



审判案例自动抽取与标注模型研究

余贵清 张永安

(北京工业大学经济与管理学院 北京 100124)

【摘要】针对刑事判决书文本,结合刑事审判本体,构建基于本体的案例自动抽取与标注模型。基于法律案例文本的半结构化特征,依据文档组织结构和线索词,运用正则表达式构建抽取规则模板;同时结合自然语言处理技术进行相关语义信息的精准抽取。运用语义标注技术构建刑事审判本体实例库,实现大量案例文本向语义信息网络的转化,便于运用语义信息进行相似案例检索和审判推荐。实验证明,该模型的抽取结果基本达到预期效果。

【关键词】语义标注 本体 规则抽取 自然语言处理

【分类号】D926.22 TP399

Study on the Model of Automatic Extraction and Annotation of Trail Cases

She Guiqing Zhang Yongan

(School of Economics and Management, Beijing University of Technology, Beijing 100124, China)

【Abstract】This paper constructs an Ontology - based automatic extraction and annotation model for the massive texts of criminal judgments combined with the case - Ontology. It uses regular expressions to construct extraction rules and templates for the semi - structured characteristics of the texts of legal cases, according to the structure of the documents and the clue words. Besides, it applies natural language processing techniques for the accurate information extraction, then gives semantic annotation of the results of extraction for building an Ontology knowledge base of legal cases, to realize the transformation of case texts to semantic information Web, for the further similar case retrieval and judge recommendation. And the experiment shows a good result.

【Keywords】Semantic annotation Ontology Rule extraction Natural language processing

1 引言

信息量的剧增使得对信息处理和运用的重要性在很多领域凸显出来。在司法领域,法律案例已经构成了相当大规模的案例集,这些案例集包含了不同的法律机构及个人对不同案例的评判标准和结果,蕴含着丰富的行业知识和专业智慧。对案例相关信息进行提炼,为法律机构提供相似案例的审判结果,进行判决推荐,可以更好地实现历史案例的价值。而如何对大规模的案例信息进行自动或半自动的抽取和标注,是形成判决推荐的基础,是构建案例本体实例库时重要的步骤。

本研究拟在构建一种基于本体的信息抽取与标注模型,结合刑事判决书本体,针对刑事判决书案例的文本特征进行实体抽取、属性识别和关系抽取,并将抽取结果经过语义标注,构造案例信息本体库,以供进一步检索推荐。

2 信息抽取研究现状

信息抽取是构建语义网^[1,2]的基础,是指从文本中抽取指定的事件、事实等信息,并形成结构化存储的过程。Cardie^[3]描述的信息抽取通用过程包括符号化和标注、句法分析、抽取、合并与模板生成5个部分。其中,语义标注与信息抽取是密不可分的两个关联过程,是在一个领域本体的指导下为文档添加规范化知识表示的过程,通常包括类型标注和关系抽取两个部分。

目前已有的语义标注工作主要包括三种类型：

(1) 基于机器学习和关系元数据的传统方法,主要是对标注好的训练集进行训练,但在实际应用中需要定义匹配规则等来实现知识的转化;

(2) 基于本体的方法, 利用本体中已有的实例来简化抽取过程中对概念实例的识别;

(3)自然语言处理的方法,借助语法结构和句法分析实现实体关系的抽取。

KIM^[4]中的语义标注模块,利用 GATE 实现自然语言处理的一系列集成,如分词、词性标注、命名实体识别、规则匹配和指代识别等。荆涛等^[5]通过构造基

于正则表达式规则的通用类型标注器和基于词汇构建的本体类型标注器进行标注,并通过分析句中词汇间的依存关系构造语法关系三元组。高琦^[6]提出一种基于弱监督的 Bootstrapping 和规则的本体标注模型,通过不断循环本体解析、文本分类和信息标注抽取的过程来实现对本体内容的丰富。Pandit^[7]形成了基于解析树和依赖树的信息抽取方法,通过句法结构对实体进行识别,再与本体模型进行匹配,实现内容扩充。

本文基于对多种语义标注和抽取算法的研究,针对法律案例文本的特点,综合运用本体、正则表达式、自然语言处理技术等多种方法构建信息抽取与标注模型,实现对法律案例信息的语义化表述。

