



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；序列模型；word2vec；机器学习

**LEGAL ISSUES KEYWORDS EXTRACTION SYSTEM**

**ABSTRACT**

Keywords extraction of legal issues is the application of current keyword extraction technology in China's legal scope,legal issues are closely related to people's daily lives and social norms.At present, China's legal documents and terms are constantly improving and the amount of data is huge due to many reasons such as China's large population and complex social conditions.The vast majority of people seek their own legal advice or assistance through the use of various online legal question and answer systems due to professional restrictions.How to extract the accurate data that people need from the large amount of data of the legal question and answer library is the focus of current research.Since most of the current keyword extraction techniques are based on statistical methods in long texts, the keyword extraction in legal questions and answers (short text) is not good effective.

Therefore, this paper mainly generates keywords from a short legal issue by constructing a sequence-to-sequence (seq2seq) model based on the Recurrent neural network (RNN).First, from the legal Q&A website, we use Crawler technology to collect a large number of legal questions and keywords that have been artificially labeled, and pre-process the (question-keyword) corpus.After the Chinese word segmentation of the corpus, the word2vec model is used to generate the word vector. Then use the already built seq2seq model to train, so that the trained model can predicts accurate keywords.

**Key words:** Neural Networks; Sequence model; Word2vec; Machine Learning

目录

[ABSTRACT 4](#_Toc7719)

[1 绪论 1](#_Toc730)

[1.1 课题背景及研究的目的 1](#_Toc12570)

[1.2 国内外研究现状及发展趋势 1](#_Toc28996)

[1.2.1 国内外研究现状 1](#_Toc28728)

[1.2.2发展趋势 3](#_Toc5239)

[1.3 课题主要研究内容 3](#_Toc26131)

[本课题主要研究一种基于序列到序列的法律问题关键词抽取系统，系统主要分为 3](#_Toc4386)

[2 开发工具及相关技术 4](#_Toc24446)

[2.1 系统开发工具 4](#_Toc14396)

[2.1.1 Python3.6 4](#_Toc2875)

[2.1.2 Pycharm 4](#_Toc16518)

[2.1.3 NVIDIA GEFORCE GTX 850M 4](#_Toc1028)

[2.2 中文分词 4](#_Toc1245)

[2.2.1 基于字符匹配的分词方法 4](#_Toc7)

[2.2.2 基于理解的分词方法 4](#_Toc13918)

[2.2.3 基于统计的分词方法 4](#_Toc30983)

[2.2.4 jieba分词 4](#_Toc31426)

[2.3 word2vec模型 4](#_Toc19994)

[2.3.1 词向量表示 4](#_Toc29005)

[2.3.2 CBOW模型和Skip-gram模型 4](#_Toc31601)

[2.3.3 gensim框架 4](#_Toc24449)

[2.4 Pytorch机器学习框架 4](#_Toc21193)

[3 带注意力机制的seq2seq模型 6](#_Toc26040)

[3.1 循环神经网络 6](#_Toc17676)

[3.2 seq2seq模型原理 6](#_Toc858)

[3.3 attention机制 6](#_Toc1406)

[4 需求分析 7](#_Toc4634)

[4.1 数据集需求分析 7](#_Toc5274)

[4.2 功能需求分析 7](#_Toc12241)

[4.2 性能需求分析 7](#_Toc25337)

[5 系统实现 8](#_Toc13387)

[5.1 语料收集和预处理 8](#_Toc21606)

[5.2 分词模块 8](#_Toc31379)

[5.3 词向量训练 8](#_Toc27779)

[5.4 搭建基于强化学习的序列到序列模型 8](#_Toc16761)

[5.5 训练模型 8](#_Toc31154)

[5.5.1 带监督的训练 8](#_Toc17767)

[5.5.2 使用GPU加速训练 8](#_Toc30746)

[5.6 系统运行与测试评估 8](#_Toc1472)

