规则与统计相结合的案件名称识别

乔春庚， 肖诗斌 ， 孙丽华 ， 施水才

（北京信息科技大学中文信息研究中心，北京 100101）

摘 要： 在公安领域信息中，案件名称有着举足轻重的作用。因此，如何准确的识别出文本中的案件名称是一个非常重要的研究课题。在对公安领域文本进行了深入地分析和研究的基础上，总结出了案件名称的结构特征及其上下文信息，建立了用于识别案件名称的知识库。在知识库的基础上，首先对案件名称进行模板识别，然后进行结构分析和上下文分析，并利用禁用词库对案件名称进行排歧，从而识别出候选案件名称。我们再使用统计方法对识别出的候选案件名称计算权值，过滤权值比较低的实体，这样能大大提高系统的准确率。初步实验结果表明，在封闭测试中犯罪案件名称抽取的精确率可以达到 95.26% ，召回率可达 89.14%；在开放测试中精确率可以达到 84.47% ，召回率可达 75.56%。

关键词： 案件名称；公安领域；信息抽取

Recognition of Case-Name Based on Rule and Static

Qiao Chungeng[1], Xiao Shibin[2], Sun Lihua[3], Shi Shuicai[4]

(Chinese Information Processing and Research Center, Beijing Information Science & Technology University ,Beijing 100101)

Abstract: Identifying case names in running texts plays a significant role in police information extraction. Based on the thoroughly investigations of police articles, the relevant structural features and contextual constraints were obtained. In this paper, a case name identification system was proposed, which is built on the knowledge bases. Based on the knowledge bases, first, we recognized the Case-name using templates, then we analyzed the structural features and contextual constraints, and we removed the branching using dictionary. Then we used static-method to calculate the value of the case name, we filtrated the lower entity, so the accuracy of the system was improved consumedly. The experiment achieves 95.26 % precision and 89.14 % recall respectively by close test, and 84.47 % precision and 75.56 % recall respectively by open test.

Keywords: Case-name; police domain; information extraction

1 引言

命名实体识别是自然语言处理中的一项基本工作，命名实体的识别也是句法分析、机器翻译、信息抽取等任务的一个非常重要的预处理模块。一般来说，命名实体识别[6]的任务就是对于一篇待处理文本，识别出其中出现的人名（Person）[4]、地名（Location）[3]、机构名（Organization）[1][5]、日期（data）、时间(time)、百分数(percentage)、

基金资助：国家自然科学基金项目(60272084)；北京市教育委员会科技发展计划重点项(KZ200310772013)；北京市教委项目(KM200510772008, KM200610772008)

作者简介：乔春庚（1981-），男，河北省霸州市，在读硕士 qcg1981@163.com .

货币(monetary value)等命名实体。但是对于一些特定的领域，又有其独特的实体，比如公安领域，案件名称就是实体。

随着公安领域信息急剧增长，如果让侦查人员费时费力去查卷宗，对于一般的案件就勉为其难，还会增加破案成本。在如此多的资源信息中如何获得破案线索信息，如何提高破案效率，如何快速有效的找到相关案件信息，已经成为刑侦工作迫切解决的问题。在案件侦破过程中，刑侦破案的重要手段就是把调查到的各种线索联系起来，查出犯罪分子作案规律，将几个相对独立的案件放到一起进行调查，就是串并案。这样，只要破一案，就能带一批案件，大大提高破案效率。其中如何正确的识别案件名称就成为了一个关键问题。目前国内外重要刊物上还没有关于案件名称识别研究工作的报导。

本文在从专名识别处理的角度进行研究工作的同时，结合了前人研究[2]上的可取之处，充分利用了公安领域的特征，专门针对犯罪案件名称的识别问题进行研究。在识别策略上综合考虑了案件名称的结构特征和文本上下文信息，建立了用于识别的知识库，并使用统计方法进行权值计算，很好地过滤掉了一些噪音，实验结果是令人满意的。