3 基于本体的案例自动抽取与标注模型设计

3.1 研究基础

(1) “刑事审判本体”简介

本研究是基于本体的案例信息抽取,故以“刑事审判本体”中实体、属性和关系的内容要求为基础进行抽取内容和抽取策略的制定。“刑事审判本体”使用 Protégé 软件^[8]构建,该本体的类框架与语义关系如图 1 所示:

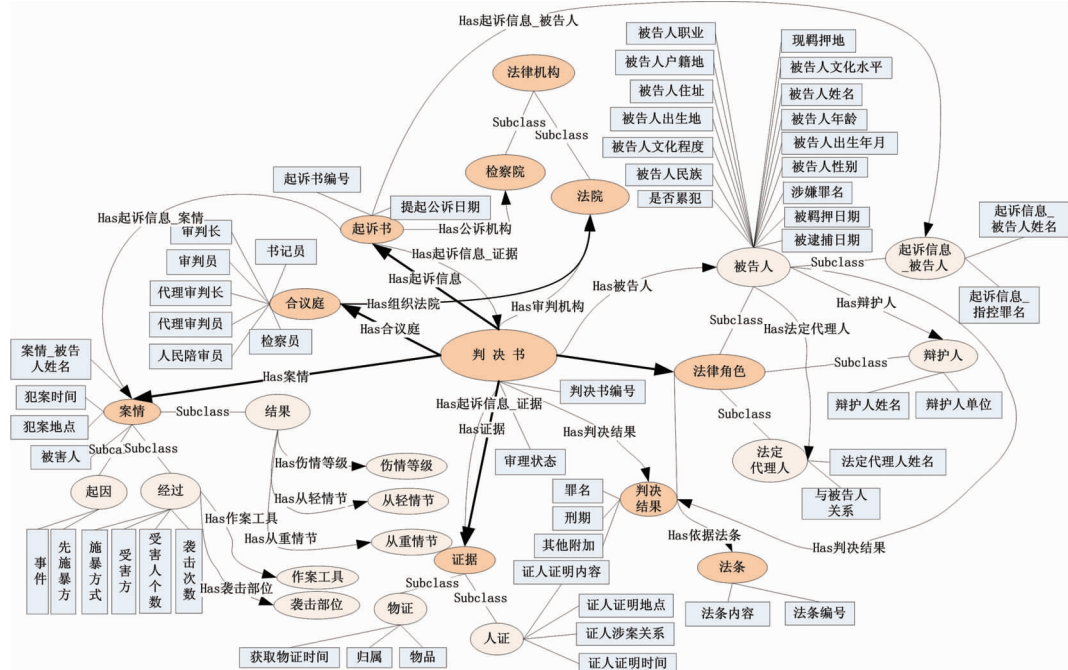


图1 刑事审判本体的类框架与语义关系

“刑事审判本体”中包含实体、实体属性及实体关系等内容。其中“判决书”为顶层本体,包含“起诉

书”、“法律角色”、“案情”、“证据”、“判决结果”等实体子类。在抽取过程中的篇章分析部分,结合刑事判

判决书案例文本的组织结构,将案例文本拆分为与上述实体子类相对应的内容,并以判决书或起诉书编号或角色姓名作为实例名进行标记。

同时,在本体结构中,各实体类中包含子类或不同的数据属性信息。如“法律角色”类中包含“被告人”、“辩护人”等实体类,且各实体都有自身不同的属性设计。在句级抽取和词级抽取的部分对实体下的各子类内容进行抽取标注,同时对应本体设计中的属性设置,从各部分的内容中识别属性信息,并添加相应的标记。

此外,本体结构中还包含各类实体之间的关系,即对对象属性的描述。如在“被告人”与“辩护人”之间有“Has 辩护人”的对象属性。在抽取过程中的关系抽取部分,对案例中的各种实体关系进行对应,并添加关系标记。

本文结合“刑事判决书”文档的结构特征和书写规范,对建立在“刑事审判本体”基础上的抽取方法和抽取模型等内容进行详述。

(2) 案例文本结构分析及抽取策略制定

通过阅读大量的案例文本,可以发现:

①不同法院的判决书写法虽存在一定差异,但内容的组织形式基本相同,重要的信息内容基本相同,组织流程大致都为判决书基本信息、法律角色信息、起诉书信息、案情信息、证据信息、判决信息几个大部分。

