

· 复杂性科学 ·

基于上下文相关的未知实体词识别方法

夏 虎^{1,2}, 黄文茜²

(1. 电子科技大学大数据研究中心 成都 611731; 2. 电子科技大学互联网科学中心 成都 611731)

【摘要】 现有的未知实体词识别方法主要针对人名、地名、机构名等具有特定结构的实体词进行识别, 而随着电子商务和社交网络的快速发展, 出现了大量结构不确定的专有领域未知实体词。针对该问题, 提出两种基于上下文相关的未知词识别算法, 通过计算词(字)和词(字)之间的上下文相关性, 得到其潜在组合的支持度, 并通过过滤模块过滤掉错误的组合, 实现具有非确定型结构的未知实体词识别。实验表明, 该算法具有较高的准确率, 并且可以通过调整参数适应不同的应用场景。

关键词 关联规则; 上下文相关; 未知词识别; 词义消歧

中图分类号 TP181 **文献标志码** A **doi**:10.3969/j.issn.1001-0548.2016.05.022

Unknown Words Recognition Based on Context-Sensitive Algorithm

XIA Hu^{1,2} and HUANG Wen-qian²

(1. Big Data Research Center, University of Electronic Science and Technology of China Chengdu 611731;

2. Web Sciences Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Existing unknown words recognition methods mainly focus on unknown words with some specific structure, such as names, places and organizations. However, with the booming of e-commerce and social networking, more and more unknown entity words with uncertain structures appear in specific areas. In order to handle this problem, this paper presents two algorithms of unknown words recognition based on context-sensitive method. We first calculate correlations between any two words in sequence to get support of any potential combination, then filter out wrong combinations by filtering module, and achieve the recognition aiming at the non-deterministic structure of unknown words. Experiment results indicate that two algorithms can achieve a high accuracy. Besides, they can adapt to different application scenarios by adjusting the parameters.

Key words association rules; context-sensitivity; unknown word recognition; word sense disambiguation

命名实体是文本中承载信息的重要语言单位, 命名实体的识别在网络信息抽取、网络内容分析和知识工程等领域都占有非常重要的地位。传统的命名实体识别主要针对人名、地名、机构名以及产品命名实体等具有特定结构的实体词^[1]。然而, 随着互联网的快速发展, 网络上出现了大量结构不确定的专有领域未知实体词, 例如电子商务中大量出现的新商品名称、网络用语“酱紫(这样子)、斑竹(版主)”等, 这类未知词结构多样, 没有特定的规律, 用传统的未知词识别方法难以有效识别。

目前未知词识别领域的研究主要有3种方法: 基于统计的方法、基于规则的方法以及两者结合的方

法。基于统计的方法认为: 如果若干个相邻的字或词经常同时出现, 它们则可能是一个新词。这种方法简单高效易实现, 但需要大量训练数据, 而且由于未考虑不同词的构词能力^[2]和构词模式, 识别的准确率不高。基于规则的方法通过标注词典和成词规则来识别新词, 这些规则往往需要专家针对特定领域来具体制定, 该方法准确率高, 但规则制定费时费力, 且不同领域需要重新制定相应规则, 领域适应性差。针对上述两种方法中的问题, 越来越多的研究者采用统计与规则相结合的思路, 取得了许多显著的成果, 本文采用的基于上下文相关的算法即为其中一种。

收稿日期: 2015-02-06; 修回日期: 2015-06-15

基金项目: 国家自然科学基金(61250110543); 中央高校基本科研业务费(ZYGX2013J079, ZYGX2014Z012, ZYGX2011J067); 四川省科技项目(2012RZ0002, 2013TD0006)