[7 总结 9](#_Toc30207)

[参考文献 10](#_Toc31329)

[致谢 11](#_Toc19406)

# 1 绪论

## 1.1 课题背景及研究的目的

在互联网大数据时代背景下，互联网使用人数以及各种终端设备如智能手机、平板、笔记本爆发式增长。信息时代为人们带来便捷的同时，网络上庞大且冗杂的数据也让人们常常需要花费许多精力去筛选得到想要的信息。百度、Google等搜索引擎在一定程度上协助人们在如此庞复的网络数据中找到需要的信息，然而搜索引擎只支持使用关键词进行检索，且在日益增长的数据量情况下对关键词精准度的要求也越来越高。关键词是简短且具有高度总结性的文本，需要用户根据自身知识来概括提取。但在法律问题方面，由于用户专业限制，无法保证依靠人工提取关键词的精准性，从而也影响了搜索引擎返回信息结果的准确性。

我国坚持依法治国的基本方略，保证社会治安稳定、社会文明进步以及国家长治久安。要求人们人民和执法部门有法必依、违法必究、执法必严，人们的法律意识和知识水平也日益提高。随着我国法律制度体系越来越完善，法律条款及文件的数量越来越大且复杂，非法律专业人士由于专业限制无法定位具体的法律范畴。法律社区问答网站（如：www.51wf.com）的出现给人们提供了一个很好的专业法律咨询的平台，用户在网站上提交一段较为简短、口语化的法律问题，会有法律专业人士提供非常专业的回答，并将该问题打上标签作为其关键词，能够迅速定位用户所咨询问题所属的法律专业范畴。本课题研究的目的是通过收集上述法律社区问答网站上的大量法律问题-关键词数据，运用自然语言处理、机器学习等技术训练，得到一个可准确挖掘出法律问题中包含的关键词的模型。该模型能够根据用户输入的法律问题，自动生成关键词，帮助用户定位问题的范畴并指引用户搜索的方向，能够大大提高用户获取需要的法律信息的效率。

## 1.2 国内外研究现状及发展趋势

### 1.2.1 国内外研究现状

关键词是一段简短的总结性内容，用来表达文本的主要语义，高质量的关键词可以促进对文本内容的理解，组织和访问。因此，国外许多研究都集中在从文本内容中自动提取关键短语的方法，它已广泛应用于许多应用。如Jones和Staveley在1999年提出信息检索、Hulth和Megyesi在2006年提出的文本分类、Berend提出的意见挖掘等应用。大多数现有的关键词提取算法通过两个步骤解决了这个问题，第一步是获取关键词候选列表，研究人员试图使用具有某些词性模式的n-gram或名词短语来识别潜在的候选关键词，如Medelyan从维基百科等重要语料库中提取n-gram，并根据预先建立的规则提取名词或短语。第二步是通过有监督或无监督的机器学习方法和一系列人为定义的特征将候选关键词对文本内容的重要性进行排名。在带监督的机器学习中，关键词提取的任务可以归为文本分类问题，如Frank使用朴素贝叶斯训练分类器；在无监督的机器学习中，Mihalcea等人提出计算候选关键短语之间的相关性。

上述关键词提取方法主要存在两个主要缺点：首先，这些方法只能提取源文本中出现过的的关键词；它们无法预测具有略微不同顺序的有意义的关键词或使用同义词的关键词。然而，法律问答系统通常根据其语义来分配关键词，而不是根据文本内容。由于法律问题是由非专业用户口语化提出的，关键词往往不包含在问题中，所以通过上述的方法提取效果不佳，这进一步促使开发更强大的关键词预测模型。其次，在对候选关键词进行排名时，以前的方法通常采用机器学习算法，如TF-IDF和PageRank。然而，这些特征仅基于词语出现和共现的统计来检测文档中每个词语的重要性，并且不能揭示作为文本内容基础的完整语义。为了解决短文本中关键词难以使用统计方法抽取的问题，诸多学者已着手研究从短文本中生成关键词的技术，如Zhang等人提出了一种用于短文本提取关键词的递归神经网络模型。但是，这些模型仍然不能抽取原始文本中未出现的关键词。