②各部分包含相对规范的线索词,可以在构造抽取规则时作为信息提取点。

③对于案情、证据及判决等部分,由于存在陈述方式的多样性,需要结合自然语言处理技术对句法结构进行深入分析,同时在实体识别的基础上进行相关信息的提取。

3.2 案例自动抽取与标注模型设计

从整体来看,本案例自动抽取标注模型分为三个层次:输入层、本体标注层及输出层,其整体架构如图 2 所示。

以案例文档集和已建好的“刑事审判本体”为输入,在本体标注层通过运用自然语言处理中的语法分析,结合模式识别方法,运用正则表达式,识别命名实体。如案例中人物角色、时间等信息,抽取识别出的实体进行类型标注;结合正则表达式构造规则抽取器,结合实体实例,进行实例属性值的抽取,并在实例抽取后进行实例间关系的抽取;最终将案例组织为 XML 格式,将案例文档集表示为标记的 XML 文档集,形成案例本体库。

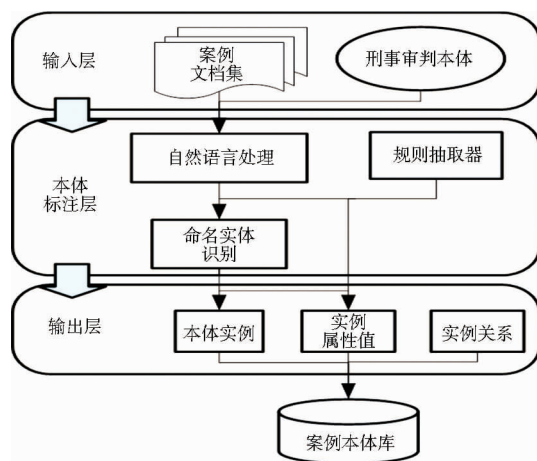


图 2 案例自动抽取标注模型框架

3.3 标注抽取过程设计

从整个信息抽取与标注过程纵向来看,主要分为 4 个模块:自然语言处理模块、模式识别模块、信息抽取模块及格式转化模块。各模块相互配合进行横向的信息内容抽取,主要包括命名实体、实体属性及实体关系三类信息的抽取与标注。需要结合文本挖掘工具进行文本分析、正则表达式进行规则构建及 C#编程进行抽取和格式转化。详细流程如图 3 所示:

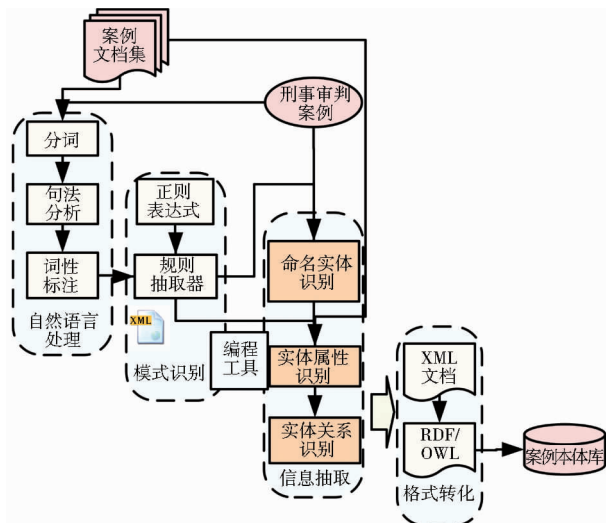


图 3 信息抽取与标注过程

(1) 命名实体的识别

运用 ICTCLAS^[9] 软件结合本体构造分词词典,对案例文档集进行分词,然后进行句法分析和词性标注,运用正则表达式构建 XML 文档形式的规则抽取器,结合规则抽取器和上下文信息进行命名实体的识别。这

里的命名实体主要包含被告人、辩护人等案件角色信息。因为这些信息会在文档中多次提及,但有些情况下未给出角色标注,所以需要先对这些信息进行抽取,以便在实体属性抽取时进行识别。