作者简介: 夏虎(1981-), 男, 博士, 主要从事数据挖掘、复杂网络方面的研究。

一个字或词的上下文是指出现在它前后的那些字或词,在文本中相邻的字词共同出现的次数越多,它们越有可能是一个“未知词”,例如“清仓/圣/丽/奴/时尚/女/挎包”、“横款/圣/丽/奴/两用/包”、“高级/提花布/深/咖/圣/丽/奴/女/挎包”的分词结果可以看出,“丽”的上下文信息中总是包括“圣”和“奴”,也就是说“圣”、“丽”、“奴”3个字经常依此顺序共同出现,而“圣丽奴”整体并没有固定的上下文信息,因此本文认为“圣丽奴”有较大概率为一个未知实体词。

以上述理论为基础,本文提出了两种基于上下文信息进行未知词识别的方法。其中,基于最大组合的上下文相关算法(MC)利用统计的手段,获取由二元组、三元组、四元组、五元组构成的候选未知实体词集,然后利用上下文信息对候选未知实体词进行支持度过滤、歧义过滤和最大组合过滤,获取真正的未知实体词。

进一步,本文提出了一种基于关联规则的上下文相关算法(FPC),在FP树构建和频繁模式挖掘过程中加入各“项”(分词后的字或词)在文中出现的下标信息,利用此信息保证挖掘出的频繁模式中各项在文中的相邻关系以及前后顺序。从而避免了传统FP-growth算法不能保证挖掘出各项之间原始的相邻关系和前后顺序而不适合用于未知实体词识别的问题。

实验结果表明,在某电子商务网站的2 000个商品网页源文件上进行的3个类别数据集上,本文的两种方法均能有效地对结构不确定的专有领域未知实体词进行识别,具有较高的准确率。

1 相关研究

文献[3]提出了一种基于角色标注的中文未登录词识别通用方法。该方法依据角色,即未登录词的内部组成成分、上下文及句子中的其他成分来识别未登录词。算法简单可行,具备较好的准确率和召回率,尤其适用于中国人名和音译名的识别。

文献[4]提出了一种隐马尔科夫模型(hidden Markov model, HMM)和一个基于HMM的块标注器,并在此基础上建立了命名实体识别系统(NER)以识别姓名、时间以及数字量。系统整合了四方面的证据:词语包含的简单且确定性的内部特征,如大写、数字、触发器等内部语义特征以及外部上下文特征。该系统在蛋白基因(MUC-6和MUC-7)的英文命名实体识别任务中分别达到了96.6%和94.1%的

准确率。

文献[5]提出了一种基于支持向量机(SVM)的命名实体识别系统。该系统从文档中提取名称、数字信息并将其分类成人名、组织名以及日期。该系统取得了较高的准确率,并且解决了传统SVM效率不高的问题。文献[6]则提出利用SVM进行生物医学命名实体识别。该系统采用了字词缓存以及HMM状态两个新特征,在GENIA语料库上取得了令人满意的结果。

文献[7]提出了一种组合分类器的实验框架以识别命名实体。该框架组合了4个不同的分类器:鲁棒的线性分类器、最大熵模型、迁移学习及隐马尔科夫模型。文献[8]提出基于最大熵模型的命名实体识别系统,该系统直接利用整篇文档的全局信息来分类每一个具体的词,并且仅使用了一个分类器而不是二级分类器。

文献[9]提出了一种基于网络资源的未登录词扩展识别方法。该方法利用统计的思想,以左右邻信息判断未登录词边界,对已识别出的二元候选未登录词进行扩展,找出具有更完整语义的无限长度复合未登录词。该方法简单高效,但没有充分考虑不同词的构词能力和构词模式,容易因成词率低的高频词引发扩展错误,因此准确率不高。

文献[10]提出了一种基于统计和规则的未登录词识别方法。该方法将文本分词后的碎片切分形成临时词典,再利用规则和词频对其赋以不同的权值,最后用贪心算法得到碎片的最长路径,从而识别出未登录词,并进一步利用互信息提取若干个词组成未登录词(组)。该方法能正确识别出碎片中的大部分未登录词,但是识别正确性依赖于分词性能且对人名的识别规则不够完善。

文献[11]提出先将文本进行分词,再利用N-Grams方法得到候选未登录词集,之后通过概率统计的手段从中识别出未登录词。但这种方法在各个阈值的设定、中文词组的确定规则以及噪音字的选取方面仍需进一步完善。