在国内，法律问题关键词自动提取也有不少研究。由于中文不像英文那样有明显的分词界限，且中文是世界上语法最复杂困难的语言之一，这在一定程度上增加了中文关键词提取的难度。在中文关键词提取之前，需要对语料进行分词处理，目前许多研究人员已经提出若干优秀的中文分词工具，如HanLP、jieba分词、FudanNLP等。在基于语义的关键词提取方法中，王立霞等提出使用《同义词词林》计算每个词语的语义相关程度,构建一个语义相似度网络,利用中介密度来衡量每个词语的重要性并将词语的语义特征整合到关键词提取的过程中,效果较好。

### 1.2.2发展趋势

由于我国疆域辽阔人口众多，社会情况和地方文化复杂，我国的法律文献及条款众多并且将来还会不断的修订和完善。因此整个法律体系所涉及的法律问题类型分类以及各种法律问题检索的的复杂度会越来越高。当人们碰到法律问题时往往只能够做出口语化的描述，并不能准确的定位某个法律问题所属的法律专业范畴。法律社区问答网站的出现使人们咨询法律的途径越来越便捷，但是同时由于网络信息数量巨大，且大部分法律社区问答网站需要通过专业人士人为解答、人为标记关键词，网站无法立即反馈准确关键词信息给用户。当前的法律问题关键词提取技术和法律问题分类还不是特别完善，这就有可能产生用户被误导，混淆法律概念的后果，因此要想真正满足人们和社会的需求，法律问题关键词提取方面还需进一步探索，主要体现在：

（一）准确性：法律问题与个人、公司或部门、政府等单位的利益以及社会制度密切相关，当涉及到这些内容时，法律条款和文件起到了决定性的作用。这就要求法律问题关键词抽取系统能够给出精准的关键词来定位问题所属法律范畴。

（二）易用性：法律问题一般是由非专业人士提出的，由于我国法律文件包含的专业词汇数量巨大，普通人无法准确使用专业词汇描述某一法律问题，描述内容一般偏口语化。再者中文本身语法复杂多变且不断有新的网络流行词汇诞生，所以如何从用户口语化描述的问题中提取专业法律词汇关键词将成为今后研究的重点。

（三）完整性：社会不断的发展，我国法律文献也随之不断修订和完善。法律关键词抽取模型必须参照当前最新最完整的法律文献库，这就要求研究人员以及专业法律机构共同努力收集全面的法律语料，保证模型能够涵盖所有用户需要的法律问题关键词。

## 1.3 课题主要研究内容

本课题主要研究一种基于序列到序列的法律问题关键词抽取系统，系统主要分为语料收集和预处理、中文分词、词向量训练、搭建基于强化学习的序列到序列模型、训练模型等模块。

语料收集和预处理：从法律社区问答网站（www.51wf.com）收集大量的法律问题语料，并对数据进行特定格式化以及清理，提供给之后的模块训练。

中文分词：采用综合了基于字符串匹配的算法和基于统计的算法的jieba分词工具进行法律语料的分词工作。

词向量训练：使用Word2vec的CBOW模型，对分词后的法律语料进行训练，生成所有词汇的向量模型。

搭建基于强化学习的seq2seq模型：构建基于seq2seq的RNN模型，将处理完的预料作为输入输出序列，在其中引入强化学习用来训练语料库。

训练模型：在NVIDIA GEFORCE GTX 850M上使用Pytorch调用其CUDA架构进行GPU加速训练，生成一个能够预测关键词的模型。

# 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 中文分词

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 搭建基于强化学习的seq2seq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

## 5.6 系统运行与测试评估

# 7 总结

参考文献

致谢