警务应用中基于双向最大匹配法的中文分词算法实现

文/陶伟

摘要

中文分词是信息提取、信息检索、机器翻译、文本分类、自动文摘、自然语言理解等中文信息处理领域的基础。目前中文分词依然是中文信息处理的瓶颈之一，本文对常见中文分词算法进行研究，并针对警务应用的场景，在经典的Jieba中文分词算法的逆向匹配法基础上提出双向最大匹配法，最后验证改进后的算法在中文分词准确度方面的提升。

【关键词】中文分词 双向最大匹配法 警务应用

1 研究背景

公安机关日常工作中采集到的数据，大多为碎片化数据，具多源、量大、且又离散如何有效提取这些非结构化数据中的有效信息以方便警务应用系统进行进一步分析处理，为案件侦破、情报分析等提供服务，关键技术就是利用中文分词算法将这些描述性的中文语句转变为结构化数据。

2 中文分词技术简介

2.1 中文分词算法分类

中文分词技术属于自然语言处理技术范畴，现有分词算法分为基于规则的分词方法、基于统计的分词算法和基于理解的分词方法。

基于规则的分词方法中占主流地位的是正向最大匹配法和逆向最大匹配法。由于汉语单字成词的特点，正向最小匹配和逆向最小匹配一般很少使用。逆向匹配的切分精度一般高于正向匹配，遇到的歧义现象也比较少。由大数据量的统计表明正向和逆向最大匹配的错误率分别为1/169和1/245，但这种精度还远远不能满足实际的需要。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

基于统计的方法是基于多个汉字同时出现的概率，通过对语料库有监督或无监督的学习，得到描述一种语言的语言模型（常用一阶隐马尔可夫模型（1stHMM）），该方法优点是只要有足够的训练文本即可降低未登录词的影响。

2.2 Jieba分词算法

Jieba分词算法基于规则与统计相结合，利用前缀词典实现高效的词图扫描，生成句子中所有可能成词情况所构成的有向无环图(DAG)，要生成有向无环图必须有语料库的辅助，语料库中每条记录会包含词、词频、词性等属性。Jieba分词试图查找树结构中的最大概率路径，找出基于词频的最大切分组合；对于未登录词，采用基于汉字成词能力的HMM模型，使用Viterbi算法。

<< 上接 152 页

对于一些结构较为新颖的桥梁来说，开展桥梁工程质量检测工作，除了需要检测桥梁的安全性与耐久性以外，还需要对桥梁使用的科研价值以及理论意义进行判定。因为桥梁施工建设通常需要在相关理论指导下才能进行，进而通过准确的计算来得到相应的数据支撑。所以，检测人员需要对桥梁工程建成后的实际受力情况以及现实承载力进行理论分析。桥梁工程具体质量是否过关，需要根据国家的设计文件以及设计标准来判断。通常来讲，计算机桥梁检测系统需要对两个方面进行判定，首先是所采集到的图像信息，因为周围的拍摄环境跟图像的实际拍摄效果之间有重要联系，只有在良好的拍摄环境下得出的图像才具有较高的研究价值。其次是对裂缝的判断，在具体检测桥梁工程质量的过程中，裂缝的实际走向跟桥梁的现实承载力之间具有重要联系。在计算机桥梁工程检测系统中，计算机可以将裂缝在横、纵坐标上的投影值进行比较分析，进而能够准确得出相应的桥梁裂缝走向。

3.3 计算机桥梁检测系统的未来展望

计算机桥梁工程检测系统是在近年来逐渐发展起来的一项新型检测技术，它逐渐在各个建筑工程检测中得到了广泛应用。计算机检测系统具有效率高、速度快、精度准等特点，将计算机检测系统应用到桥梁工程质量检测工作当中去。一方面可以有效减轻检测工作人员的工作负担，使得检测工作人员的桥梁工程检测工作能够得到先进技术的辅助支撑，同时有效降低了桥梁质量检测时的危险性；在另一方面，可以将原本复杂的数据测量工作交由计算机检测系统来分析与处理，有效降低了在具体检测桥梁工程质量过程中的人力资源消耗，节省了大量的检测成本。计算机检测系统还具有无损无接触的特点，在具体开展桥梁工程质量检测工作时，不会对桥梁产生影响。因此，计算机检测系统在未来的应用范围将会逐步得到拓展。

