表示,使中文分词在某种意义上更接近基于语料库的机器学习过程。分词竞赛中通常采用如下五项评测指标: P,R,F,R_{oov},R_{Iv}[16]。其中各个指标的定义如下: F 综合值: $F = 2PR/(P+R)$ (F 综合值是衡量分词系统综合切分性能的首要指标),其中 P = (正确切分的词组数/系统切分的词组数) $\times 100\%$,表示分词系统的整体准确率; R = (正确切分的词组数/切分文本中的词组数) $\times 100\%$,表示系统的整体召回率; 另外还有 $R_{oov} = \frac{\text{正确切分的未登录词语数}}{\text{切分文本中的未登录词语数}} \times 100\%$,表示分词系统对词典未登录词识别的召回率; $R_{Iv} = \frac{\text{正确切分的词典已登录词语数}}{\text{切分文本中词典已登录的词语数}} \times 100\%$,表示系统对词典已登录词识别的召回率。各项指标分值越高,代表系统在这方面的能力就越强。这三次竞赛的分词结果如表 1 ~ 3 所示[16-18]。

表 1　第一届中文分词竞赛的测试结果
(基线得分/最高得分)

语料库	P	R	F	R_{oov}	R_{Iv}
AS	0.912/0.993	0.917/0.990	0.915/0.992	0.000/0.988	0.938/0.990
CTB	0.663/0.988	0.800/0.982	0.725/0.985	0.062/0.990	0.962/0.980
HK	0.830/0.991	0.908/0.986	0.867/0.989	0.037/0.996	0.974/0.985
PK	0.829/0.996	0.909/0.995	0.867/0.995	0.050/1.000	0.972/0.994

表 2　第二届中文分词竞赛的测试结果
(基线得分/最高得分)

语料库	P	R	F	R_{oov}	R_{Iv}
AS	0.857/0.985	0.909/0.979	0.882/0.982	0.004/0.996	0.950/0.978
CITYU	0.790/0.991	0.882/0.988	0.833/0.989	0.000/0.997	0.952/0.988
MSR	0.912/0.992	0.955/0.991	0.933/0.991	0.000/0.998	0.981/0.990
PKU	0.836/0.988	0.904/0.985	0.869/0.987	0.059/0.994	0.956/0.985

表 3　第三届中文分词竞赛的测试结果
(基线得分/最高得分)

语料库	P	R	F	R_{oov}	R_{Iv}
CITYU	0.882/0.985	0.930/0.982	0.906/0.984	0.009/0.993	0.969/0.981
CKIP	0.870/0.987	0.915/0.980	0.892/0.983	0.030/0.997	0.954/0.979
MSRA	0.900/0.993	0.949/0.991	0.924/0.992	0.022/0.999	0.981/0.991
UPUC	0.790/0.976	0.869/0.961	0.828/0.968	0.011/0.989	0.951/0.958

其中语料库有台湾中研院(AS) 、中国知识信息处理实验室(CKIP) 、香港城市大学(CITYU 或 HK) 、微软亚洲研究院(MSRA 或 MSR) 、宾夕法尼亚大学(UPUC 或 CTB) 和北京大学(PKU 或 PK) 语料库。分词竞赛加快了分词进程的步伐,分词标准不断得到完善,使得分词比赛更加国际化。各科研人员相继开展了大量与中文分词相关的理论和实践工作,各种性能优良的中文分词系统不断出现,分词的准确率不断得到提高。

2　词性标注

所谓词性标注就是为每个词的词性加上标注,也就是确定该词属于名词、动词、形容词还是其他词性的

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

过程,它也是自然语言处理领域的基础,也是像机器翻译、信息检索等应用中一个不可缺少的环节[19]。中文词性标注方法主要包括两种:基于统计的方法和基于规则的方法。

2.1 基于统计的词性标注方法

基于统计的自然语言处理方法在消除歧义和句法分析等方面得到越来越广泛的应用,是近年来兴起的一种新的也是最常使用的方法。对于给定的输入词串,该方法先确定其所有可能的词性串,选出得分最高的作为最佳输出。其中应用比较广泛的主要有隐马尔可夫模型(HMM)方法和条件随机场(CRF)的方法。

