计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):169−174 [doi: 10.15888/j.cnki.csa.006676]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于混合模型的新闻事件要素提取方法①

虞金中, 杨先凤, 陈　雁, 李　娟

(西南石油大学 计算机科学学院, 成都 610500)
通讯作者: 陈　雁, E-mail: carly.chenyan@gmail.com

摘　要: 为了帮助读者从大量新闻报道信息中迅速地把握其主要内容, 本文分析了事件要素对新闻主要内容的影响, 结合新闻报道的基本原则和要求, 提出了一种基于混合模型的事件要素提取方法. 该方法首先对新闻数据中识别的实体进行加权, 然后使用依存句法树分析实体在新闻事件中扮演的角色, 并对关于要素的指代现象进行消解, 最终融合频率及角色关系对实体加权的方法进行改进, 有效地提取出新闻事件关联性较为重要的要素. 实验结果表明, 本文所述方法能够准确地提取出与新闻事件关联性较强的事件要素, 提高了读者快速筛选新闻事件要素的效率.

关键词: 中文命名实体识别; 词性标注; 条件随机场; 依存句法分析; 混合模型

引用格式: 虞金中,杨先凤,陈雁,李娟.基于混合模型的新闻事件要素提取方法.计算机系统应用,2018,27(12):169–174. http://www.c-s-a.org.cn/1003-3254/6676.html

News Event Element Extraction Method Based on Mixed Model

YU Jin-Zhong, YANG Xian-Feng, CHEN Yan, LI Juan

(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract: In order to help readers quickly grasp the main content of a large amount of news report information, this paper analyzes the impact of event elements on the main news content, and combines the basic principles and requirements of news reports, proposes a method of extracting event elements based on hybrid model. The proposed method first weighs the entities recognized in the news data, and then uses the dependency syntax tree to analyze the role of entities in news events, and dispels the reference phenomenon of elements. Finally, the fusion frequency and role relationship are used to improve the entity weighting method and effectively extract the important elements of news event relevance. The experimental results show that the method described in this study can accurately extract event elements with strong relevance to news events and improve the efficiency of readers' rapid selection of news event elements.

Key words: Chinese name entity recognition; POS tagging; Conditional Random Fields (CRF); dependency syntax; hybrid model

1　概述

近年来, 随着数据库技术和网络技术的广泛应用, 新闻文本数据增长迅速, 数据的种类也逐渐增多. 在这些海量的文本信息中, 仅有很少的一部分信息是刻画新闻事件的主要信息, 因此对于每天接触大量信息的现代人, 快速筛选有用信息, 提取事件要素, 提高阅读效率, 无疑是很有意义的.

现有的新闻事件要素提取方法容易受到新闻数据稀疏性的影响, 虽然基于语义分析实现效果不错, 但是可移植性差、对语料库有很大的依赖性. 针对提取新闻事件要素存在的不足, 许多研究者提出了改进的算法. 裴东辉等人[1]提出了通过新闻中的子事件与事件因

① 基金项目: 国家自然科学青年基金项目 (61503312)
Foundation item: Young Scientists Fund of National Natural Science Foundation of China (61503312)
收稿时间: 2018-05-03; 修改时间: 2018-05-24; 采用时间: 2018-06-14; csa 在线出版时间: 2018-12-03

© 中国科学院软件研究所 http://www.c-s-a.org.cn

素的关联性抽取新闻要素的方法, 以子事件元素与元素间关联关系分别表征为节点、边, 构建新闻事件提取无向图模型. 最后, 求解无向图中节点的权重, 实现对新闻事件要素的提取. 该方法没有涉及新闻中的子事件之间的关联关系. 朱青等人[2]提出了一种通过生成标题的要素关联树对包含地点进行关联度评价的方法, 依次从新闻正文中抽取地点要素. 该方法由于依赖于地名关系数据库, 因此具有对地名因素抽取的细粒度有限、可移植性不高的缺点. 涂子令等人[3]提出了一种基于超图的 PageRank 随机游走的方法提取新闻话题要素, 通过该方法计算后, 对新闻事件要素集合给出一个信息重要性的排序. 由于这类方法没有考虑新闻数据中的指代, 容易产生错误.