2 犯罪案件名称的特征

犯罪案件名称众多，规律各异，长短不一，所以案件名称的识别有一定的困难。

(1) 没有明确规范的案件名称定义，随着时代的发展，会有新的案件名称不断出现。

(2) 案件名称的用词比较自有、分散。案件名称中既可以有名词，也可以有动词。例如："贩卖|毒品|案"。

(3) 案件名称的长短不一，短的如"杀人案"，长的如"中国长城资产管理公司广州办事处诉中山大松通用机电有限公司及火炬集团(担保方)借款合同纠纷一案"。

(4) 案件名称有时同一些介词、动词、方位词之类的指示词出现，但有些指示词可以作为案件名称的组成。例如 "制造'3.25'杀人案"，其中"制造"为指示词；"非法制造贩卖毒品案"，其中"制造"为案件名称的一部分。

(5) 有的案件名称中包含标点符号，识别上产生困难。例如"杀人、抢劫案"，"平安证券有限责任公司（简称：平安证券）偷税案"。

(6) 案件名称结尾一般有案件名称特征词出现，例如"案"就是一个明显的特征词。

其中，1、2、3、4、5增加了案件名称识别的难度，并可能产生歧异；4、6有助于案件名称的识别。

3 规则与统计相结合的案件名称识别

规则与统计相结合的案件名称识别主要包括三个过程（见图 1）：文本预处理；基于规则的识别；基于统计的识别。文本预处理阶段就是对原始文本经过简单的分词、分句之后，得到初加工文本。基于规则的识别就是利用建立的知识库对案件名称进行识别。对识别出的案件名称，通过统计方法计算权值，过滤权值较低的案件名，最后得到识别结果。

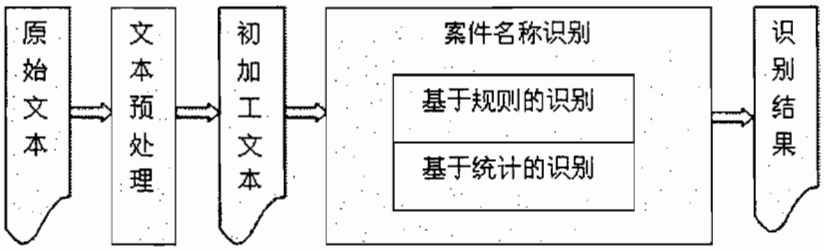

图 1 案件名称识别策略

Fig.1 the Strategy of the Recognition of Case-name

3.1 文本预处理

原始文本首先进入分词系统，该分词系统进行了改造，添加了部分公安领域词词典。部分公安领域词词典中收集了公安领域中出现频率比较高的词汇。经过分词，然后进行分句后，我们得到初加工文本。

3.2 基于规则的案件名称识别

◆ 知识库的建立

本文使用的语料来自互联网上，共 10，026 篇真实文本。其中案件名称 21，500 个。在这些资源的基础上，我们利用机器统计和人工筛选相结合，建立了如下知识库，用来指导案件名称的识别。

(1)关键词库

案件名称结尾有案件名称特征词出现。所以，在我们的系统中，对于案件名称的识别是从确定案件名称的右边界开始。以案字结尾的这些词可以提供准确的案件名称右边界信息，将这些信息收集整理，建立案件名称关键词库，作为案件名称识别的触发条件。

(2)前缀库

案件名称有时同一些介词、动词、方位词之类的指示词出现，例如“审理”、“判决”、“抓获”，“涉嫌”等。为了确定案件名的左边界，我们建立了案件名称前缀库。但是前面提到的特征 4 已经表明，有些词既可以是案件名称的前缀，也可以是案件名称的一部分。所以在确定案件名称前缀库时，一定要确保该词不可以作为案件名称的一部分，例如“制造”、“提供”、“参与”、“组织”等词都不能作为案件名称前缀。

(3)关键词前词禁用词库

我们经过实验发现，在案件名称中，有一些词不能出现在案件名称关键词前面，我们收集了这些词，建立了案件名称关键词前词禁用词库。如果关键词前词在禁用词库中，表明这不是案件名称，不用继续识别，从而可以大大减轻系统的工作量，也可以提高系统准确率。

(4)禁用词库

有一些词不能作为案件名称的组成部分，主要是一些介词，例如“随着”、“因而”、“此外”、“其余”等等，我们建立案件名称禁用词库。在识别过程中，碰到库中的词则中断案件名称的识别，认为当前识别对象不是案件名称成分。

