Web 新闻语料分词和标注错误分析

张永奎[1,2],张 彦[1,2],安增波[3],刘 睿[1,2]
ZHANG Yong-kui[1,2],ZHANG Yan[1,2],AN Zeng-bo[3],LIU Rui[1,2]

1.山西大学 计算机与信息技术学院,太原 030006
2.计算智能与中文信息处理省部共建教育部重点实验室,太原 030006
3.中国人民解放军 91708 部队 自动化工作站,广州 510320
1.Department of Computer & Information Technology,Shanxi University,Taiyuan 030006,China
2.Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing,Taiyuan 030006,China
3.Workstation Automation of 91708 PLA,Guangzhou 510320,China
E-mail:zjm1203@sxu.edu.cn

ZHANG Yong-kui,ZHANG Yan,AN Zeng-bo,et al.Analysis of inaccurate style in processing Web true news text——about word segmentation and part of speech tagging.Computer Engineering and Applications,2007,43(15):166-169.

Abstract: Eleven inaccurate styles are obtained through analyzing the processing of Web accidental news text,we propose resolvent for some styles.This not only illuminates the improvement of word segmentation and part of speech tagging methods in early process of corpora,but also provides references to automatic check,another branch of Chinese information processing.

Key words: Chinese information processing;word segmentation;part of speech tagging;inaccurate style;Web accidental news corpora

摘 要:通过分析 Web 突发事件语料库文本的加工统计得出 11 类错误类型,并对其中的一些错误提出了解决方案。研究结果不仅对语料库加工初期分词、标注方法的改进有启发作用,而且对中文的自动校对方法,提供一定的借鉴。

关键词:中文信息处理;分词;词性标注;错误类型;Web 突发事件新闻语料库
文章编号:1002-8331(2007)15-0166-04 **文献标识码**:A **中图分类号**:TP391

1 引言

随着语料库语言学研究的兴起,建设高质量的大规模语料库已成为首要任务。语料库的加工包括分词和词性标注。对语料库标注得越准确,语料库的价值就越高。近年来国内外对词性标注的研究有很多,大多是采用基于规则和基于统计的方法。对错误标注结果进行分析可以看出,无论哪种标注算法都有其固有缺陷:概率标注方法总会抑制小概率事件的发生,而规则方法本质上说是一种确定性的演绎推理方法,因此它们很难对词性标注的准确率进行进一步的提高。显然,这样的准确率仍然严重影响语料库的加工质量[1]。词性标注是进一步自然语言处理的基础,在许多应用领域,如文本索引、文本分类、语言合成、语料库加工,词性标注都是一个重要环节。汉语自动分词已经被研究了二十多年,但是目前仍然是制约汉语信息处理发展的一个"瓶颈"。汉语的分词还没有形成一个公认的分词标准[2-4]。目前国内已有多个分词系统在应用,它们的正确率一般能在 95%以上,但只有限制在一定的范围内,这些系统才能取得相对较好的效果,并且这些都是在对已登录进词典的词进行处理的结果。

为了进一步探讨分词和标注的错误原因及其解决方法,我们选用了 Web 突发事件新闻语料进行实验研究。

2 突发事件新闻语料的加工实验

实验所用数据集合全部是来自互联网的新闻网页,包括 2000 年至 2004 年的 1 633 篇突发事件新闻(大约 700 万字),并将其分为海啸、道路交通、飞机失事、爆炸、沉船、禽流感、SARS 等 7 类。作者使用 ICTCLAS 分词软件[5]对新闻文本进行分词处理和词性标注,所选用的是二级标注操作,北京大学的标准输出[6]。

在实验过程中,发现真实页面本身存在一些明显的错误,以及使用分词软件后带来的一些错误。这些错误直接影响我们统计某些参数的准确率,有必要在此对语料库文本加工过程中常见的错误类型进行分析,这样对以后分词软件的改进可以提供一定的借鉴和依据。

3 突发事件新闻语料加工过程中的错误类型及其分析

3.1 录入时引起的错误

这里指使用录入设备(键盘等)录入文稿时所产生的一些

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60475022);山西省自然科学基金(the Natural Science Foundation of Shanxi Province of China under Grant No.20041041);山西省回国留学人员基金(No.2002004)。