HMM 模型是一种特别适合处理随机序列数据的统计模型。它需要大量训练语料来达到较高的标注准确率,还存在标记偏置问题[20]。CRFs 模型是在给定输入节点条件下计算输出节点条件概率的无向图模型,是序列标注和切分的统计模型。CRFs 虽然能克服最大熵和 HMM 等有向图模型的标记偏置[21]问题,但它倾向选择概率较高的词进行标记,这可能影响标注的准确率。因此在实际应用中研究人员多对其进行改进以此来提高准确率。2010 年,Moon T 等[22]提出了一种利用边界条件的 HMM 模型,利用文本内容和功能词之间的不同对局部文本快速进行词性标注,提高了无监督词标注的准确率。2011 年,孙静等[23]提出了一种基于 CRF 模型的无监督词性标注方法,先利用词典对已分词文本进行词性标注,再利用 CRF 对语料进行迭代标注,逐步优化标注结果。2012 年,袁里驰[24]提出了一种改进的 HMM 词性标注方法,把马尔可夫族模型和句法分析结合进行词性标注,在相同测试条件下,马尔可夫族模型性能明显优于 HMM 模型。

基于统计的词性标注方法是目前应用较多的方法,通过对大规模语料库进行训练得到,覆盖面很广,标注结果有很好的一致性和较高的覆盖率,因此被广泛用于自然语言处理领域。但当训练语料达到一定规模后,通过扩充语料规模来提高正确率也变的不实际。需要设计更为严格精细的特征统计模板,以便参考更多的特征信息来进行词性标注。另外在进行标注的过程中要处理好概率参数的获取以及如何应用所获得的概率参数对文本进行词性标注等问题。

2.2 基于规则的词性标注方法

基于规则的词性标注方法是一种传统的方法,获取的规则集精度直接影响标注结果的优劣[25]。该方法能充分利用现有语言学成果,总结出许多有用的规则。先利用词典对语料进行基本切分和标注,列出该对象所有可能的词性,然后依据上下文信息,结合规则库消除歧义,最终保留唯一合适词性[26]。

基于规则的方法表达清晰,应用范围较广,但不能方便地通过机器学习来自动获取规则,人工构造又是一项艰难耗时的任务,如果把规则描述过细,规则的覆盖面就会大大减小,很难根据实际情况进行调整。如果不根据上下文仅根据规则判断词性又可能会出现歧义。它不属于统计模型,所以适应性也较差。而统计方法正好能弥补这个缺点,所以在实际应用中常把两者结合起来或对基于规则的方法进行改进。这样既能充分利用现有语言学的成果,还能利用统计模型来增强方法的适应性。标注的准确率也比单一使用一种要高。1995 年,Eric Brill[27]提出了基于转换的错误驱动的方法。利用初始标注器来标注训练语料库,得到的结果与正确结果进行比较,从中选出效果最好的变换模式作为系统的标注规则重新标注语料库,重复该过程直到获得所有规则,再用这些规则对待标注语料进行标注。2008 年,王广正等[28]提出基于规则优先级的词性标注方法,对每条词性标注规则加上优先级,通过控制优先级来完成兼类词的词性标注。2010 年,姜尚仆等[29]提出了一种基于规则和统计的分词和词性标注方法,使用基于单一感知器的联合分词和词性标注算法作为基本框架,以基于规则的词语邻接属性为特征。虽然该方法是针对日语进行词性标注的,但该方法的思想和技术同样值得中文词性标注研究人员借鉴。2010 年,陈小芳等[30]提出了一种基于统计和规则相结合的汉语术语语义分析方法。首先以词、词性、距离信息、上下文信息、词语的第一个原义信息为特征,基于这些特征得到支持向量机分析模型,在此基础上利用统计和规则相结合的方法进行术语语义分析。