◆ 基于知识库的识别

在识别系统的核心部分“案件名称识别”中，我们首先借助于知识库的进行规则识别（如图 2 所示）。

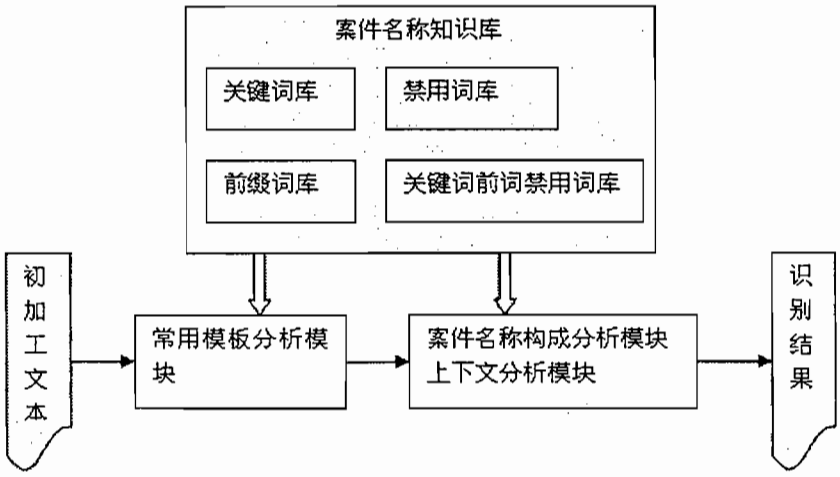

图 2 基于知识库的案件名称识别

Fig.1 Recognition of Case-name Based on Knowledge-bases

对于那些左边界比较清晰的案件名称，我们经过案件名称构成分析以及上下文分析，同时利用禁用词库对识别出的案件名称进行排歧，将最后结果保存到识别结果中。识别方法如下：首先，我们根据关键词库确定案件名称的右边界；然后从案件名称右边界开始向左搜索，判断关键词的前词是否在前词禁用词库中，如果在的话，就停止向左扫描，这不是案件名称；继续向左扫描，直到该词在前缀词库中或者到一句话的开头，此时识别出候选的案件名称；再从左向右判断案件名称的组成词是否在禁用词库中，如果在的话，就删去该词，重复执行，如果不在的话，就停止。这样我们就利用知识库识别出了候选案件名称。

在识别过程，有些词既可以作为案件名称的左边界，又可以作为案件名称的一部分，但是当这些词与某些词搭配出现时，那它只能作为案件名称的左边界，例如“与[案件名称]有关”，此时“与”就能确定为案件名称的左边界。所以，我们首先对案件名称进行常用模板分析，凡是符合模板的，就识别为案件名称。我们利用机器学习和人工选择相结合的方法，定义了 35 套模板，模板形式为：[案件名前词][案件名称][与前词搭配的尾词]。例如：“和[案件名称]有牵连”，这样“ [和].....[有牵连]”，就定义为一套模板。

3.3 基于统计的案件名称识别

在利用规则对案件名称进行识别后，我们发现一些噪音，组成这些噪音的词大都是些比较常见的词，我们利用统计方法计算识别出的候选案件名称的权重，从而过滤掉那些权值比较低的案件名称。我们的训练语料来自人民日报，共 26，987 篇文本。

在信息检索中最常用的确定一个词在文档中重要性的方法是TF*IDF 的方法。TF 即该词在一篇文档中出现的频率，IDF 称为反文档频率，一个词在越多的文档中出现它的IDF 就越小，反之就越大。公式如下（1）：

$$W(f_i,d)=TF(f_i,d)*IDF(f_i)=N(f_{id})*\log(N(f_i)/N) \qquad (1)$$

其中，$W(f_i,d)$ 是特征 f_i 在文本 d 中的权重，$N(f_i)$ 是出现 f_i 的训练文本数，N 是总训练文本数，$N(f_{id})$ 是文本 d 中出现 f_i 的次数。

TF*IDF方法识别原理：(1) 文档集中包含某一词条的文档越多，说明它区分文档类别属性的能力越低，其权值越小；反之，其权值越大。在犯罪案件名中出现的词，一般都是犯罪信息领域中的词，而我们的训练语料来自人民日报，所以包含词条的文档较少，其IDF值较大。(2) 另一方面，某一文档中某一词条出现的频率越高，说明它区分文档内容属性的能力越强，其权值越大。在犯罪信息文本中，犯罪案件名称中出现的词条一般都是文本中比较重要的词，其出现的次数也比较多，所以其TF值较大。根据上面两个原因，组成案件名称的词汇的TF*IDF值比普通词汇的大很多。我们利用该方法计算案件名称的权重，然后过滤权重比较低的噪音，实验表明，该方法能大大提高识别的准确率。