作者简介:张永奎(1945-),男,教授,博士生导师,主要研究方向:中文信息处理与人工智能;张彦(1981-),女,在读硕士,主要研究方向:中文信息处理与人工智能;安增波(1979-),男,助理工程师,主要研究方向:人工智能。

错误,主要包括以下6类。

(1)误输现象。主要是由于文稿在输入过程中粗心大意造成的,也与录入者所使用的汉字输入法有一定的关系。比如,拼音输入时可能会造成相同发音的错字。利用部首字型输入法(五笔字型输入法等)时可能会造成形相似的错字。

例 1 第一/m　笔/q　25万/m　元/q　赔付/v　金/tg　已/d　于/p　5月/t　10日/t　交/ng　到/v　了/u　遇难/v　副/b　架/v　驶/vg　员/ng　林/nr　友贵/nr　的/u　家属/n(录入)

应为:第一/m 笔/q　25万/m　元/q　赔付/v　金/tg　已/d　于/p　5月/t　10日/t　交/ng　到/v　了/u　遇难/v　副/b　驾驶员/n 林/nr　友贵/nr　的/u　家属/n

(2)漏输现象。主要是由于文稿在输入过程中的疏忽造成的,比如,录入者进行录入时跳过一个字或者多个字,或者发现输错字时,在删除过程中多删掉了字和词产生了漏字现象。

例 2 德国/ns　中部/f　地区/n　的/u　一/m　条/q　高/a　公路/n　2日/t　发生/v　严重/a　交通/n 事故/n　(录入)

应为:德国/ns 中部/f　地区/n　的/u　一/m　条/q　高速公路/n　2日/t　发生/v 严重/a 交通/n 事故/n

(3)换位错误(这种提法参照文献[7])。主要是由于文稿在输入过程中粗心造成的,比如,录入者使用汉字输入法录入文稿,遇到同音的词就易产生这类错误。

例 3 送/v　往/p　附近/f　医院/n　极/d　积/v　救治/v(录入)

应为:送/v 往/p　附近/f　医院/n　积极/ad 救治/v

(4)重复现象。主要是由于文稿在输入过程中粗心造成的,比如,在文稿录入过程中录入者进行了多余的敲击键盘、粘贴复制的动作产生了同一个字、词甚至段的重复。

例 4 法国/ns　发生/v　重大/a　交通/n　事故/n　至少/d　6/m　人/n　至少/d　6/m　人/n 丧生/v　20/m　人/n　身/ng　受/v　重伤/n　(录入)

应为:法国/ns　发生/v　重大/a　交通/n　事故/n　至少/d　6/m　人/n　丧生/v　20/m　人/n　身/ng　受/v　重伤/n

(5)整篇不分段现象。主要是由于文稿在输入过程中录入者没有给文稿进行分段,像直接记流水账目一样没有分开层次。还可能是由于录入者在录入过程中使用某种超文本标记语言时漏掉了分段标记造成的,比如,在使用HTML语言时,漏掉了分段标记
,就会造成不分段现象。

(6)其他错误。包括数字、标点或括号的不匹配现象,当然,这类错误对于我们前期语料库的建设不构成太大的影响。

3.2　某些地名不能识别

例 5 大/a　垸/ng　乡/n　派出所/n　所长/n　刘/nr　云跃/nr　等/u　当即/d　赶赴/v。

"大/a　垸/ng　乡/n"应为"大垸乡/ns"

例 6 准备/v　飞往/v　巴/j　拉那州/ns　的/u。

"巴/j　拉那州/ns"应为"巴拉那州/ns"

3.3　某些专用名词不能识别

例 7 前/f 苏联/ns 图波/nr 列/v 夫/rg 设计局/n。

"图波/nr 列/v 夫/rg 设计局/n"应为"图波列夫设计局/nz"

例 8 这/r　架/q　二十二/m　年/q　机龄/nr 的/u 菲/j　鹰/j 航空/n　公司/n　客机/n。

"菲/j　鹰/j　航空/n　公司/n"应为"菲鹰航空公司/nz"