(2) 对实体属性的识别

主要是运用正则表达式构造的规则抽取器,按照案例文档结构对文档内容进行分块,对应本体形成6个信息块,即判决书基本信息、法律角色、起诉书、案情、证据和判决。然后结合之前抽取出的本体实例对各信息块中的详细属性进行抽取。如“北京市人民检察院第二分院以京检二刑诉[2007]183号起诉书指控被告人XXX犯故意伤害罪”中,可以依据模式“(? <= 以) [\u4E00 - \u9FFF] + \[\d {4} \] \ (\d {3, 6} , ?) + 号”抽取出起诉书编号。但有些信息涉及到语法结构及上下文意义,需要在正则表达式识别后,在词性标注的基础上拆分出具体属性相对应的内容,如对于时间格式的信息,在分词后需要对其上下文进行分析,来确定时间信息为“犯案时间”、“被逮捕时间”或“起诉时间”,并添加相应标注。

(3) 对实体关系的抽取

即根据在结构块中的共现规则,判断实体之间的对应关系或从属关系等。如判决书与起诉书的包含关系、合议庭与法院的从属关系、被告人与辩护人的对应关系等。

最终对抽取结果进行汇总,对相应的内容添加XML标记,对XML结构进行整理,构建标注库。再将标注库中的信息转换成RDF或OWL的形式,便于向本体库中导入,形成最终的案例本体实例库。

3.4 抽取算法设计

从纵向来看,信息抽取主要针对刑事判决书基本的写作规范和写作格式,结合自然语言处理的通用流程进行抽取算法的设计,主要分为篇章分析、句级抽取、词级抽取和标注4个部分,如图4所示。

(1) 篇章分析:依据刑事判决书的组成及内容分布,结合正则表达式将刑事判决书拆分为判决书基本信息、法律角色等6个部分并进行标注。

(2) 句级抽取:对包含多项相似模式内容的部分,识别起始模式对其进行内容子块的划分。

(3) 词级抽取:参照3.3节内容实现对命名实体、实体属性及关系等详细信息的抽取。

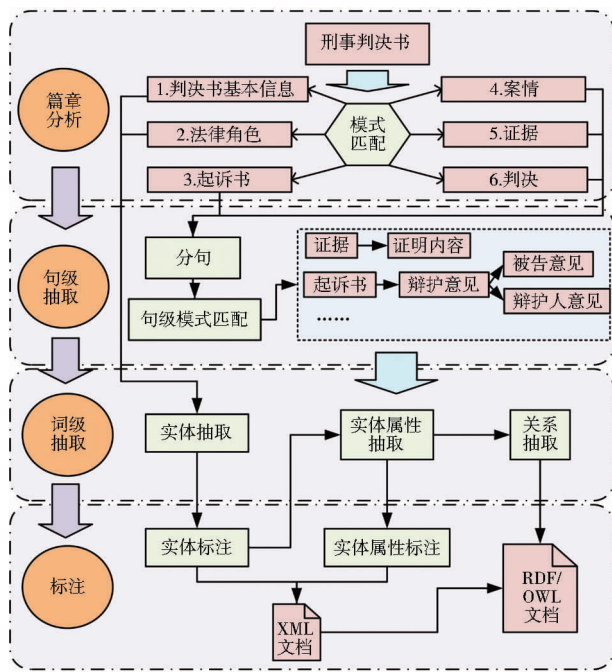


图4 抽取算法

(4) 对抽取结果进行标注,实现本体的实例化。

该算法中各信息项的处理基本都采用辨识、抽取、标注的流程,针对审判书文本自身的特点,设计正则表达式抽取规则,并对由中文表述的复杂性而造成的问题,结合自然语言处理技术进行文本预处理,如采用分词等技术对如判决书中的各类角色姓名的命名实体进行分词和词性标注等处理。今后还将结合自然语言处理技术对抽取模型进行优化。

仅以“判决书”中“法律角色”部分为例,进行算法展示,“……”省略了其余5个部分的信息抽取过程,其抽取标注方法与“法律角色”类基本相同,如下所示:

For 文档集中的每个“刑事判决书”文档 |

if 匹配“法律角色”的模式 |

抽取“法律角色”子类

if 匹配“被告人”的模式 |

抽取“法律角色”下的“被告人”子类,抽取被告人姓名,并标记为“@被告人姓名”

抽取性别、年龄、出生年月、出生地、民族、住址、职业、户籍地、涉嫌罪名、被羁押日期、被逮捕日期、现羁押地的信息,并标记 |

if 匹配“辩护人”子类 |

抽取“法律角色”下的“辩护人”子类,抽取辩护人姓名,并标记为“@辩护人姓名”