TF*IDF方法识别：对于根据规则识别出的案件名称，我们计算每一个词的TF*IDF值，然后取平均值，再和设定的阈值进行比较，如果大于阈值，那就是案件名称，如果小于阈值，就过滤掉。根据统计，我们取阈值value =3. 1。

4 实验结果和分析

本文使用的语料库由10，026篇犯罪信息文本构成，我们从中随机选出1000篇犯罪信息文本进行封闭测试。实验结果如下表1。

表 1 封闭测试的实验结果

Tab.1 The result of experiment by close test

	测试文本	案件名称个数	识别出的个数	正确	错误	准确率%	召回率%	F_1
进行统计识别前	1000	2118	2105	1897	208	90.11	89.57	89.83

进行统计识别后	1000	2118	1982	1888	94	95.26	89.14	92.10

注：上表中F_1 是β=1的F测试

F测试的公式为：
$$F=(\beta^2+1)RP/\beta^2R+P \quad (2)$$

R是召回率(Recall)，P是准确率(Precision)。

同时，我们还对300篇文本进行了开放性测试，实验结果如下表2：

表2 开放测试的实验结果

Tab.2 The result of experiment by open test

	测试文本	案件名称个数	识别出的个数	正确	错误	准确率%	召回率%	F_1
进行统计识别前	300	540	531	415	106	78.15	76.85	77.49
进行统计识别后	300	540	483	408	75	84.47	75.56	79.76

我们对识别结果中的错误进行了整理分析，发现错误主要有以下几种类型：

(1) 标点符号导致的识别错误

在识别的开始，我们对文本进行了分句处理，所以当案件名称中包含标志一句话结束的标点符号时，就导致了识别错误。例如："由广州市人民检察院提起公诉的郑洪钧等36名被告人走私'红油'（香港地区专用的添加红色染色剂的免税柴油，俗称红油）案"，在这段话中，案件名称识别为"俗称红油）案"，这就是由于逗号导致的错误。

(2) 分词错误导致的识别错误

在文本预处理阶段，对文本进行分词，由于分词错误，也可能导致识别错误。例如："南充市中级人民法院近日分别对方吉和罗小林两起涉黑犯罪大案做出一审判处"，在这句话中分词为"分别|对方|吉|和|罗|小|林|两起|涉|黑|犯罪|大|案"，由于分成了"对方"，所以导致了识别错误。

(3) 有些案件名称没有明显左边界，导致不能正确识别

如"我国加入世贸后知识产权案大幅增加"、"公交分局立扒窃案160起"、"使这个银行连续12年无经济罪案发生"，这些句话中，没有明显的左边界导致识别错误。

5 结论

在公安领域的信息抽取问题中，如何正确识别文本中出现的犯罪案件名称是一个非常重要的问题。本文在对公安领域中犯罪案件名称的结构及其在文本中的出现的上下文进行了深入研究的基础上，建立了四个识别用的知识库，并且使用TF*IDF方法对识别出的案件名称进行过滤。经过初步实验，结果表明我们的识别策略是很有效的。

参考文献：

[1] 张小衡，王玲玲. 中文机构名称的识别与分析. 中文信息学报，1997，11(4)：21-32.

[2] 王宁，葛瑞芳，苑春法，等. 中文金融新闻中公司名的识别. 中文信息学报，2002，16（2）：1-6.

[3] 黄德跟，岳广玲，杨元生. 基于统计的中文地名识别. 中文信息学报，2003，17（2）：36-41.

[4] 张华平，刘群. 基于角色标注的中国人名自动识别研究. 中文信息学报，2004，27（1）：85-91.

[5] Keh-Jiann Chen，Chao-jan Chen. Knowledge Extraction for Identification of Chinese Organization Names. In Proceedings of ACL workshop on Chinese Language Processing，2000，15-21.

[6] Hobbs J. The Generic Information Extraction System. In Proceedings of the Fifth Message Understanding Conference (MUC-5)，1993，87-91