**面向公共法律服务的自然语言理解技术及语料库构建**

**研究背景：**

**法律语言**是表达了法律上有拘束力的符号系统。从语言学的角度看，法律语言是“领域语言”，而且可以说是最有现实意义的领域语言。法律语料库是用于法律活动和法律语言学研究的语言信息库。法律语料库是为了法律语言学研究和法律实践收集的，由书面语和口头语构成，经过计算机处理和存储的、大量真实的法律文本库集合。法律语料库的建设有着非常明确的目的。建立完备的法律术语语料库，可以规范法律术语，可以用于研究某个法律术语在法律文本中的出现频率，相关搭配及其出现时的语境，并根据统计结果确定法律术语的使用方法及适用范围。

**研究内容**：

公共法律服务是由司法行政机关统筹提供的，旨在保障公民基本权利，维护人民群众合法权益，实现社会公平正义和保障人民安居乐业所必需的法律服务。公共法律服务包括宣传和援助两个主题，涉及的法律面广、对象丰富。为规范公共法律服务内容和降低公共法律服务成本，有必要针对公共法律服务建立领域词典、设计有关法律法规和具体案件卷宗的自然语言处理工具，并完成对应的语料库和本体库的构建。

针对公共法律服务的具体特点，我们研究的内容包括初始语料信息的采集和预处理、中文文本语料信息的分词和词性标注、语料语义分析、根据语料语义筛选出正确信息、构建语料库和搜索引擎、基于语料库构建本体库六个方面的研究内容。

**研究方法：**

1. 基于自然语言处理技术的语料采集和预处理

由于信息化程度的提高，近年来的法律法规以及法律有关卷宗大多有电子存档，同时历史卷宗的数字化工作也在稳步推进，因此法律相关的语料素材是可以搜集的，需要注意的是要保证语料收集的准确、全面和均衡。根据搜集素材的多少、素材的存在形式（文字、图片、音频、视频等）确定素材的存储方式，针对当下语料素材格式丰富、体量大的特点拟采用分布式大数据平台（如阿里云）进行存储。

存储的大量数据是构建语料库的原始数据，这些数据格式不一，对后面的进一步处理会造成困扰，因此需要进行一些数据的预处理工作。针对多媒体数据，可以依据其标签，或者借助图像和音频的识别分类技术将其抽象为特殊代表的文字进行存储。针对不同格式的电子文档，编写相应的处理程序，将文字部分转成相同的格式。随后，将这些处理之后的文字数据以UTF-8编码存储于传统的关系型数据库。

2、基于分词和词性标注的语料库语料标注及分类

在语料素材采集和预处理的基础上，拟运用先进的自然语言处理工具对以文字形式存在的语料进行分词、词性标注、以及词频分析，添加法律特有的词性标注，主要需要解决的问题是如何提高分词的准确度。词性标注与分词通常是一起处理，处理的方式目前主要分为两类：

（1）管道式串行处理。在分词结束后，对已分好的词继续进行标注；

（2）联合模型处理。在分词的同时，对词性进行标注。

现有的分词方法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。近年来基于统计的分词方法在分词领域取得了很好的效果，比较常用的模型包括隐马尔可夫模型（Hidden Markov Model ，HMM），最大熵模型（ME），条件随机场模型（Conditional Random Fields，CRF）等。词性标注常用的方法分为马尔科夫方法和隐马尔科夫方法两类，具体方法有maxFreq、greedyTran和HMM等，其中实验显示隐马尔科夫类模型在词性标注方面有较好的效果，针对法律相关文本需要根据具体情况选择适用的方法。并在此基础上进行词性分类的分析与研究。

3、基于语料规范语义的法律领域词典构建

在完成词性标注之后，需要进一步理解由词组成的句子的含义，也就是需要进行语义分析。法律术语必须遵照其真实性原则，对于由专家定义证实的法律术语可以直接纳入信息库，而那些语义上模棱两可的信息将不能纳入术语库。尤其，从网络上采集的法律术语语料有时存在乱码、标点符号混乱(全角半角不统一，英文标点与中文标点混杂)、语料重复甚至错误等问题。因此需要更行之有效的手段来保证法律术语符合规范，模糊有歧义的句子在术语库中不能出现，以此构建法律领域词典。

领域词典的构建需要研究从语料库中自动抽取术语的算法。目前术语抽取中常用算法有C-Value与NC-Value算法、基于频率、假设检验、思然比和相对频率比、互信息、共现分布等等。实现法律领域词典的构建，需要对法律领域词汇的特征进行描述，如词频、出现的位置、上下文信息等，选择合适的算法并进行优化。

4、**语料库文本存储与搜索：**

经历上述阶段，搜集的多种格式的语料素材已经被处理成了结构化的、有明确语义的语料。这些具有结构化特征的数据，将以知识图谱等形式存储，并依据合适的搜索策略编写对应的搜索程序，完成语料库的构建。

（1）存储方案的研究。目前针对语料库，没有一个完整的存储方案。需要针对法律语料库，选择合适的存储介质，适当的组织结构对语料库的原始数据进行存储。同时，如何有效存储分词以及词性标注信息、词频信息、不同的索引信息，也需要进行针对性研究；

（2）搜索策略研究。针对用户需对特定词语进行全文搜索，或进行相似性搜索，或需要对特定词性的词语进行搜索。针对不同的需求，需要研究不同的搜索规则。

5、**法律本体库构建**：

法律本体库来源于语料库，是对语料库的进一步抽象，在语料库建成之后，采用自底向上或者自顶向下的方法，以发现领域内的概念、概念的继承层次、潜在的关系和公理等。法律本体库主要包括以下7 个部分，并可根据实际情况进行扩充：

1)Legal Action：法规中的处理行为及适用条款

2)Legal Explain：法规中的各种解释

3)Legal Punishment：法规中的处罚标准

4)Legal Righ：法律中规定的权利

5)Legal Obligation：法律中规定的义务及职责

6)Legal Proces：法律事务执行时的各种流程

7)Legal Prohibition：法律中禁止的事项

而本体库的构建，主要分为本体分析、本体表示、本体评估、本体建立四个步骤。构建方法有从局部到全体、统计分词、特征提取、FCA构建文档本体等方法，在此基础上进行本体之间的关联分析。