综上所述,目前未知词识别的研究对象主要集中在人名、地名、机构名或者产品命名实体等具有特定结构的实体词上,对于近几年网络中出现的大量结构不确定的专有领域未知实体词的研究较少,本文特针对该问题提出两种识别方法。

2 基于最大组合的上下文相关算法(MC算法)

一个字或词的上下文是指出现在它前后的那些

字或词, 在文本中相邻的字词共同出现的次数越多, 它们越有可能是一个“未知词”。本文算法充分利用字词的上下文关系统计获取候选未知词集, 然后通过支持度过滤、歧义过滤以及最大组合过滤筛选出最终的未知词, 具体流程如下:

- 1) 对于输入文档集 D 中的任一文档 d , 首先将文本中“、,.;:”5种标点替换为换行符得到文档 d' ;
- 2) 对文档 d' 分词, 得到文档 d'' , 将 d'' 中的每个词/字作为基本单位“项”, 对于每一行文本, 统计该行相邻项之间形成的 n 元组($2 \leq n \leq 5$)出现的次数count, 形成集合 $P <n$ 元组, count>;
- 3) 将 P 中具有相同 n 元组的count值合并, 作为该 n 元组在文档 d 中的总支持度, 并过滤掉count<minSup的组合, 其中minSup是算法设定的最小支持度参数;
- 4) 进行歧义过滤(参考2.1节)及最大组合过滤(参考2.2节), 得到最终识别出的未知实体词;
- 5) 相同未知词可能出现在单一文档的不同位置, 也可能出现在文档集 D 的任一文档中, 因此需要针对所有文档遍历完后得到的组合集totalPat中再进行一遍歧义过滤和最大组合过滤。最后得到的结果保存在未知词集unKnown中, 算法结束。

2.1 歧义过滤

歧义过滤是指若识别出两个“歧义组合”, 仅保留count值最大的未知词组合。歧义组合定义如下:

定义 1 歧义组合

两个候选未知词组合 $U_1\{w_i, w_{i+1}, \dots, w_j\}$ 和 $U_2\{w_m, w_{m+1}, \dots, w_n\}$, 其中 w 为单字, U_1 和 U_2 为字串, 若满足 $w_j = w_m$ 或者 $w_i = w_n$, 则认为组合 U_1 和 U_2 是歧义组合。

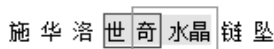


图1 歧义组合示例

如图1所示, 在“施华洛世奇水晶链坠”的分词字符串中, “世奇”和“奇水晶”就是一对歧义组合, 两种划分方式必然只有一种正确。根据“世奇”与“奇水晶”在全文中的支持度, 可以过滤掉支持度较低的“奇水晶”这样的错误组合。

2.2 最大组合过滤

最大组合过滤是指若识别出若干个具有“歧义父子串关系”的组合, 则保留歧义父串而去掉歧义子串。歧义父子串关系定义如下。

定义 2 歧义父子串

识别出两个候选未知词组合 $U_1\{w_i, w_{i+1}, \dots, w_j\}$ 和 $U_2\{w_m, w_{m+1}, \dots, w_n\}$, 其中 w 为单个的字, 如果 U_1 和 U_2 具有相同的支持度, 且 U_1 中存在子串 $U'\{w_p, w_{p+1}, \dots, w_q\} (i \leq p \leq q \leq j)$ 与 U_2 完全相同, 则 U_1 为歧义父串, U_2 为歧义子串。

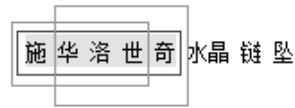


图2 歧义父子串示例

如图2所示, 在“施华洛世奇水晶链坠”的分词字符串中, “施华洛世奇”与“施华洛世”、“华洛世奇”、“华洛世”等具有相同的支持度, 构成了歧义父子串关系, 根据最大组合过滤规则只保留“施华洛世奇”这一歧义父串组合。

