BeltaGo—法律智慧服务者

作品简介

本作品是一个以深度学习、机器学习、大数据等为支撑的自主系统，旨在设计一个面向法律工作者的人工智能系统，具体包括智能信息抓取、机器学习以及智能分析预测等模块。通过对语音、视觉以及自然语言的识别和理解，利用小样本训练集建模，实现对案件资料进行有效的内容抓取和信息分类。通过数据和算法对法律规范、案件事实、历史诉讼类型和请求、判决依据等内容进行自我学习和决策，提供对案情具体走向以及胜败诉概率数据，为法律工作者提供合理的驳论依据参考，以提高审查分析以及判决结果的公平性和客观性。

本作品符合国家新一代人工智能发展规划中的智慧法庭建设内容，将提供一种人工智能在证据收集、案例分析、法律文件阅读与分析中应用的新思路。



第1章 立项依据

1.1 项目研究背景和意义

随着科技的发展与进步，社会生产力水平的不断提高，新一轮的科技革命和相关产业变革，使得人们的生产生活方式出现了前所未有的改变。各种各样的最新技术，正在逐步的进入人们的生活，以其特有的方式，在各个方面展现出其特有的优势，人工智能、大数据、云计算、区块链等关键词更是成为时下最热。

同时，随着人们生活水平的提高，法律意识也开始不断的增强，越来越多的人把社会、经济生活中的纠纷诉诸于法律。但在现实生活中，基层法官的人数十分有限。与此同时，对于法律工作者来说，越来越复杂的法律条款和法律解释，也成为他们工作效率和服务难度的巨大考验。并且，由于业务水平的良莠不齐，律师们对于同一案件的处理往往并不相同，甚至可能出现错误。

因此，发明一种可以辅助法律工作者们进行法律条款的检索，甚至帮助律师和法官们对案件进行预判决的系统就尤为重要。而“人工智能+法律”的组合，则可以很好的满足这样的要求。

1.2 国内外研究现状

从2016谷歌公司的阿尔法狗之后，人工智能就开始频繁出现在公众视野之中，截至目前，以“法律人工智能”为检索关键词，百度搜索的结果已经超过500万条。但要追溯起来，人工智能与法律的结合，却已经有30年的发展历史。1987年，首届国际人工智能与法律会议在美国召开，开启了“法律+人工智能”的研究篇章。发展到如今，越来越多的互联网公司推出了自己的相关产品。

在不同的法律领域中，不同的企业也各有侧重，其中最具有代表性的主要有以下几种：

1. 美国的硅谷的黑石发现公司(BlackStone Discovery)研发了一种能够进行自然语言处理的软件，专门用来处理法律文件。该软件与初级水平的律师相比，在准确性上提高了40%，而在效率层面上更是提高了500倍，综合而言，将官司成本节约了99%，适合有大量法律文件需要分析和审查的大型公司和个人。
2. 美国的ROSSIntelligence公司率先推出的人工智能律师ROSS。ROSS应用本身与Siri类似，可以通过语音识别客户咨询的问题，并对其进行回答。当客户进行咨询后，ROSS会匹配数据库，找到相匹配的答案，给到用户。并当用户进行反馈后，ROSS会根据反馈进行修正自己的结果。
3. 世界上第一个机器人律师Do Not Pay，在英国、美国等被人普遍使用，它可以帮普通民众处理交通罚单，且成本极低。

而在我国，虽然“人工智能+法律”的概念是近几年兴起，但“司法智能”在我国的发展却也十分迅速。例如，我国的法律数据查询类APP“法律家法律数据库”，该APP收录了100余万国家、地方法律法规文件和全部最高人民法院、最高人民检察院发布的指导性案例。但也需要意识到，这些软件或者系统，与真正的人工智能仍有一定的距离。

1.3 主要实现功能

(1)基于语音识别和图像识别的信息录入、提取和保存。法律工作者可以直接通过语音或者上传合同照片的形式，在APP上进行信息的识别和保存，并支持对合同类文件的格式修订。该功能使法律工作者可以减少很多重复而繁琐的简单劳动。

(2)法律条款和法律解释的检索审阅。这是本APP的基本功能，在理解了用户上传的信息后，系统会根据特定算法，检索数据库，提供检索结果。并且同时提供了用户评价的接口，在用户对结果进行评价后，将数据传回云服务器进行更新，优化系统。

