



CHANGSHA UNIVERSITY OF SCIENCE & TECHNOLOGY

**毕业设计（论文）**

**题目：** **法律问题关键词抽取系统**

**学生姓名： 刘嘉挺**

**学 号： 201450080334**

**班 级: 计算机卓越班15-1**

**专 业： 计算机科学与技术**

**指导教师： 曾道建**

**2019 年 4 月 18日**

法律问题关键词抽取系统

**学生姓名： 刘嘉挺**

**学 号： 201450080334**

**班 级： 计算机卓越班15-1**

**所在院(系): 计算机与通信工程学院**

**指导教师： 曾道建**

**完成日期: 2019年4月18日**

法律问题关键词抽取系统

摘要

法律问题关键词抽取是当前关键词抽取技术在我国法律范畴内的应用，法律问题与人们日常生活以及社会规范制度息息相关。目前我国的法律文件及条款在我国人口众多、社会情况复杂等诸多原因下不断完善且数据量庞大，且伴随着高速网络信息和大数据时代的到来，绝大部分的人们由于专业限制通过使用各种在线法律问答系统来咨询自身所需的法律咨询或援助。如何从法律问答库的海量数据中抽取出人们需要的、准确的数据已经是当前研究的重点。关键词抽取是法律信息挖掘最基本的任务，准确的关键词能够快速引导人们关于法律问题的参考方向，能够告知人们某个法律问题所属的具体法律条款范畴。由于当前关键词抽取技术大多在长文本中基于统计学的方法，在法律问答（短文本）的关键词抽取效果不好。

因此本文主要通过构建一个基于循环神经网络(RNN)的序列到序列(seq2seq)模型从某一简短的法律问题中生成关键词。首先从法律问答网站上利用爬虫技术收集大量法律问题以及已经人为对应标记好的关键词，并对（问题-关键词）语料库做预处理。对语料库进行中文分词之后，利用word2vec模型训练生成词向量。然后使用已经构建好的seq2seq模型训练，从而能够让训练好的模型预测生成准确的关键词。

关键词：关键词抽取；神经网络；序列到序列模型；word2vec；机器学习

**LEGAL ISSUES KEYWORD EXTRACTION SYSTEM**

**ABSTRACT**

Keywords are usually a short and summative content that can summarize the topic information in the text content. High-quality keywords can provide people with highly summarized and valuable information. Keyword extraction is one of natural language processing. Important task. In the legal question answering system, the questions are submitted by ordinary users rather than by legal experts. Due to the limitations of the professional field, the texts submitted by ordinary users are more colloquial, and they do not use professional legal vocabulary. The purpose of this design is to generate a professional legal question keyword that the user wants to consult based on the colloquial legal questions submitted by the user.

**Key words:** Legal issues; keyword extraction; python; machine learning

目录

[ABSTRACT 4](#_Toc12649)

[1 绪论 1](#_Toc28257)

[1.1 课题背景及研究的目的和意义 1](#_Toc26669)

[1.2 研究现状及发展趋势 1](#_Toc25183)

[1.2.1 研究现状 1](#_Toc4082)

[1.2.2发展趋势 1](#_Toc31219)

[1.3 课题主要研究内容 1](#_Toc20024)

[2 开发工具及相关技术 2](#_Toc6976)

[2.1 系统开发工具 2](#_Toc5266)

[2.1.1 Python3.6 2](#_Toc3465)

[2.1.2 Pycharm 2](#_Toc8525)

[2.1.3 NVIDIA GEFORCE GTX 850M 2](#_Toc8627)

[2.2 中文分词 2](#_Toc1629)

[2.2.1 基于字符匹配的分词方法 2](#_Toc7164)

[2.2.2 基于理解的分词方法 2](#_Toc28856)

[2.2.3 基于统计的分词方法 2](#_Toc20589)

[2.2.4 jieba分词 2](#_Toc31620)

[2.3 word2vec模型 2](#_Toc12235)

[2.3.1 词向量表示 2](#_Toc30151)

[2.3.2 CBOW模型和Skip-gram模型 2](#_Toc12558)

[2.3.3 gensim框架 2](#_Toc30326)

[2.4 Pytorch机器学习框架 2](#_Toc9138)

[3 带注意力机制的seq2seq模型 3](#_Toc14870)

[3.1 循环神经网络 3](#_Toc6484)

[3.2 seq2seq模型原理 3](#_Toc24139)

[3.3 attention机制 3](#_Toc18814)

[4 需求分析 4](#_Toc29455)

[4.1 数据集需求分析 4](#_Toc10383)

[4.2 功能需求分析 4](#_Toc22115)

[4.2 性能需求分析 4](#_Toc29471)

[5 系统实现 5](#_Toc14002)

[5.1 语料收集和预处理 5](#_Toc14677)

[5.2 分词模块 5](#_Toc14666)

[5.3 词向量训练 5](#_Toc4607)

[5.4 搭建基于RNN的序列到序列模型 5](#_Toc28050)

[5.5 训练模型 5](#_Toc7030)

[5.5.1 带监督的训练 5](#_Toc15071)

[5.5.2 使用GPU加速训练 5](#_Toc24722)

[5.5 系统运行与测试评估 5](#_Toc9760)

[7 总结 6](#_Toc10802)

[参考文献 7](#_Toc30134)

[致谢 8](#_Toc18991)

# 1 绪论

## 1.1 课题背景及研究的目的和意义

## 1.2 研究现状及发展趋势

### 1.2.1 研究现状

### 1.2.2发展趋势

## 1.3 课题主要研究内容

# 2 开发工具及相关技术

## 2.1 系统开发工具

### 2.1.1 Python3.6

### 2.1.2 Pycharm

### 2.1.3 NVIDIA GEFORCE GTX 850M

## 2.2 中文分词

### 2.2.1 基于字符匹配的分词方法

### 2.2.2 基于理解的分词方法

### 2.2.3 基于统计的分词方法

### 2.2.4 jieba分词

## 2.3 word2vec模型

### 2.3.1 词向量表示

### 2.3.2 CBOW模型和Skip-gram模型

### 2.3.3 gensim框架

## 2.4 Pytorch机器学习框架

# 3 带注意力机制的seq2seq模型

## 3.1 循环神经网络

3.1.1 RNN模型结构及原理

3.1.2 双向RNN

3.1.3 GRU模型

## 3.2 seq2seq模型原理

## 3.3 attention机制

# 4 需求分析

## 4.1 数据集需求分析

## 4.2 功能需求分析

## 4.2 性能需求分析

# 5 系统实现

## 5.1 语料收集和预处理

5.1.1 数据爬虫

5.1.2 数据预处理

## 5.2 分词模块

5.2.1 添加自定义词典

5.2.2 去除停用词

5.2.3 jieba分词

## 5.3 词向量训练

5.3.1 使用gensim训练word2vec模型

5.3.2 保存词向量模型

## 5.4 搭建基于RNN的序列到序列模型

5.4.1 编码器-解码器

5.4.2 NLLLoss损失函数

## 5.5 训练模型

### 5.5.1 带监督的训练

### 5.5.2 使用GPU加速训练

## 5.5 系统运行与测试评估

# 7 总结

参考文献

致谢