2.3 MC算法总结

基于最大组合的上下文相关算法MC利用统计信息构造候选未知词集, 然后通过支持度过滤、歧义过滤以及最大组合过滤, 删除候选未知词集中绝大部分错误的候选词, 从而识别出正确的未知实体词。

MC算法简单高效, 可以有效识别出网页中的未知实体词。MC算法的主要思想是认为在文本中相邻的字词共同出现的次数越多, 它们越有可能是一个“未知词”。而关联规则算法是挖掘数据项共同出现关系的经典算法。因此, 下文基于关联规则的上下文相关算法FPC提出利用关联规则挖掘字词间的共现关系来识别未知实体词。

3 基于关联规则的上下文相关算法(FPC算法)

FP-growth算法^[12]是一种高效的关联规则挖掘算法, 但是由于未保证挖掘出的频繁模式中各项间的相邻关系和前后顺序而不适合直接用做未知词识别。本文提出的基于关联规则的上下文相关算法改进了FP-growth算法, 在FP树构造过程以及频繁模式挖掘过程均充分利用了文档中各项出现位置的下标信息, 有效地保证了所挖掘频繁模式中的各项间具备正确的相邻关系以及前后顺序, 亦即保证了识别出的未知词在上下文意义上的正确性。

与MC算法类似, 本文算法首先对输入文档集 D 中的每一个文档 d 进行文本切分处理, 即将其中的“、,.;:”5种标点换为换行符得到文档 d' , 分词后得到文档 d'' 。 d'' 中每一个分词后的单位称为“项”, 每一行称为一条“事务记录”。为了存储每个项在文

档 d'' 中出现的所有位置的下标,将每一项的数据结构定义为 $\text{Item}\{\text{name},\text{index},\text{flag}\}$,其中 name 是该项的名字, index 是该项在文档 d'' 中出现的位置编号数组, flag 是排序的标志,用于将之后挖掘出的频繁模式按照在文中出现的先后顺序排序。对于文档 d'' ,FPCTree构造与频繁模式挖掘的过程如下。

1) FPCTree的构造

① 扫描文档 d'' ,得到频繁1项集,对它们的支持度计数,统计 index 信息,将频繁1项集按照支持度递减排序,若支持度相同,则按照各项在文中出现的先后顺序排序。删除支持度小于 minSup 的项,得到 $F1$ 项集。

② 构造树的结构为: $\text{TreeNode}\{\text{item},\text{count},\text{parent},\text{children},\text{nextHomonym}\}$,其中 item 表示该树节点对应的项, count 是该节点的支持度, parent 为父节点, children 是子节点数组, nextHomonym 指向下一个同名节点。创建树的根节点 T ,以 null 标记。

③ 第二次扫描文档 d'' ,每条事务记录中的项按照 $F1$ 中的顺序排序,设排序后的频繁项表为 $[p|P]$,其中 p 为频繁项表的第一项, P 为频繁项表中的剩余项。调用函数 $\text{addNode}(T,[p|P])$ 递归的将每一项加入到FP树中。 $\text{addNode}(T,[p|P])$ 执行过程如下:首先判断 T 的儿子节点中是否存在 p 的同名节点,即存在一儿子节点 t ,满足 $t.\text{Item.name} = p.\text{Item.name}$ 。若存在,则 t 节点的 count 计数加1,将 p 节点 index 数组中的所有下标加入到 t 节点的 index 数组中去;若不存在,则创建一个新节点 t ,将其 count 值设为1,链接到它的父节点 T ,并通过 nextHomonym 链接到下一个同名节点。将 t 加入到 T 的子节点数组中。

2) 从FPCTree中挖掘候选频繁模式

对 $F1$ 中的每一项 item 执行以下步骤:

① 生成条件模式基。利用 nextHomonym 信息,找到所有 item 同名节点的祖先路径,路径上所有节点 count 值均设为 item 的 count 值。