通过对中文新闻数据进行提取关联事件要素方面的分析与研究, 本文提出了一种混合模型提取事件要素的方法 ERCDSPEE(Extraction of event elements entity recognition combining dependency syntactic parsing), 实质是综合新闻内容实体识别、依存句法分析提取新闻事件要素. 本文以提取事件人名要素为例对方法进行分析验证, 首先, 通过命名实体识别[4,5]技术识别出相关新闻事件中的人名实体, 对新闻数据中的实体进行加权, 然后使用依存句法树[6,7]分析实体在新闻事件中扮演的角色, 并且对关于要素的指代现象进行消解, 进一步根据改进的 Sigmoid 函数对事件要素赋予权重, 有效地提取出新闻事件关联性较为重要的人名要素.

2 相关工作

2.1 挖掘要素的方法

新闻文本中通常包含一些描述事件发生的对象、时间、地点等要素信息, 但是怎么对数据所隐藏的价值进行充分挖掘和利用, 带着这样的思路对新闻数据进行深入分析, 有利于找到解决问题的关键. 考虑到新闻事件中人名实体的比重以及人名实体与事件的关联关系, 本文提出得研究方法 ERCDSPEE 是在实体识别[8]和依存句法算法的基础上构建一个抽取刻画事件要素的模型, 实现了新闻要素的提取.

2.2 命名实体识别方法与依存句法分析

2.2.1 命名实体识别

命名实体识别 (NER) 是自然语言处理 (NLP) 的一个基础任务, 它的目的是识别文本数据中时间、人名、地名、组织机构名等命名实体. 本文使用条件随机场[9] (Conditional Random Field, CRF) 模型进行实体识别, 条件随机场是由 Lafferty[10] 等人在最大熵模型和 HMM 模型的基础上提出的统计序列标注算法. 条件随机场模型不仅放宽了 HMM 模型的条件独立性, 在一定程度上, 还解决了标记偏置的问题, 并且具有时间复杂度低、准确度高等优点.

CRF 是一种概率无向图模型, 它能够被用来定义在给定一个观察序列 x 的条件下, 标记序列 y 的条件概率 $P(y|x)$, 是一种判别模型. 但在现实应用中, 尤其是对标记序列建模时, 最常采用线性链 (linear-chain) CRF 模型, 其图模型如下图 1 所示的结构.

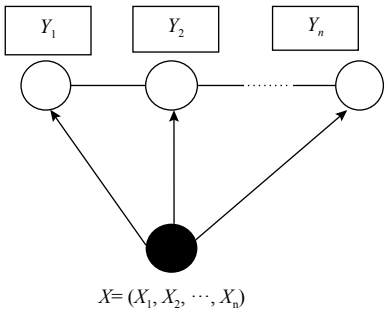

图 1 链式条件随机场的图结构

给定观测序列 x, 图 1 所示的链式 CRF[11]主要包括单个标记变量$\{y_i\}$和其相邻的标记变量$\{y_{i-1}, y_i\}$两种. 关于标记变量的团在条件随机场中, λ_j 通过选用指数势函数并引入特征函数, 条件概率被定义为:

$$P(y|x)=\frac{1}{Z}+\exp\left(\sum_j\sum_{i=1}^{n-1}\lambda_j t_j(y_{j+1},y_j,x,i)+\sum_j\sum_{i=1}^{n}u_k s_k(y_i,x,i)\right) \quad (1)$$

其中, $t_j\{y_{i+1}, y_i, x, i\}$是在观测序列的两个相邻标记位置处定义的转移特征函数, 其目的是表示相邻标记变量之间的相关性和观测序列对它们的影响, $s_k\{y_i, x, i\}$是定义在观测序列的标记位置 i 处的状态特征函数, 以此表示观测序列对标签标量的影响, λ_j 和 u_k 为参数, Z 为规范化因子, 用于确保式 (1) 是正确定义的概率.

2.2.2 依存句法分析

依存句法分析是基于依存句法的一种自动句法分析方法, 它将句子解析成一颗依存句法树, 描述出句子中词与词之间直接关系, 这种关系被称为依存关系, 一个依存关系连接两个词 (核心词和修饰词). 在依存句法树中不含终节点, 只有由具体词构成的终结点, 一条

© 中国科学院软件研究所 http://www.c-s-a.org.cn

依存边连接两个节点, 核心词所对应的节点为父亲节点, 而修饰词所对应的节点为树中的孩子节点. 两个词之间的依存关系可以细分为十几种类型, 如主谓关系(SBV)、并列关系 (COO)、动宾关系 (VOB) 等等. 例如,

依存句法分析的任务是针对已经分词和词性标注完成的句子, 进行其依存句法结构的分析. 给定输入为一个分词、词性标注完的句子, 进行依存句法分析后, 得到一个依存句法树. 依存句法分析器的输入如图 3 所示.