4 结语

总而言之，随着现代科学技术的不断发展，计算机桥梁检测系统将越来越完善。计算机桥梁检测系统的应用，有效提高了桥梁检测结果的准确性与有效性。能够及时发现桥梁工程中存在的一些问题，同时采取有效措施来解决问题，进而保障人民群众的出行安全。

参考文献

[1] 彭玲丽，黄少旭，张申申，李乾．浅谈无人机在桥梁检测中的应用与发展[J]．交通科技，2015,06:42-44.

[2] 殷迅．PDA在桥梁检测中的应用[J]．北方交通，2013,07:81-84.

[3] 赵赛雷．论动静荷载试验在桥梁检测中的应用[J]．黑龙江交通科技，2014,03:158-160.

[4] 唐文昌．关于桥梁检测车在桥梁检测中应用的分析[J]．黑龙江交通科技，2014,07:103-104.

[5] 吴军．计算机在桥梁检测中的应用[J]．科技资讯，2013,32:28.

[6] 何文．安全管理在桥梁检测中的应用分析[J]．黑龙江交通科技，2015,08:143.

作者简介

向阳（1988-），男，湖北省咸宁市人。现为中铁大桥局武汉桥梁特种技术有限公司助理工程师。研究方向为计算机技术应用。

作者单位

中铁大桥局武汉桥梁特种技术有限公司　湖北省武汉市　430074

Jieba 分词在 HMM 中有两种状态，一种是具有决定性的隐含状态，另一种的显性输出的状态。在 Jieba 分词中状态有 4 种，分别是 B,M,E,S，对应于一个汉字在词语中的地位即是 B（开头）,M（中间）,E（结尾）,S（独立成词），而输出就是一个汉字。

在 HMM 中还有三种概率分别是状态分布概率，状态转移概率和发射概率（发射概率是一个条件概率，表示在某一状态下得到某一输出的概率）。图 1 为二状态的马尔科夫模型。

要对一串汉字进行分词，首先使用 Viterbi 算法判断这串汉字最有可能的 BMES 组合形式。在 Jieba 分词中作者经过大量的实验在文件中中预存好了汉语的一些概率值。prob_start.py 中预存了每种状态的概率。

对于一个中文语句，第一个汉字的状态概率称为初始概率，可以用贝叶斯公式得到：

$$P(i) \times P(\frac{k}{i}) = P(k) \times P(\frac{i}{k})$$

其中 P(i) 表示状态的概率，在文件 prob_start.py 中可以找到，

$P(\frac{k}{i})$ 即发射概率，而 P(k) 即某个汉字出现的概率，忽略不计。则有：

$$P\left(\frac{i}{k}\right) = P(i) \times P(\frac{k}{i})$$

第二个字的状态概率是：

$$P(i_2) = \frac{P(i_1) \times P(i_2|i_1) \times P(k_2)}{P(i_2)} = P(i_1) \times P(i_2|i_1) \times P(k_2|i_2)$$

其中 $P(i_1)$ 表示第一个字的状态概率，$P(i_2)$ 表示第二个字的状态概率，$P(i_2|i_1)$ 表示状态 i_1 到 i_2 的转移概率，$P(k_2|i_2)$ 表示发射概率。

以此类推，由于每一个状态都有 4 种选择，所以根据每种选择导致的状态转移路径计算得出的概率值也不同，Viterbi 算法的目的是找出概率最大的一种转移路径。如果语句长度为 2，那么算法的目的就是使上面的 $P(i_2)$ 最大化。到达某一种中间状态的路径可能有多条，如在第三个节点到达状态 M 可能路径有 S->B->M、B->M->M 等，Viterbi 算法在中间这一步中就进行“剪枝”，只记住路径中概率最大的那条。