② 构建条件FP树。将条件模式基作为事务记录生成条件FP树。

③ 对于条件FP树中的每一条长路径生成项的任意组合方式,得到组合集 P 。过滤掉 P 中支持度小于 minSup 的组合,得到组合集 P' 。对于 P' 中的每一个组合,利用各项的 index 信息判断组合的上下文顺序是否正确。若正确,则获取该组合的支持度,并且将该组合按照在文中出现的先后顺序排序;若不正确,删掉该组合。得到候选频繁模式集 Pat 。

④ 挖掘出所有 item 的候选频繁模式后,将相同

的模式合并。

⑤ 识别出文档 d'' 中的候选未知词集 Pat 后,同MC算法一样,仍然需要在文档内部以及文档间进行歧义过滤与最大组合过滤,得到最终的未知词集 unKnown ,算法结束。

4 实验与分析

4.1 实验数据和工具

本文利用爬虫程序采集了某电商网站2 000个商品源文件,涉及项链、凉鞋、包、羽绒服、帽子、连衣裙、围巾、灯饰、针织衫和牛仔裤等10个类别的商品,每个类别中商品数量均为200。按商品类别等比例选取其中1 000份作为数据集1,剩余1 000份作为数据集2。

实验首先针对网页进行数据预处理,去除包括网页标签在内的无效字段,处理过程非本文重点,在此不再赘述。

为检验本文算法对不同分词工具的适应性,实验过程分别采用MMAnalyzer和IKAnalyzer^[13]进行测试。本文实验采用Precision(准确率)和Recall(召回率)作为评价指标。

4.2 实验过程及结果

1) 不同数据集结果比较

表1为MC算法和FPC算法使用不同分词工具在不同数据集上识别效果。对于每一个(算法,分词工具,数据集)的组合,随着支持度阈值 min_sup 阈值的增加,Precision和Recall也不断变化,表1中所有结果均选取最佳识别效果时的准确率召回率。其中MMAnalyzer和IKAnalyzer分词工具分别简称为MM和IK。

表1 不同数据集上的结果

复合特征模板	特征模板含义描述		
MC算法-MM	准确率/%	92.90	91.50
	召回率/%	60.30	57.90
MC算法-IK	准确率/%	99.60	95.80
	召回率/%	68.80	67.40
FPC算法-MM	准确率/%	93.60	91.00
	召回率/%	49.70	50.20
FPC算法-IK	准确率/%	100	94.80
	召回率/%	60.50	55.70

由上表可以看出:对于MC算法、FPC算法、MMAnalyzer分词工具、IKAnalyzer分词工具的任意组合,均有较好的准确率和召回率。

2) 不同分词工具结果比较

观察两个算法在分别使用两个分词工具时识别结果的好坏, 实验结果如图3所示。

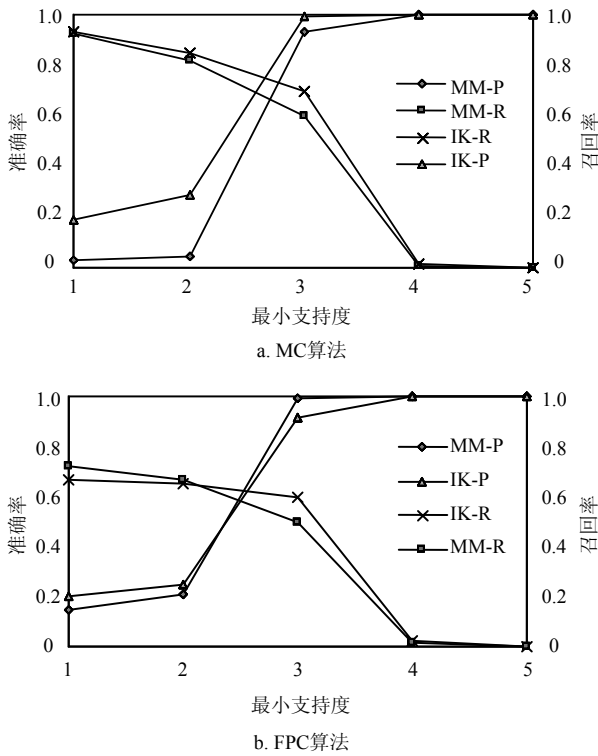


图3 两个算法分别在分词工具上的对比

由图中可以看出, MC算法和FPC算法在两个分词工具上Precision和Recall的走势一致, Precision随着最小支持度参数min_Sup的增加而呈现上升趋势, 在min_Sup=3时突变到一个高点, 并在min_Sup>3后趋于稳定; Recall随着min_Sup的增加而呈现下降趋势, 在min_Sup=4时突变到0%附近, 并在之后稳定于0%。