经过依存句法分析之后, 结果如图 2 所示. 其中小海 (修饰词) 和吃 (核心词) 之间存在依存关系 SBV(主谓关系), Root(核心词) 和吃 (修饰词) 之间存在依存关系 HED(核心关系), 吃 (核心词) 和鱼 (修饰词) 之间存在依存关系 VOB(动宾关系).

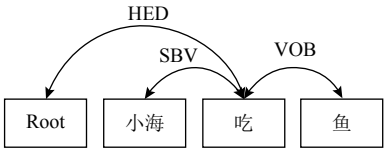

图 2 依存句法分析例子

图 3 依存句法分析器输入格式

3 提取中文新闻事件要素

3.1 语料收集与语料自动标注方法

首先抽取中文新闻实体, 然后分析新闻事件句[12]中的重要要素. 本文选用 1998 年人民日报语料作为实验语料, 将该语料分成训练语料和测试语料, 大小为 80% 和 20%. 通过训练语料建立实体识别模型, 使用测试语料测试模型, 准确率达到 97%. 以网络爬虫抓取的新闻数据作为实验测试数据, 其来源网站包括微博、头条、搜狐新闻、网易新闻、新华网, 该数据有 86 655 篇新闻.

由于中文新闻文本内部人名[13]关系不多, 名称形成的规律性不突出, 单词词性的识别需要基于准确的分词结果. 如果分析不明确, 相反, 它会干扰识别过程和结果, 因此这个实验任务是在单词级粒度进行建模, 1 个单词是一个标记. 中文实体识别任务是一个序列标注任务, 本文使用 4tag(S 表示单个词、B 表示词首、M 表示词中、E 表示词尾) 的标注方式来确定序列标注集. 通过 1998 年人民日报语料训练的模型识别新闻文本实体的效果并不是很理想, 其原因是当今新闻文本中出现很多新颖的名字等因素. 为解决此问题, 本文采用增加新语料来提高模型准确率的方法[14], 首先使用已训练好的模型测试少量的新闻数据, 并对其错误的词性标注进行手动修改标注, 然后把修改后的语料扩充到已有的训练数据来训练新模型, 再使用新模型测试少量的新闻数据, 循环往复, 最终获得性能良好的模型.

3.2 提取刻画新闻事件要素

为了从大量且繁杂的数据中挖掘出与新闻事件关联性较为强的人名, 本文基于 ERCDSPEE 方法构建出提取刻画新闻事件人名要素的模型. 提取刻画新闻事件要素的对象即针对新闻文本数据, 提取刻画新闻事件要素的具体流程如图 4 所示. 在识别新闻实体的基础上, 通过对新闻文本进行依存分析, 消除不同关系类型的人称代词, 进一步调整模型的参数, 使模型能够有效识别新闻人物与新闻事件的关联性; 最后, 把依存分析的要素与实体要素权重相融合, 实现新闻事件人名要素的抽取.

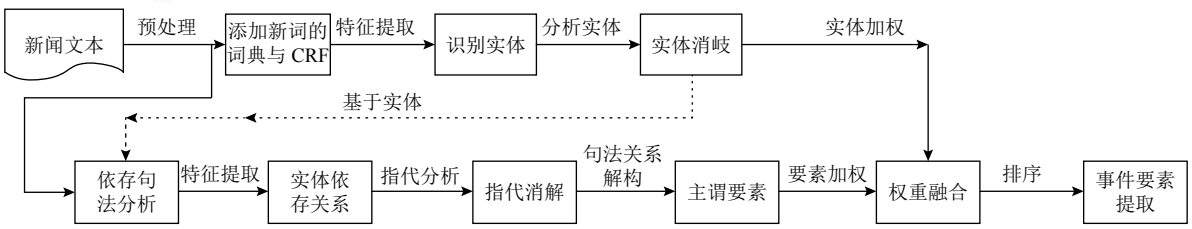

图 4 提取新闻事件要素的流程

3.2.1 构建识别新闻实体模型

首先基于命名实体规则挖掘的相关概念、过程和方法, 使用了工具 CRF++(CRF++是一个 CRFs 模型的实现) 提取新闻文本中的实体.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