表 4 是对规则、统计标注方法的比较。

表 4 三类词性标注方法的性能比较

方法名称	标注依据	机器效率	正确率	实用水平
基于规则的方法	规则库	一般	一般	不能达到
基于统计的方法	统计模型	低	较高	基本达到
规则与统计相结合的方法	规则库 + 统计模型	高	高	达到

3 分词和词性标注一体化模型

中文分词和词性标注有两种方案:先进行分词再进行词性标注;分词和词性标注一起进行。传统方法是分开处理这两个阶段,但分词的精度和词性标注的准确度密切相关,分词产生的错误可能会影响标注的准确率,有机地将两者结合起来有利于消除歧义和提高整体效率。文献[31]用实验证明将分词和词性标注统一在一个架构中,会大幅提升中文词法分析的性能。目前构建一体化模型的方法主要有基于实例的方法(Example Based Chinese word Segment and Tagging, EBST)和基于无向图模型的方法。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

较早对 EBST 进行研究的是香港城市大学的 Kit[32]。基本思想是在语料库中查找与分词和词性标注文本相匹配的片段当作候选结果,再按照某种准则对候选结果进行优化,从中找出最优的分词 - 词性标注序列。无向图模型给输入序列中每个位置上的字符都分配一个词法成分标注,再利用后处理规则对词法进行分析和处理。该模型中的任何一个连通子图,都可以作为依赖关系,能够大幅提高词法分析的性能。2007 年,姜涛等[33]构建了一个基于 EBST 的分词 - 词性标注系统,使用 best - first 算法对候选序列进行优化,最大词匹配器来处理实例匹配不能处理的数据稀疏情况。2010 年,朱聪慧等[34]实现了一种基于无向图序列模型的分词词性标注一体化系统。该系统在 1998 年人民日报语料上进行测试,分词的准确率达到 97.19%,词性标注的准确率达到 95.34%。

EBST 系统的语料库中的实例片断通常包含多个词,比基于词典的分词和词性标注方法考虑了更多的上下文信息,输出的结果也更可信[35];对于和训练语料相关的文本进行分词和标注会有极高的准确率,结果与训练语料中的分词和标注结果有很好的一致性。但该方法存在严重的数据稀疏问题,作为通用的分词和词性标注系统,它性能并不理想,需要做进一步的细化,比如构建规模大且包含多领域语料知识的语料库,笔者认为通过构造语料库该方法可以达到极好的性能。基于无向图模型的一体化模型将中文分词和词性标注以序列标注的形式,真正统一在一个架构中;它考虑了当前标注位置和已完成词法成分标注的位置之间的依赖关系,还考虑了更深层次的依赖关系来提升中文词法分析的性能。实验结果表明,该模型可以大幅提升中文词法分析的性能。现在越来越多的研究表明,将中文分词和词性标注统一起来进行处理比单一分开处理准确率要高[36],这样能充分利用两个阶段相互之间的依赖关系,有利于消除歧义和提高系统的整体效率。这对于今后进行中文词法处理提供了很好的启示和解决途径。

4　结束语

由于中文的独特性,目前还没有十分完善的中文分词和词性标注算法。分词和词性标注算法的进一步发展应该在已经取得的成绩的基础上,综合运用多种方法,引入新的模型和算法,通过不断探索,使其越来越完善。目前歧义切分和未登录词仍是分词和词性标注领域所面临的最大问题。因此下一步的工作将主要围绕以下两个方面展开:

(1) 充分发挥统计模型在解决未登录词和歧义切分问题的优势,对现有的统计模型进行优化或者设计更加合理的统计模型。基于已经取得的成绩,综合运用多种方法来完善分词和词性标注工作。

(2) 对于微博短文本来说,它数据规模较小,信息含量较少,内容比较口语化,还具有实时更新和动态变化的特点。与传统文本有相同更有许多不同之处,对其进行分词和词性标注就更加困难[37]。对此有必要在对其进行分词和标注之前进行预处理,将无结构的原始文本转化为结构化的、能被计算机识别和处理的信息表示形式。另外还需要对一般文本的特征选择算法和学习算法进行优化,利用主动学习或者半监督方式来解决短文本特征稀疏和数据分布不平衡等问题,弥补短文本信息量少的缺陷[38]。

参考文献:

[1] Wang Kun, Zong Chengqing, Su K Y. A character - based joint model for Chinese word segmentation [C] //Proceedings of the 23rd international conference on computational linguistics. Beijng: Association for Computational Linguistics, 2010: 1173 - 1181.