抽取辩护人单位等信息,并标记 |

```
if 匹配“法定代理人”子类 {
    抽取“法律角色”下的“法定代理人”子类,抽取法定代理人
    姓名,并标记为“@法定代理人姓名”
    抽取法定代理人年龄、出生日期、民族、职业、住址、与被告
    关系等信息,并标记 |
}
.....
标注关系:被告人——辩护人
标注关系:被告人——代理人
}
```

3.5 抽取规则设计

主要是对需要抽取的内容项进行抽取模式的总结,包括实体抽取和关系抽取。

(1) 实体抽取

针对案例本体的组成,对判决书文本进行实体抽取,主要采用正则表达式构建模式的方法,如法律角色部分的抽取模式,如表 1 所示:

表 1 抽取模式——被告人

XML 标识	类名	模式
Role_defendent_name	被告人姓名	(? <= 被告人) [\u4E00 - \u9FFF] { 2, 4 } (? = \ (,)
Role_defendent_sex	被告人性别	(? <= ,) [男女] (? = ,)
Role_defendent_age	被告人年龄	(? <= ,) \d { 1, 2 } 岁 (? = ,)
Role_defendent_birthdate	被告人出生年月	(? <= ,) \d { 4 } 年 \d { 1, 2 } 月 \d { 1, 2 } 日 (? = 出生于)
Role_defendent_birthplace	被告人出生地	(? <= 出生于) [\u4E00 - \u9FFF] + (? = ,)
Role_defendent_nation	被告人民族	(? <= ,) [\u4E00 - \u9FFF] + 族 (? = ,)
Role_defendent_education	被告人文化水平	(? <= ,) [\u4E00 - \u9FFF] + 文化 (? = ,)
Role_defendent_career	被告人职业	(? <= 文化 ,) [\u4E00 - \u9FFF] + (? = ,)
Role_defendent_address	被告人住址	(? <= , 住 暂住) . * ? (? = , 。)
Role_defendent_residence	被告人户籍地	(? <= , 户籍所在地 户籍地为) . * ? (? = , 。 ()
Role_defendent_crime	涉嫌罪名	(? <= 因涉嫌) [\u4E00 - \u9FFF] + (? = 于 \d { 4 } 年 \d { 1, 2 } 月 \d { 1, 2 } 日被)
Role_defendent_detained	被羁押日期	(\d { 4 } 同) 年 \d { 1, 2 } 月 \d { 1, 2 } 日 (? = 被羁押)
Role_defendent_arrested	被逮捕日期	(? <= 于) \d { 4 } 年 \d { 1, 2 } 月 \d { 1, 2 } 日 (? = 被逮捕)
Role_defendent_detainplace	现羁押地	(? <= 现羁押 (于 在)) [\u4E00 - \u9FFF] + (? = , 。)
Judge_punish_lighter	从轻情节	自首 积极赔偿 投案 积极救治 从犯 如实供述 揭发 认罪悔罪 未成年 (从判决依据中抽取)
Judge_punish_heavier	从重情节	逃逸 累犯 (从判决依据中抽取)

对于每一份刑事判决书,都包含一个或多个被告

人、辩护人等信息,需要先匹配其出现模式,再对这些人物的基本信息进行标注。

(2) 关系抽取

按照上述实体抽取规则,能够识别实体及类的包含关系,但在转化为 OWL 文件前,需要对实体之间的关系进行抽取。这里所指的关系,一方面包含普通关系,如判决书对起诉书、被告人、案情、证据、判决等信息的显性包含关系;另一方面还有复杂关系,如法院 - 合议庭的包含关系等这种隐含的机构之间的对应关系。当存在多个被告人时,被告人与辩护人的对应关系也需抽取,同理还有被告人与代理人的对应关系,抽取的源与目标模式如表 2 所示:

表 2 抽取模式——关系

XML 标识	关系名	源类	目标类
Relation_JudgeProcue	Has 起诉书	刑事判决书	起诉书
Relation_JudgeDefendent	Has 被告人	刑事判决书	被告人
Relation_JudgeCase	Has 案情	刑事判决书	案情
Relation_JudgeProof	Has 证据	刑事判决书	证据
Relation_JudgeJudge	Has 判决	刑事判决书	判决
Relation_Defend	Has 辩护人	被告人	辩护人
Relation_Agent	Has 代理人	被告人	代理人
Relation_CourtCouncil	Has 合议庭	法院	合议庭