单一的 CRF 是根据词之间关系、词性等特征来区分专有名词和非专有名词, 难以识别一些特征不明显的专有名词. 结合具有新词的自定义词典的 CRF 能识别出来一些特征不明显的人名, 词典可以自定义扩充那些特征不明显的人名和新颖的人名新词, 在正确分词方面具有良好的可控性, 可以提高了抽取实体人名的准确率. 针对人名实体识别存在的不足, 本文采用词典与 CRF 相结合的方法来识别实体; 通过生成的命名实体识别模型, 实现了对新闻文档人名实体的抽取, 进一步进行实体消岐, 从而对相同的实体进行统计并通过公式 (2) 和 (3) 对实体特征赋予权重.

关于新词的识别, 根据 Qiu[15]等提出一种中文未知单词自动 POS 猜测的方法建模. 首先使用机器学习方法根据其内部组件特征预测未知词的 POS, 然后测量预测结果的可信度, 对于低可信度的单词, 进一步根据这些单词的全局上下文信息对标注结果进行校正. 使用模型对当代的新闻文本数据抽取出新词, 进一步结合词典更新词语.

3.2.2 依存句法分析及指代消解

虽然识别出了某新闻事件中大量的人名, 但是哪些人物是此新闻事件句主要刻画的人? 通过依存句法树分析实体在新闻事件中扮演的角色, 并根据其角色有效地提取出新闻事件关联性较为重要的要素. 本文由于依存树的结构复杂, 分析文本句子时间复杂度比较高, 因此使用基于神经网络的高性能依存句法分析器[16]来分析实体之间的依存关系.

考虑到两个词之间的依存关系可以细分为十几种类型[17], 通过对新闻事件中人物特点 (例如, 人物在事件中的形态多样性) 的分析, 本文只考虑与中文新闻事件直接相关联的关系 (主谓关系与非主谓关系), 主谓关系包括施事关系、当事关系, 非主谓关系包括涉事关系、受事关系以及源事关系等.

通过依存句法分析器提取新闻事件的因素、因素词性、关系类型. 虽然根据句中的词与词之间的关系可以分析出这一句话中的事件人物, 但是句中含有一些动宾关系、介宾关系、主谓关系等的人称代词, 它对提取刻画事件人物有一定影响. 结合以上描述, 本文考虑到人称代词的词性及不同的关系类型对句中事件关联性人物的影响, 可以通过对人称代词的处理来加重事件人物权重的方法, 不仅对人称代词进行了消解, 而且更能考虑到人物在事件中扮演角色的重要性. 综上所述, 消除人称代词方法如下:

(1) 首先如果句中人名是主谓关系或者非主谓关系, 并且句子中含有人称代词, 然后对此人名的权重增加 1.

(2) 在句中没有人名且含有人称代词的基础上, 尽管句中含有一系列的职位及称呼的名词, 但是考虑到事件人物的多样性, 实行零代词消解, 更能提高识别的容错率.

(3) 句中含有人名和人称代词, 经判断得知识别的人名是一个单个姓氏词, 为减少人名识别的错误率, 使用 jieba 抽取的人名, 同 (1) 可以达到消除人称代词的效果.

3.2.3 权重融合及要素提取

一般一篇人物报道新闻讲究绿叶配红花的原则, 在人物报道中, 主要人物是红花, 次要人物是绿叶, 通过次要人物的活动衬托主要人物, 可以使主要人物形象更加鲜明. 如果文档中的人物是事件关联的主要人物, 他 (她) 一定会在文档重复出现 (至少两次) 且在句中做主语; 如果主要人物的人名仅在文中出现一次, 一般是次要人物来衬托主要人物的. 有些新闻报道为了突出主要人物, 常常多次提及次要人物, 本文根据主谓关系和非主谓关系来区分主要人物和次要人物, 达到提取事件关联性人物要素的目的.