[2] Xue Nianwen, Shen Libin. Chinese word segmentation as LMR tagging [C] //Proceedings of the second SIGHAN workshop on Chinese language processing. [s. l.]: Association for Computational Linguistics, 2003: 176 - 179.

[3] 何国斌,赵晶璐. 基于最大匹配的中文分词概率算法研究 [J]. 计算机工程,2010,36(5):173 - 175.

[4] 张海营. 全二分快速自动分词算法构建 [J]. 现代图书情报技术,2007(4):52 - 55.

[5] Mai F J, Li D P, Yue X G. Research on Chinese word segmentation based on bi - direction marching method and feature selection algorithm [J]. Journal of Kunming University of Science and Technology: Natural Science Edition, 2011, 36(1): 47 - 51.

[6] 曹月雷,纪文彦,贾　斌. 词典与后缀数组相结合的中文分词方法 [J]. 硅谷,2012(21):151 - 154.

[7] Li Qinghu, Chen Yujian, Sun Jiaguang. A new dictionary mechanism for Chinese word segmentation [J]. Journal of Chinese Information Processing, 2003, 17(4): 13 - 18.

[8] Shi W. Chinese word segmentation based on direct maximum entropy model [C] //Proceedings of the fourth SIGHAN workshop on Chinese language processing. [s. l.]: [s. n.], 2005: 193 - 195.

[9] Wang Xinhao, Lin Xiaojun, Yu Dianhai, et al. Chinese word segmentation with maximum entropy and n - gram language model [C] //Proc of the 5th SIGHAN workshop on Chinese language processing. Morristown, NJ: ACL, 2006: 138 - 141.

[10] Zhang Liyan, Qin Min, Zhang Xuemei, et al. A Chinese word segmentation algorithm based on maximum entropy [C] //Proc of international conference on machine learning and cybernetics. Qingdao: IEEE, 2010: 1264 - 1267.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[11] Zhang Meishan, Deng Zhilong, Che Wanxiang, et al. Combining statistical model and dictionary for domain adaption of Chinese word segmentation [J]. Journal of Chinese Information Processing, 2012, 26(2): 8 – 12.

[12] Song H, Yao T. Active learning based corpus annotation [C] //Proc of IPS – SIGHAN joint conference on Chinese language processing. Beijing: [s. n.], 2010: 28 – 29.

[13] 朱红斌, 蔡 郁. 基于主动学习支持向量机的文本分类 [J]. 计算机工程与应用, 2009, 45(2): 134 – 136.

[14] 陈锦禾, 沈 洁. 基于信息熵的主动学习半监督分类研究 [J]. 计算机技术与发展, 2010, 20(2): 110 – 113.

[15] Li Shoushan, Zhou Guodong, Huang Churen. Active learning for Chinese word segmentation [C] //Proceedings of COLING 2012. Mumbai: [s. n.], 2012: 683 – 692.

[16] Sproat R, Emerson T. The first international Chinese word segmentation bakeoff [C] //Proceedings of the second SIGHAN workshop on Chinese language processing. [s. l.]: Association for Computational Linguistics, 2003: 133 – 143.

[17] Emerson T. The second international Chinese word segmentation bakeoff [C] //Proceedings of the fourth SIGHAN workshop on Chinese language processing. [s. l.]: Association for Computational Linguistics, 2005: 123 – 133.

[18] Levow G A. 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: [s. n.], 2006: 108 – 117.

[19] Ekbal A, Saha S. Simulated annealing based classifier ensemble techniques: application to part of speech tagging [J]. Information Fusion, 2013, 14(3): 288 – 300.