准确率突变点的存在是因为电商网站商品网页经过数据预处理后的待识别的未知词支持度普遍大于等于3, 而其他候选未知词中错误的未知词的支持度普遍小于3, 从而导致当min_Sup<3时识别出许多错误的未知词并拉低准确率。召回率突变类似。

MC算法和FPC算法在使用IKAnalyzer分词工具时, 均可以得到更好的准确率和召回率。这主要是由于算法1和算法2均先对输入文本进行了分词处理, 分词的效果将直接影响到未知词识别的效果。如果分词工具将一个待识别未知词的某一部分和其他词分到了一起, 则通过两个算法都无法识别出正确的未知词。例如, 若未知词xyz(其中x、y、z为单字或者字串)被分成了xy和zh, 则经过算法1和算法2都无法识别出xyz, 而分成xy和z则可以很容易地被两

个算法识别出来。IKAnalyzer分词工具比MMAnalyzer分词工具更能避免此类错误的分词结果, 故而具备更高的准确率, 又由于在同等情况下能识别出更多的未知词而具备更高的召回率。算法表现仍然依赖于分词效果, 粒度越细的分词工具理论上将获得越好的表现。

3) 算法的对比

将使用相同分词工具时两个算法的结果进行对比, 如图4所示。

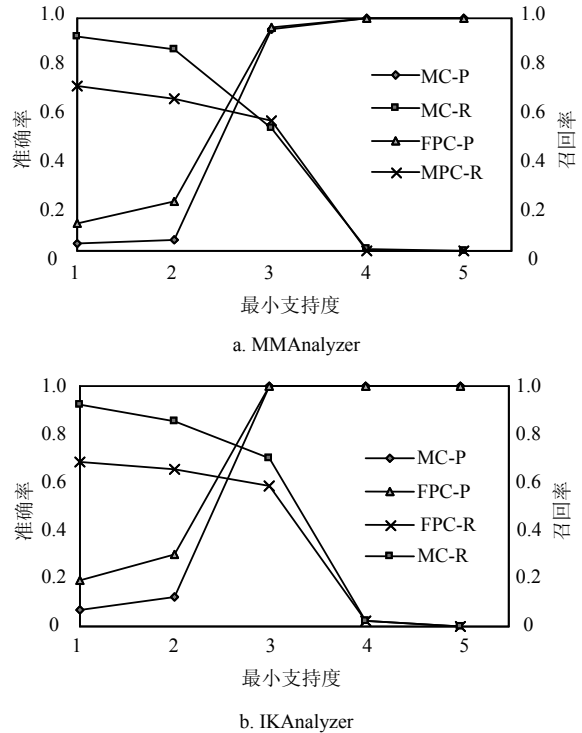


图4 使用相同分词工具时算法间的对比

由图4可以看出, FPC算法准确率明显优于MC算法, 但召回率则明显弱于MC算法。由于本文所述的未知词识别更为强调较高的准确率, 因此本文实验最终选取minSup=3, 牺牲部分召回率换取令人满意的准确率。

综合整个对比分析过程, 本文实验中最终未知词识别的最佳组合方式为: FPC算法, IKAnalyzer分词工具, min_Sup=3。

5 结束语

本文针对网络中新出现的大量未知实体词, 提出了两个未知词识别算法: 基于最大组合的上下文相关算法(MC)和基于关联规则的上下文相关算法(FPC)。两个算法均充分利用了字词的上下文关系信息, 可以有效识别专有领域具有非确定型结构的未知实体词, 对于只能识别具有特定结构实体词的现