本文称主谓关系的人物为主语 (简称主), 非主谓关系的人物称为宾语 (简称宾). 相同人名不同关系所占的比例计算大概分为四个方面: 有主无宾、无主有宾、有主有宾和无主无宾. 通过上面的指代消解, 再根据不同方面的主宾人名比率以及分别所占个数的范围设置不同的权重. 结合以上特点, 分析出此新闻中主谓关系的人名数目和非主谓关系的人名数目, 并对同时含有主谓和非主谓关系的人名进行消解, 根据新闻人物报道的特点, 使非主谓关系权重的 0.4 倍相叠加到主谓关系人名的权重. 根据相同人名不同关系所占的比例设置一定的权值 W, W 的计算方法如下:

(1) 使用 Sigmoid 函数把输入值 (主谓关系类型的不同人名个数)“压缩”到 0~1 之间, 输出的值是相对应于人名的权重. 公式如下:

$$\varphi_i = \frac{x_i}{person_{\max}} \times interval_\theta \tag{2}$$

$$f(\varphi_i) = \frac{1}{1+e^{-\varphi_i}} \tag{3}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

$person_{max}$ 表示此新闻人名权重的最大值, $interval_{\theta}$ 表示使用 Sigmoid 函数的区间长度, x_i 表示统计的主谓关系的人名个数; 通过式 (3) 对重要性不同程度的人名赋予权值, 根据权值抽取刻画新闻事件的人名.

(2) 如果不考虑实体本身的权值, 直接对 (1) 所得人名根据权值抽取新闻事件人名要素; 否则, (1) 所得与其对应的实体人名的权值 (权值获取的方法与主谓关系人名计算权值一样) 相融合, 然后抽取与事件关联密切的人名.

考虑到两个人名的个数都很大, 经过 Sigmoid 函数输出的值基本上接近于 1 且两者之间的差异性不明显, 然而又基于实体的权值有可能会造成偏差. 为了避免丢失新闻事件的主要人物信息, 并放大主要人物和次要人物的差距, 所以本文把统计的主谓关系的人名个数 x_i 归一化到 0~6 区间的 φ_i 值作为 Sigmoid 函数的输入值.

4 实验结果及分析

实验 1 提出基于条件随机场方法来识别新闻文本中的实体, 本文以人名实体识别为例对训练模型进行分析验证. 以人民日报数据和扩展新闻数据的语料库作为训练语料训练模型, 选取预处理后的 86 655 篇新闻作为测试数据. 基于训练完成的模型进行实验, 多次随机选取 100 条新闻实验结果进行分析.

实验结果表明, 只使用人民日报新闻作为训练数据训练出的模型对应的 F 值为 63%, 而添加当今新闻数据后的语料库训练出的模型, 对新闻数据进行测试, 准确率明显提高了 22%, 其原因是现今的新闻文本和 1998 年的人民日报语料存在一些新意的专有名词和语境环境的偏差, 扩展语料库进一步提高了模型的预测能力.

实验 2 提取刻画新闻事件的要素, 以提取事件人名为例对方法进行验证. 首先, 在实验 1 识别出人名实体的基础上, 通过依存句法算法分析人名实体在新闻事件中扮演的角色, 根据实体之间的依存关系, 通过提取刻画事件要素模型对新闻事件要素设置不同的权重, 根据权重进行排序, 并提取出新闻事件关联性较为重要的人名. 实验把测试数据分成社会、时政、财经、娱乐与体育五大类别. No weight 表示基于实体不带权值的基础上提取事件人名要素的准确率, Weight 是基于人名实体 (有权值) 的基础上提取人名要素的准确率. 实验 2 结果如表 1 所示.

通过对实验结果和新闻人物报道的研究与分析, 最后, 通过提取刻画新闻事件人名要素的模型抽取前三项要素作为与事件密切相关的人物, 经过多次实验结果表明提出的方法能够有效地提取事件要素.

表 1 基于实体的事件要素提取

类别	文本数	No weight(%)	Weight(%)
财经	18 065	75	76
社会	23 504	65	68
时政	73	77	80
体育	26 986	76	83
娱乐	18 027	85	89
全部	86 655	75.6	79.2

从表 1 可以看出, 基于实体识别和依存句法算法两者产生的新思路 (建立一个提取刻画事件要素的模型) 比传统提取事件要素的算法更能体现新闻事件的主题, 更符合用户的需求, 且算法的性能较优; 在带有权值实体的基础上提取新闻要素的准确率有明显的提升, 主要是因为本文除了考虑实体之间的关系外, 还考虑了事件要素与新闻事件关联性; 测试数据分为社会、时政、财经、娱乐与体育五大类别, 关于社会生活新闻的要素识别准确率明显低于娱乐、时政新闻, 其主要原因是娱乐与时政新闻刻画事件人物比较明显、深刻.