[20] Zin K K, Thein N L. Part of speech tagging for Myanmar using hidden Markov model [C] //Proc of international conference on the current trends in information technology. [s. l.]: IEEE, 2009: 1 – 6.

[21] 黄德根, 焦世斗, 周惠巍. 基于子词的双层 CRFs 中文分词 [J]. 计算机研究与发展, 2010, 47(5): 962 – 968.

[22] Moon T, Erk K, Baldridge J. Crouching Dirichlet, hidden Markov model: unsupervised POS tagging with context local tag generation [C] //Proceedings of the 2010 conference on empirical methods in natural language processing. [s. l.]: Association for Computational Linguistics, 2010: 196 – 206.

[23] 孙 静, 李军辉, 周国栋. 基于条件随机场的无监督中文词性标注 [J]. 计算机应用与软件, 2011, 28(4): 21 – 23.

[24] 袁里驰. 基于改进的隐马尔科夫模型的词性标注方法 [J]. 中南大学学报: 自然科学版, 2012, 43(8): 3053 – 3057.

[25] Saharia N, Das D, Sharma U, et al. Part of speech tagger for Assamese text [C] //Proceedings of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing of the AFNLP. Singapore: Association for Computational Linguistics, 2009: 33 – 36.

[26] Schmitz S. A note on sequential rule – based POS tagging [C] //Proceedings of the 9th international workshop on finite state methods and natural language processing. [s. l.]: Association for Computational Linguistics, 2011: 83 – 87.

[27] Brill E. Transformation – based error – driven learning and natural language processing: a case study in part – of – speech tagging [J]. Computational Linguistics, 1995, 21(4): 543 – 565.

[28] 王广正, 王喜凤. 一种基于规则优先级的词性标注方法 [J]. 安徽工业大学学报: 自然科学版, 2008, 25(4): 426 – 429.

[29] 姜尚仆, 陈群秀. 基于规则和统计的日语分词和词性标注的研究 [J]. 中文信息学报, 2010, 24(1): 117 – 122.

[30] 陈小芳, 张桂平, 蔡东风, 等. 基于统计和规则相结合的汉语术语语义分析方法 [C] //第六届全国信息检索学术会议. 出版地不详: 出版者不详, 2010: 481 – 488.

[31] Ng H T, Low J K. Chinese part – of – speech tagging: one – at – a – time or all – at – once? word – based or character – based? [C] //Proceedings of EMNLP 2004. Barcelona, Spain: Association for Computational Linguistics, 2004.

[32] Kit C, Pan H, Chen H. Learning case – based knowledge for disambiguating Chinese word segmentation: a preliminary study [C] //Proceedings of the first SIGHAN workshop on Chinese language processing. [s. l.]: Association for Computational Linguistics, 2002: 1 – 7.

[33] 姜 涛, 姚天顺, 张 俐. 基于实例的中文分词 – 词性标注方法的应用研究 [J]. 小型微型计算机系统, 2007, 28(11): 2090 – 2093.

[34] 朱聪慧, 赵铁军, 郑德权. 基于无向图序列标注模型的中文分词词性标注一体化系统 [J]. 电子与信息学报, 2010, 32(3): 700 – 704.

[35] Kit Chunyu, Liu Xiaoyue. An example based Chinese word segmentation system for CWSB – 2 [C] //Proc of SIGHAN 4. Jeju Island: [s. n.], 2005: 146 – 149.

[36] Zhang Yue, Clark S. Joint word segmentation and POS tagging using a single perception [C] //Proceeding of 46th annual meeting of the association for computational linguistics. Columbus, Ohio, USA: Association for Computer Linguistics, 2008: 888 – 896.

[37] Wang L, Wong D F, Chao L S, et al. CRFs – based Chinese word segmentation for micro – blog with small – scale data [C] //Proceedings of the second CIPS – SIGHAN joint conference on Chinese language. [s. l.]: [s. n.], 2012: 51 – 57.

[38] 王连喜. 微博短文本预处理及学习研究综述 [J]. 图书情报工作, 2013, 57(11): 125 – 131.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net