3 警务应用系统中对Jieba分词算法的改进

警务应用系统对分词算法准确度的要求远远高于对分词算法速度的要求。本文基于 Python3.4+Jieba0.37 分词包，选择在中文词库、最大匹配算法等方面进行改进以提高分词准确度。

3.1 词库优化

词库是中文自动分词的基础，词库机制的优劣直接影响到分词的速度和准确度。

（1）枪械，盗窃，抢劫这些事警务应用中的常见词汇，通过合理调高这些词的比重，能够快速将其识别提取。

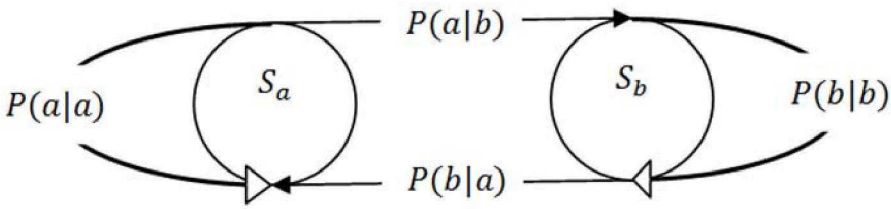

图 1：二状态马尔科夫模型

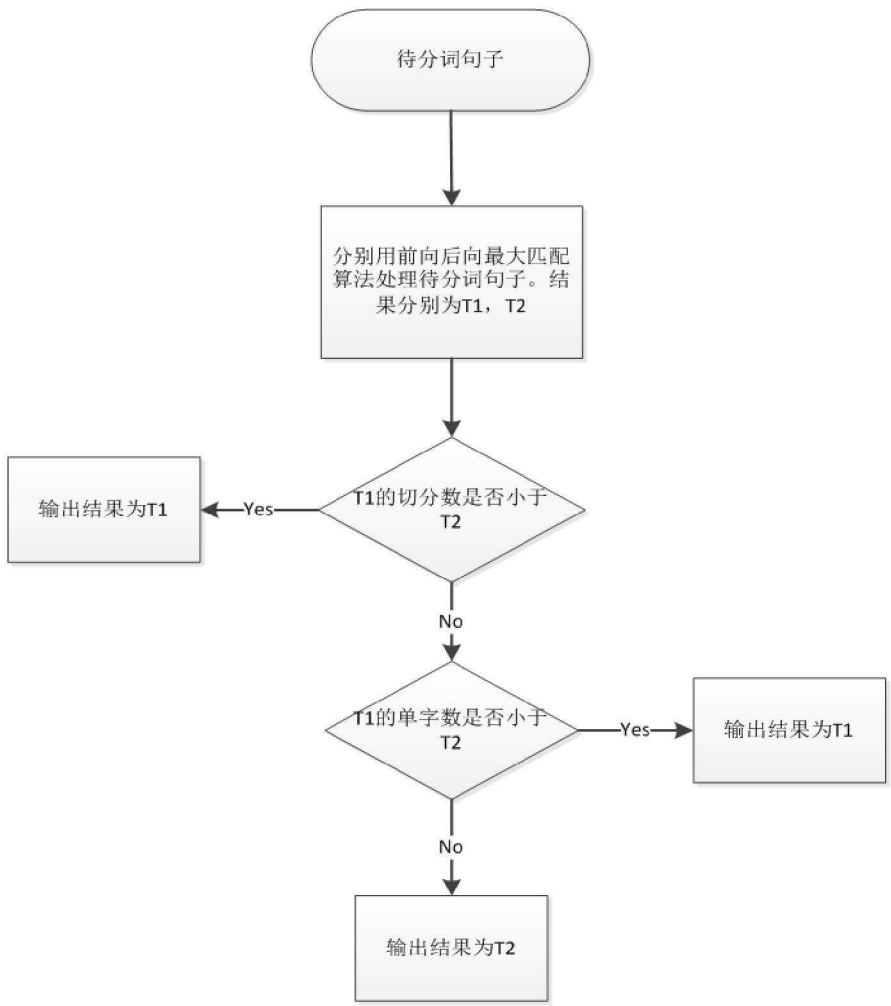

图 2

（2）利用二次分词法可以提高城市名、路名、车牌号、时间等警务应用中较为关心的词汇的识别度。在基本词库中添加当地地名和地址库，地名地址词库保存城市特有地名、路名、机构名、小区名、兴趣点和大地名。通用词主要用于对词典中有重复的词或是分词后的具有歧义或不准确的词进行修复，如“华夏二路”在词典并不存在，但是“华夏”却在词典中，“二”为数量词，所以自动分词结果为华夏 / 二 / 路。这时可对地址进行二次查找，当有单字存在于地名通名词典中时，将其与前面未登录单词合并。如“路”存在于通用地址库中，可将其和前面的单词“二”合并，分词结果为华夏 / 二路。同理，对于某些特殊格式的词，如车牌“沪 XXXXXX”、时间等，由于词库不能枚举所有车牌或日期，同样可采取二次提取，将特定格式的字符串整合。如“沪 XXXXXX”按常规分词会分为“沪 / XXXXXX”，利用二次提取，将所有省简称与后面跟着非汉字字符串长度为 7 的字符串合在一起成为新词，即可被识别为车牌信息。

3.2 双向最大匹配法

最大匹配算法主要原理是切分出单字串，然后和词库进行比对，如果确定是一个词就记录下来，否则通过增加或者减少一个单字，继续比较，直到剩下一个单字为止，如果该单字串无法切分，则作为未登录词处理。