3.4　时间写法不够严谨,数字与汉字混用

例 9 当地/s　时间/n　晚上/t 9/m　点零/m 8分/t　。

"9/m　点零/m　8分/t"应为"9点/t　08分/t"

3.5　外国的地名、人名混淆

例 10 飞机/n　是/v　从/p　位于/v　克罗地亚/ns　东南部/f　的/u　著名/a　旅游/vn　城市/n　杜布罗夫尼克/nr　飞来/v　的/u。

"杜布罗夫尼克/nr"应为"杜布罗夫尼克/ns"

例 11 这/r　架/q　双/m　引擎/n 飞机/n 从/p 凯恩斯/nr 起飞。

"凯恩斯/nr"应为"凯恩斯/ns"

3.6　应该分为人名的分成了语素

例 12 工作/vn　人员/n　张/nr　淑珍/nr　以及/c　报务员/n　李/n　来/f　清/a 的/b。

"李/n　来/f　清/a"应为"李/nr 来清/nr"

例 13 莫/d　晓/vg。

"莫/d　晓/vg"应为"莫/nr 晓/nr"

3.7　繁体字不能准确识别

例 14 因/p 爆炸/v 後/nr 现场/d 温度/n 极/d　高/a

"後/nr　"应为"後/p"

3.8　其他成分被划分为人名

例 15 远东/ns 西伯利亚/ns 滨海/n 城市/n 符拉迪沃斯托克/nr(/w 海参崴/ns)/w 方/nr 向飞/nr 去/v

"方/nr　向飞/nr　去/v"应为"方向/n 飞/dg 去/v"

3.9　一些简称无法识别

例 16 中/f　新/a　社/n　网站/n。

"中/f　新/a　社/n"应为"中/j　新/j　社/n"(这里为"中国新闻社"的简称)

例 17 以/p　安全/an　部队/n　的/u　汽车/n。

"以/p"应为"以/j"(这里的"以"是"以色列"的简称)

3.10　系统编译导致的错误

这是由于录入者当时所使用的计算机的浏览器没能正常的运行代码,编辑错误就产生了。比如,编辑过程中产生的乱码。

3.11　其他错误

例 18 广/a 东海/ns 丰/ag 发生/v 特大/b 交通/n 事故/n。

"广/a　东海/ns　丰/ag"应为"广东/ns　海丰/ns"

例 19 在/p 浙江/ns 瑞安市/ns 上空/s 因/p 飞/v 机/ng 爆炸/v 坠毁/v。

"飞/v　机/ng　"应为"飞机/n"

(由于篇幅所限,只列举了少数几个例子)

经过分析研究发现出错的地名、人名对新闻事件主题提取有直接影响。总的来说,第2、3、8、11类可以划为是分词软件分词错误导致的,第5、6、9类可以划为是分词软件标注错误导致的。其中第1类错误的直接原因是录入者在输入文稿过程中粗心大意造成的,所以解决此类错误的主要方法在于提高录入者的细心程度。第2、5是错误出现频率较高的,要解决这些错误必须及时更新地名库,使地名库包括到乡、村。特别是对于国外的地名库,但是由于条件所限,很难做到地名的穷尽性搜集。人名和地名的自动识别是未登录词识别的重点和关键,人名、地名识别问题的解决必然会提高词法分析、句法分析乃至整个中文信息处理的质量。第4类错误解决的方法可以是规范时间的书写格式,"XXXX年XX月XX日XX时XX分XX秒",其中,

表 1 统计结果(总文本数 1633)

错误类型		错误类型所占比例		各类错误总数/处		出错文本所占比例	出错文本数目/篇
录入时引起的错误	误输现象	15%	60%	571	953	30%	490
	漏输现象		20%	190			
	整篇不分段现象		10%	96			
	重复现象 换位错误 其他错误		10%	96			
应该分为人名的分成了语素		9%		572		27.5%	449
某些专用名词不能识别 外国的地名、人名混淆 一些简称无法识别 某些地名不能识别		72%		4 572		60.9%	996
繁体字不能准确识别 其他成分被划分为人名 时间写法不够严谨,数字与汉字混用 系统编译导致的错误 其他错误		4%		254		7.2%	118