5 结束语

本文提出基于混合模型的新闻事件要素提取方法, 该方法借鉴命名实体识别方法的构建思想, 提取出新闻事件中关键要素 (专有名词), 进一步提取匹配概括新闻事件最为接近的要素, 取得了一个较好的实现效果. 面向新闻事件要素的分析研究迫切需要解决的问题就是新闻文本的要素语料的收集和标注问题. 随着半监督和监督学习方法不断引入该领域, 使用未标注语料集的方法将逐步解决语料库不足的问题, 也为新闻数据挖掘方面的研究提供了较好的基础. 提取新闻要素之间的关系类型比较耗时, 如何提高模型的性能并保证提取要素的效果, 是我们下一步需要研究的工作之一. 我们下一步的探讨工作将围绕新闻事件发生的时间、地点、内容以及对事件人物的情感色彩[18]展开研究, 用这些要素来表达整个新闻的核心思想.

参考文献

1 裴东辉. 中文新闻事件抽取方法研究[硕士学位论文]. 昆

© 中国科学院软件研究所 http://www.c-s-a.org.cn

明: 昆明理工大学, 2015.
2 朱青, 李贞昊. 基于要素关联树的新闻发生地抽取技术研究. 网络新媒体技术, 2015, 4(3): 28–36, 59. [doi: 10.3969/j.issn.2095-347X.2015.03.005]
3 涂子令, 周枫, 余正涛, 等. 基于超图的汉越双语新闻话题要素提取. 计算机应用研究, 2017, 34(8): 2278–2281. [doi: 10.3969/j.issn.1001-3695.2017.08.008]
4 鞠久朋, 张伟伟, 宁建军, 等. CRF 与规则相结合的地理空间命名实体识别. 计算机工程, 2011, 37(7): 210–212, 215. [doi: 10.3969/j.issn.1000-3428.2011.07.071]
5 孙镇, 王惠临. 命名实体识别研究进展综述. 现代图书情报技术, 2010, (6): 42–47.
6 徐靖, 李军辉, 朱巧明, 等. 基于短语和依存句法结构的中文语义角色标注. 计算机工程, 2011, 37(24): 169–172. [doi: 10.3969/j.issn.1000-3428.2011.24.057]
7 石翠. 依存句法分析研究综述. 智能计算机与应用, 2013, 3(6): 47–49. [doi: 10.3969/j.issn.2095-2163.2013.06.013]
8 龙光宇, 徐云. CRF 与词典相结合的疾病命名实体识别. 微型机与应用, 2017, 36(21): 51–53.
9 徐元子, 张迎新, 刘登第. 基于条件随机场的网络评论与事件中命名实体匹配研究. 计算机应用研究, 2016, 33(6): 1642–1647. [doi: 10.3969/j.issn.1001-3695.2016.06.010]
10 Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA. 2002. 282–289.
11 隋臣. 基于深度学习的中文命名实体识别研究[硕士学位论文]. 杭州: 浙江大学, 2017.
12 王雍凯, 毛存礼, 余正涛, 等. 基于图的新闻事件主题句抽取方法. 南京理工大学学报, 2016, 40(4): 438–443.
13 邱莎, 段玻, 申浩如, 等. 基于条件随机场的中文人名识别研究. 昆明学院学报, 2011, 33(6): 64–66.
14 Ghani R, Probst K, Liu Y, *et al*. Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter, 2006, 8(1): 41–48. [doi: 10.1145/1147234]
15 Qiu LK, Hu CJ, Zhao K. A method for automatic POS guessing of Chinese unknown words. Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK. 2008. 705–712
16 Chen DQ, Manning CD. A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP. Doha, Qatar. 2014. 740–750.
17 郭江. 依存句法分析的置信度研究[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2012.
18 庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法. 计算机工程, 2012, 38(13): 156–158, 162.

© 中国科学院软件研究所 http://www.c-s-a.org.cn