据 SunM.S. 和 Benjamin K.T. 的研究表明，中文中 90.0% 左右的句子，正向最大匹配法

和逆向最大匹配法完全重合且正确，只有大概9.0%的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功），只有不到1.0%的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两者都错（歧义检测失败）。

本文提出的双向最大匹配法将正向和逆向最大匹配法得到的分词结果进行比较，决定正确的分词方向。首先对语句正反向各进行切分，然后根据大颗粒度词越多越好，单字词越少越好的原则，选择一种合适的方式输出。即在切分数不同时，优先取切分数少的那一个结果；切分数相同时，优先取单字少的那种；在切分数，单字数都相同时，因为逆向最大匹配法统计来说相对准确率较高，取逆向为结果。流程图如图2所示。

以下两例是双向最大匹配法的应用场景：

例一：“我们在野生动物园玩”（如表1）

表1

	正向：我们/在/野/生动/物/园/玩	逆向：我们/在/野生动物园/玩
切分数	6	4
单字数	2	2
选取结果	我们/在/野生动物园/玩	

例二：“我是北大学生”（如表2）

表2

	正向：我/是/北大/学生	逆向：我/是/北/大学生
切分数	4	4
单字数	2	3
选取结果	我/是/北大/学生	

可以看出在使用双向最大匹配算法可以在单纯使用正向与逆向匹配算法结果不同时，选择更合适的一个，保证了更高的分词准确度。

3.3 性能测试

根据国际中文分词测评的标准中对汉语分词进行测试的方法，在第二届国际汉语分词测评中共使用了四家单位提供的测试语料(Academia Sinica、City University、Peking University、Microsoft Research)，本文基于Peking University提供的语料，利用语料库自带的分词脚本进行测试。以下的测试结果显示双向最大匹配法比起Jieba算法的最大逆向匹配，虽然消耗更多计算时间和计算资源，却提升了分词准确度（如表3）。