“X”可以是0–9之间的数字,还可以是以“零一二三四五六七八九”任意组合的时间,但是“X”必须是纯数字或纯文字,不然就会产生错误。第7类错误出现的频率较小,对人们的工作影响不是特别大。但是也应该注意中文简体、繁体的使用对象和环境,根据环境和对象的不同选择合适的。第8类错误直接原因是名字中的姓和名单独也可以作其他成分,所以导致了错误的出现,只有继续提高分词软件的准确率。对第9类错误,只能加强上下文的语义的研究,使得那些简称可以被准确地标注。

4 错误类型的统计

作者以突发事件新闻语料作为统计错误类型的基础,在研究中发现了错误类型共11类。经人工统计,在这几类错误中,第1、2、3、5、6、9较为常见。某些地名不能识别、某些专用名词不能识别、外国的地名和人名混淆、一些简称无法识别通常要占72%左右,录入时引起错误的比例通常要占15%左右,应该分为人名的分成了语素这种错误仅占9%左右,而在录入时引起的错误中,误输现象占60%左右,漏输现象占20%左右,整篇不分段现象占10%左右。在这1 633篇文本中共出现错误为6 351处。各类出错文本所占比例分别为30%(第1类)、27.5%(第6类)、60.9%(第2、3、5、9类)、7.2%(第4、7、8、10、11类)。因为一个文本中可能会出现多种错误类型,当在实验文本总数确定的情况下,出错文本比例之和就不是100%,这也就是出错文本比例之和不是100%的原因(统计结果见表1)。计算所用的公式如下:

$$\text{错误类型所占比例}=\frac{\text{某一出错类的错误数}}{\text{所有错误的总数}}\times 100\% \quad (1)$$

$$\text{出错文本所占比例}=\frac{\text{出错文本的数目}}{\text{总文本数}}\times 100\% \quad (2)$$

5 简体和繁体混用错误的解决方案与实验

本文主要针对第4类和第7类错误提出一个解决方案。

对于第4类时间书写不规范错误,本文设计了一个程序,使得日期的格式保持一致性,得到了不错的效果(见图1)。不过在实验中发现:

(1)如果语料中还包括除时间之外的其他数字出现,在由数字转换成对应汉字的时候,这些数字也将转换成相对应的汉字(有一种例外情况就是:数字是全角输入时,则不会转换,这样也给了我们一个启示:当不想把语料中的某些数字信息转换成汉字时,就可以把它设为全角输入,见图2)。

图 1 时间写法的一致性处理

图 2 数字是全角输入时不会被转换

(2)如果语料中出现“一、二、三、四、五、六、七、八、九、零”这样的字,在由汉字转换成数字的时候,它们就对应转成了“1、2、3、4、5、6、7、8、9、0”(见图3)。鉴于在实验中发现的这两种情况,只要把语料中汉字及数字均转换成所需的简体中文或繁体中文即可。

对于第7类中文简体和繁体混用错误。这里同样设计了中文简体与繁体的互转(即GB2312<–>BIG5)程序,也达到非常好的效果。这个软件只能转文本格式(*.Txt)和html格式(*.htm)的文件,本文的语料也只是html格式和由html格式转换成的文本格式两种,所以符合要求。

思想是:分别建立简体字库和繁体字库,从所读取语料文件的第一个字开始在字库中找到该字的位置号,锁定该位置号,然后在另一个库中找到该位置号所对应的字进行转换。就可以把这篇语料转换成所要的简体或繁体(见图 4)。

图 3　汉字转成数字的现象

图 4　中文简体与繁体混合语料转换成简体语料

经实验得出:通过设计的程序进行一致性处理后,这样可以完全避免第 4 类和第 7 类错误的发生。

要解决第 2、5 这些错误必须及时更新地名库,使地名库包括到乡、村。特别是对于国外的地名库,但是由于条件所限,很难做到地名的穷尽性搜集。我们将针对这些错误进行进一步的探讨和实验,以解决更多的问题。

6　结束语

通过本文的研究,对语料的分词和标注工作提出以下建议:

(1)现有的分词软件对译名、专有名词、人名、地名的处理还不太完善,还有必要投入大量的人力、财力进行研究。汉语自动分词是进行中文信息处理的基础。困扰汉语自动分词的一个主要难题就是未登录词的识别[8]。未登录词分为专名和非专名两大类。专名包括人名、地名等,非专名包括新词、简称、方言词语、文言词语、行业用词、港台用词等。目前关于未登录词识别的研究,主要集中在专名上,非专名的未登录词识别问题尚未引起足够的重视。为了使分词和标注正确率得到进一步的提高,在今后的工作中,应对汉语未登录词进行深入研究,这些问题的解决必然会提高词法分析、句法分析乃至整个中文信息处理的质量。

(2)有些错误是录入者在输入文稿过程中粗心造成的,所以有必要提高录入者的细心程度,而且今后也应在自动侦错、纠错方面下功夫。

(3)应规范时间的书写格式,也应注意中文简体、繁体的使用对象和环境,避免错误的产生。(收稿日期:2006 年 12 月)

参考文献:

[1] 张虎,郑家恒,刘江.汉语语料库词性标注自动校对方法研究[J].计算机应用,2005,25(1):17-19,24.

[2] 赵曾贻,陈天娥,朱兰.一种基于语词的分词方法[J].苏州大学学报,2002,18(3):44-48.

[3] 黄昌宁,李涓子.语料库语言学[M].北京:商务印书馆,2002.

[4] 刘开瑛.中文文本自动分词和标注[M].北京:商务印书馆,2000.

[5] 张华平,刘群.汉语词法分析系统.北京:中国科学院计算技术研究所,2002.

[6] 俞士汶,段慧明,朱学锋,等.北京大学现代汉语语料库基本加工规范[J].中文信息学报,2002(5):49-64;2002(6):58-65.

[7] 张磊,周明,黄昌宁,等.中文文本自动校对[J].语言文字应用,2001(1):19-26.

[8] 陈小荷.自动分词中未登录词问题的一揽子解决方案[J].语言文字应用,1999,31(3):103-109.

(上接 165 页)

的爬行系统所表现的结果(如图 1(a)),本文提出的搜索系统在相关页面的获取速率上是呈递增趋势的(如图 1(b))。换句话说,在能正确选择爬行路线、减少下载和分析其他资源所需的时间等的情况下,本系统能正确引导主体对相关页面的探索。

5　结束语

通过对蚁群进食行为的仿生研究,提出一种反映式智能主体架构。通过一个标准测试集的实验评测显示,当该架构用于适应性 Web 搜索系统时,其呈现一定的优异性能。另外,本架构还具有对环境改变的鲁棒性和对用户信息需求变更的适应性等重要特性。这种帮助用户获取相关资源的自治搜索是基于通用相似性测量理论基础上的,因而广泛使用的 IR 技术能被简单实现,包括那些在发现和组织信息过程中为用户提供帮助的个性化搜索系统。除了上述研究成果,本文尚需进一步研究的内容包括:多智能主体理论与架构及其在海量、动态的超文本领域的具体应用。(收稿日期:2006 年 11 月)

参考文献:

[1] Bonabeau E,Dorigo M,Theraulaz G.Inspiration for optimization from social insect behavior[J].Nature,2000,406:39-42.

[2] Chakrabarti S,Van Den Berg M,Dom B.Focused crawling:a new approach to topic-specific Web resource discovery [J].Computer Networks(Amsterdam,Netherlands,1999,31(11/16):1623-1640.

[3] Menczer F,Belew R.Adaptive retrieval agents:internalizing local context and scaling up to the web[J].Machine Learning,2000,31(11/16):1653-1665.

[4] Mizuuchi Y,Tajima K.Finding context paths for web pages[C]//Proc of 10th ACM Conference on Hyper Text and Hypermedia,Darmstadt,Germany,1999:13-22.

[5] Yang C C,Yen J,Chen H.Intelligent Internet searching agent based on hybrid simulated annealing[J].Decision Support Systems,2000,28:269-277.

[6] 王小平,曹立明.遗传算法——理论、应用与软件实现[M].西安:西安交通大学出版社,2002.

[7] 蒋国瑞,孙明.基于 Lucene 的 TBT 文档管理 Agent 系统研究[J].情报杂志,2006,25(5):37-40.