有算法是一个很好补充。

实验表明,本文算法具有较高的准确率。同时,算法可通过调整支持度阈值参数 min_sup ,从而适应不同的应用场景,具备一定的通用性。

本文两个算法中均用到了歧义过滤和最大组合过滤,然而两种过滤方法均不能完全保证过滤的正确性,如何充分利用词的构词模式和构词能力形成新的过滤方法是下一步的研究内容之一。另外,网页噪声处理有多种不同的方法,多种方法对于未知词识别效果的影响也是下阶段研究的重要内容。

参 考 文 献

- [1] 秦文, 苑春法. 基于决策树的汉语未登录词识别[J]. 中文信息学报, 2004, 18(1): 14-19.
- [2] QIN Wei, YUAN Chun-fa. Identification of Chinese unknown word based on decision tree[J]. Journal of Chinese Information Processing, 2004, 18(1): 14-19.
- [3] 王文荣, 乔晓东, 朱礼军. 针对特定领域的新词发现和新技术发现[J]. 现代图书情报技术, 2008, 161(2): 35-40.
- [4] WANG Wen-rong, QIAO Xiao-dong, ZHU Li-jun. New word and technology discovery of specific domain[J]. New Technology of Library and Information Service, 2008, 161(2): 35-40.
- [5] ZHANG K, LIU Q, ZHANG H, et al. Automatic recognition of Chinese unknown words based on roles tagging[C]//In SIGHAN '02: Proceedings of the First SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics. Stroudsburg: ACM Press, 2002: 1-7.
- [6] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger[C]//In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACM Press, 2002: 473-480.
- [7] ISOZAKI H, KAZAWA H. Efficient support vector classifiers for named entity recognition[C]//In COLING '02: Proceedings of the 19th International Conference on Computational linguistics. Stroudsburg: ACM Press, 2002: 1-7.
- [8] KAZAMA J, MAKINO T, OHTA Y, et al. Tuning support vector machines for biomedical named entity recognition[C]//In BioMed '02: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. Association for Computational Linguistics. Stroudsburg: ACM Press, 2002: 1-8.
- [9] FLORIAN R, ITTYCHERIAH A, JING H, et al. Named entity recognition through classifier combination[C]//In CONLL '03: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Stroudsburg: ACM Press, 2003: 168-171.
- [10] CHIEU H L, NG H T. Named entity recognition: a maximum entropy approach using global information[C]//In COLING '02: Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg: ACM Press, 2002: 1-7.
- [11] 韩艳, 林煜熙, 姚建民. 基于统计信息的未登录词的扩展识别方法[J]. 中文信息学报, 2009, 23(3): 24-30.
- [12] HAN Yan, LIN Yu-xi, YAO Jian-min, Study on Chinese OOV identification based on extension[J]. Journal of Chinese Information Processing, 2009, 23(3): 24-30.
- [13] 周蕾, 朱巧明. 基于统计和规则的未登录词识别方法研究[J]. 计算机工程, 2007, 33(8): 196-198.
- [14] ZHOU Lei, ZHU Qiao-ming. Research on recognition method of unknown Chinese words based on statistic and regulation[J]. Computer Engineering, 2007, 33(8): 196-198.
- [15] 韩洁, 周勇, 刘少辉, 等. 基于WWW的未登录词识别研究[J]. 计算机科学, 2002, 29(12): 155-156.
- [16] HAN Jie, ZHOU Yong, LIU Shao-hui, et al. WWW-based recognition of non-login words[J]. Computer Science, 2002, 29(12): 155-156.
- [17] HAN J, KAMBER M, PEI J. Data mining: Concepts and techniques[M]. San Francisco: Morgan Kaufmann, 2006.
- [18] WANG Kun-shan. IKAnalyzer[EB/OL]. [2015-01-17]. <https://github.com/wks/ik-analyzer>.

编辑 蒋 晓