对于关系的抽取,如包含关系等,在以常规模板进行信息抽取之后,结合案例审判本体,在构造语义库时,针对不同信息项之间的意义关联,通过定义关系属性并设置相应的一对一或一对多的对应关系来实现。

4 实验及结果分析

按照本文所述的抽取方法和抽取规则,笔者结合自然语言处理和正则表达式模式构建规则抽取器,以 C# 为编程语言,实现对 XML 模板的构建。采用“北京市第二中级人民法院”2006 年 - 2012 年“故意伤害罪”相关的 1 000 个刑事判决书的文本文档为数据集(文档长度大约在 5KB - 15KB 之间)进行信息抽取算法的实现和 XML 模板的构建。由于篇幅所限,仅以某审判案例中“法律角色”部分的自然语言文本为抽取案例,并展示抽取标注后对应的 XML 文档。

审判案例部分文档的内容截取如下所示^①:

.....

被告人张 XX,男,46 岁,1962 年 5 月 16 日出生于山西省孟县,汉族,小学文化,北京市 XXX 公司职工,住北京市朝

① 案例范例的斜体字部分对应 XML 抽取出的内容。

阳区慈云寺商场院内平房(户籍所在地:北京市朝阳区关东店X巷XX号);因涉嫌犯故意伤害罪于2008年3月17日被羁押,同年3月28日被逮捕;现羁押在北京市第一看守所。

指定辩护人王建京,北京市京工律师事务所律师。

.....

抽取标注后对应的XML文档部分,如下所示:

```
<?xml version="1.0" encoding="UTF-8"?>
<Judgement>
<Entities_JudgementCName="刑事判决书">判决书_(2008)
 二中刑初字第XXXXX号
  ....
  <Roles CName="法律角色">
    <Role_defendentCName="被告人">张XX
    <Role_defendent_nameCName="被告人姓名">张XX
    </Role_defendent_name>
    <Role_defendent_sexCName="被告人性别">男
    </Role_defendent_sex>
    <Role_defendent_ageCName="被告人年龄">46岁
    </Role_defendent_age>
    <Role_defendent_birthdateCName="被告人出生年月">
      1962年5月16日
    </Role_defendent_birthdate>
    <Role_defendent_birthplaceCName="被告人出生地">
      山西省盂县
    </Role_defendent_birthplace>
    <Role_defendent_nationCName="被告人民族">汉族
    </Role_defendent_nation>
    <Role_defendent_educationCName="被告人文化水平">
      小学文化
    </Role_defendent_education>
    <Role_defendent_careerCName="被告人职业">北京市
      XXX公司职工
    </Role_defendent_career>
    <Role_defendent_addressCName="被告人住址">北京
      市朝阳区慈云寺商场院内平房
    </Role_defendent_address>
    <Role_defendent_residenceCName="被告人户籍地">
      北京市朝阳区关东店X巷XX号
    </Role_defendent_residence>
    <Role_defendent_crimeCName="涉嫌罪名">故意伤害罪
    </Role_defendent_crime>
    <Role_defendent_detainedCName="被羁押日期">2008
      年3月17日
    </Role_defendent_detained>
    <Role_defendent_arrestedCName="被逮捕日期">同年
      3月28日
```

```
</Role_defendent_arrested>
<Role_defendent_detainplaceCName="现羁押地">北京市
  第一看守所
</Role_defendent_detainplace>
</Role_defendent>
<Role_defenderCName="辩护人">王建京
  <Role_defender_nameCName="辩护人姓名">王建京
  </Role_defender_name>
  <Role_defender_unitCName="辩护人单位">北京市京
    工律师事务所律师
  </Role_defender_unit>
</Role_defender>
.....
</Entities_Judgement>
<Relations_Judgement>
  <Relation_defendRName="Has 辩护人">
    <Source CName="被告人">张XX</Source>
    <Domain CName="辩护人">王建京</Domain>
  </Reltion_Defend>
  ....
</Relations_Judgement>
</Judgement>
```