表3

	Jieba分词	改进后的分词算法
准确率（P）	0.853	0.871
召回率（R）	0.787	0.797
综合指标（F）	0.818	0.832

召回率指在分词标答中，分出正确的词语所占的比例；准确率指分词结果中，分出正确的词语所占的比例。召回率和准确率一般是此消彼长的关系，所以常用11种召回率下准确率的平均值（综合指标）来衡量一个分词系统的精度。将这两个度量值融合成一个度量值，如F度量，如图3、图4所示。

```
=== SUMMARY:
=== TOTAL INSERTIONS:   1811
=== TOTAL DELETIONS:    9894
=== TOTAL SUBSTITUTIONS:        12386
=== TOTAL NCHANGE:      24091
=== TOTAL TRUE WORD COUNT:      104372
=== TOTAL TEST WORD COUNT:      96289
=== TOTAL TRUE WORDS RECALL:    0.787
=== TOTAL TEST WORDS PRECISION: 0.853
=== F MEASURE:  0.818
=== OOV Rate:   0.058
=== OOV Recall Rate:    0.583
=== IV Recall Rate:     0.799
###     pku_seg.txt     1811    9894    12386   24091   104372  96289   0.787
0.853   0.818   0.058   0.583   0.799
```

图3：Jieba0.37结果

```
=== SUMMARY:
=== TOTAL INSERTIONS:   1065
=== TOTAL DELETIONS:    9945
=== TOTAL SUBSTITUTIONS:        11240
=== TOTAL NCHANGE:      22250
=== TOTAL TRUE WORD COUNT:      104372
=== TOTAL TEST WORD COUNT:      95492
=== TOTAL TRUE WORDS RECALL:    0.797
=== TOTAL TEST WORDS PRECISION: 0.871
=== F MEASURE:  0.832
=== OOV Rate:   0.058
=== OOV Recall Rate:    0.740
=== IV Recall Rate:     0.801
###     pku_seg.txt     1065    9945    11240   22250   104372  95492   0.797
0.871   0.832   0.058   0.740   0.801
```

图4：改进后的结果

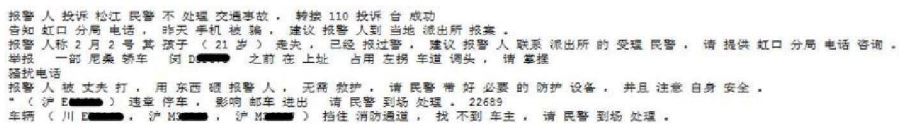

图5：Jieba0.37部分结果

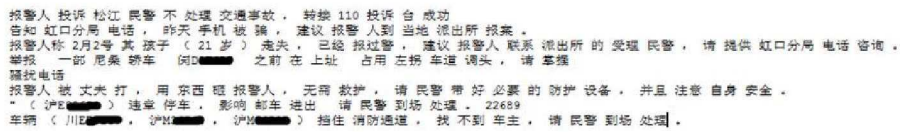

图6：改进后的部分结果

图3、图4所示。

对于警务应用系统，针对警用词汇的词库改进后，能以更准确的提取所需词汇。以接警信息作为对象进行测试，例如在接警信息中出现“报警人”、“虹口分局”、地址、日期等词汇时，在原程序下会变成“报警人”“虹口分局”无法正确识别。摘取部分实际测试结果如图5、图6。

4 小结

本文针对警用领域的分词进行了一定的研究，寻找在杂乱无序的碎片文件中以较高准确度提取关键词的方法。中文分词是识别、提取出关键词进行分析应用的基础，本文提出的词库优化和匹配算法改进能够提升原Jieba算法的准确度，由于语料库规模限制以及参数设置等因素的影响，实验结果还存在可提高的空间。

参考文献

[1] 冯书晓，徐 新，杨春梅．国内中文分词技术研究新进展[J]．情报杂志，2002(11)：29-30.

[2] 宗成庆．统计自然语言处理[M]．北京：清华大学出版社，2013(08).

[3]（美）Wesley J.Chun（陈仲才）著，杨涛，王建桥，杨晓云，高文雅，等译．python核心编程[M]．北京：机械工业出版社，2001(08).

[4] 结巴中文分词项目[EB/OL]．(2012-09-29)[2013-01-5]. https://github.com/fxsjy/jieba.

作者单位

上海市公安局静安分局 上海市 200040