对于内容抽取后的XML文档,需要依据XML对RDF文档的转化规则,用编程的方式实现对RDF文档的转化,最终将RDF文档导入本体库中,构建本体实例。

在实验中,笔者对1 000个审判案例分10组进行实验,采用召回率和准确率两个指标作为模型的评价标准,实验结果如表3所示:

表3 实验结果

组别	召回率	准确率
1	65.2%	52.3%
2	68.1%	57.8%
3	71.2%	63.5%
4	78.4%	69.4%
5	84.9%	72.6%
6	87.8%	73.9%
7	86.3%	78.1%
8	89.2%	80.2%
9	88.7%	82.6%
10	89.4%	81.4%

在实验中,依据抽取结果中出现的对抽取模板进行逐步修正,最终使抽取结果的召回率和准确率都达到80%以上。

经过实验,该模型的抽取结果基本达到预期效果,证明了该算法的有效性。但模型对信息的抽取仍不能达到完全准确,在后续研究中,会在模型中引入自然语

言处理中的句法分析,达到提高模型信息抽取精度的目的。

5 结 语

本文通过对案例信息进行抽取提炼,捕捉案例中有价值的语义信息,变案例集为计算机可以理解的语义网络,即本体实例库。笔者构建了基于本体的审判案例自动抽取标注模型,并进行了算法的设计与实现以及对基于正则表达式规则抽取器的构建工作。由于刑事判决书既有一定的书写模式结构,又兼具语言表述的多样性,所以在进行信息抽取时,对于规则结构比较清晰的内容,结合正则表达式构建抽取规则模板,实现了基于线索词的信息抽取方法。对于规则结构并不是很清晰的部分,结合自然语言处理技术,基于分词、词性标注和语义分析,来识别信息抽取重点,实现灵活语义的抽取。而如何精准地从表达灵活的信息中抽取所需信息,也将是未来研究的重点,同时还需对 XML 模板和正则表达式规则进行进一步细化与完善。

参考文献:

- [1] Uschold M, Gruninger M. Ontologies and Semantics for Seamless Connectivity[J]. *ACM SIGMOD Record*, 2004, 33(4): 58 - 64.
- [2] Berners - Lee T, Hendler J, Lassila O. The Semantic Web[J]. *Scientific American Magazine*, 2001, 284(5): 28 - 37.
- [3] Cardie C. Empirical Methods in Information Extraction[J]. *AI Magazine*, 1997, 18(4): 65 - 78.
- [4] Popov B, Kiryakov A, Kirilov A, et al. KIM - Semantic Annotation Platform[C]. In: *Proceedings of the 2nd International Semantic Web Conference (ISWC' 2003)*, Florida, USA. 2003: 834 - 849.
- [5] 荆涛, 左万利, 孙吉贵, 等. 中文网页语义标注: 由句子到 RDF 表示[J]. *计算机研究与发展*, 2008, 45(7): 1221 - 1231. (Jing Tao, Zuo Wanli, Sun Jigui, et al. Semantic Annotation of Chinese Web Pages: From Sentences to RDF Representations[J]. *Computer Research and Development*, 2008, 45(7): 1221 - 1231.)
- [6] 高琦. 基于 Bootstrapping 的本体标注方法研究[D]. 重庆: 重庆大学, 2010. (Gao Qi. A New Annotate Ontology Method Based on Bootstrapping[D]. Chongqing: Chongqing University, 2010.)
- [7] Pandit S. Ontology - guided Extraction of Structured Information from Unstructured Text: Identifying and Capturing Complex Relationships[D]. Ames: Iowa State University, 2010.
- [8] 章勇, 吕俊白. 基于 Protégé 的本体建模研究综述[J]. *福建电脑*, 2011, 27(1): 43 - 45. (Zhang Yong, Lv Junbai. The Research Review of Ontology Modeling Based on Protégé[J]. *Fujian Computer*, 2011, 27(1): 43 - 45.)
- [9] 刘克强. 2009 共享版 ICTCLAS 的分析与使用[J]. *科教文汇*, 2009(22): 271. (Liu Keqiang. The Analysis and Instructions for the 2009 Shared Version Of ICTCLAS[J]. *Education Science & Culture Magazine*, 2009(22): 271.)

(作者 E-mail: sheguiqing@263.net)