(3)案件预决策。该功能主要用于辅助律师对案件进行客观的判断。基于预先构建、不断更新的案例数据库，系统会提取案件的重要信息，并对案件的走向和结果作出预测，为律师提供一个参考。同时，当案件判决后，律师可以上传案件资料，传回数据库，以便系统学习后能够更好的服务。

(4)律师训练。这是本APP的独特功能之一，用于对新手律师熟悉案件流程，提高自己的业务能力。提供全真数据库，可以基于用户数据对律师进行评级，为新手律师进行针对性的训练。

第2章 创新点分析

2.1 整体方案设计

基于人工智能的法律APP，主要利用移动终端与云服务器的配合，构建系统。首先基于现行的法律法规和经典性案例，打造一个可供检索和咨询的大数据库，作为整个系统的基础。其次是移动终端的设计，移动终端将主要分为律师入口和基层法律服务工作者入口，根据不同的用户提供不同的服务。通过手机等移动终端的信息反馈，将用户的反馈数据同步到云服务器，利用云计算平台计算数据，评估APP给出决策的匹配程度，对给出的决策进行不断优化，从而变得更加“聪明”。

如图2-1所示，为本系统的整体结构和功能实现流程方案，主要包括移动终端模块和云服务器的模块。通过控制流以及信息流的走向展示了系统的基本运作流程。移动终端作为用户与云服务器平台的桥梁，提供给用户系统的配置功能和人机交互显示功能，并将所得到的用户反馈信息送入云服务器，而云服务器也可以借此来优化自己的服务。云服务器作为系统的功能核心，起到接收数据、保存数据、处理数据、反馈数据，并对用户的反馈信息进行数据优化的功能。

图2-1 系统总体结构图

2.2 信息识别和提取

信息的识别和提取是本APP的一个基础。对于传统的法律检索和审阅来说，首要的就是对案件的信息进行判断，比如审查合同的信息，判断其是属于刑事案件还是劳动纠纷等，或是审查涉案金额，判断案件的级别是行政级还是刑事级，而这样的简单重复的审查却需要耗费大量的人力物力。

在本APP中，对于信息的提取和识别，我们将会运用图像识别技术和语音识别技术。对于提交的合同类文件等图像信息，我们会利用图像识别技术，轻易地将合同内容提取，包括甲乙双方的协定和报酬、义务等。并且，借助人工智能的可学习性，可以智能的将所提取的信息进行逐层次的分类。例如基于Adaboost算法和利用层叠分类器，对合同的内容进行多层次结构的分类。举例来说，在对合同进行分类为经济纠纷和人身安全问题之后，可以再从经济纠纷问题中，细分出借贷、投资等不同的小类。而对于操作不便的用户，我们也提供语音识别的信息录入方式，更好的服务有需求的人。

总的来说，在对信息的识别和提取部分，我们要做的主要是将用户提供的案件信息进行更好的识别和处理、分类，将所涉及到的法律条款范围缩小，在给用户提供更精准服务的基础上，也将工作效率提高了许多。

2.3 资料检索查阅

资料的检索和审阅，是本APP提供的一个非常必要的辅助功能，在2.2的信息识别和提取之后，经过分类，我们将案件相关的法律条款锁定在一定的范围之内，通过在大数据库中检索相关的法律条款和经典案例，提供给用户。并且，我们可以基于人工智能的学习性，对所提供的答案进行一个满意度和符合度的信息呈现。在提供给用户答案时，系统会同时提供一个符合度数据和一个满意度调查接口。用户在对答案进行满意度反馈之后，移动终端将用户数据反馈给服务器，服务器整合数据之后，以符合度的形式，在答案的下方显示出来。智能化的检索、查阅和可量化的满意度数据，使得用户可以更好更直观的了解案情，并作出自己的判断。

2.4 案件预决策

法律资料的检索和审阅，对于律师和基层法律服务工作者都非常重要。而对于律师来说，如何帮助客户打赢官司，如何找到足够的证据链，以提高案件成功的几率才是他们最关心的。因此，案件的预决策可以说是本APP最为重要的功能之一。

通过2.2信息识别和提取，2.3资料检索查阅之后，对于案件的有关信息和相关条款都有了一个较为清楚和明确的界定。自从之后，就是对于案件结果的一个较为提前的预测。在预决策的部分，我们同样创造性的引入人工智能的理念，在锁定法律条款的基础上，我们通过大量的经典案例和后期加入的时事案例，对系统进行训练。系统经过对这一系列案件的分析和学习，得到较为完备的惩罚依据，结合具体案件中ABC等各方的行为和以往案例的惩罚尺度，给出一个较为合理的惩罚区间。

同时，我们还会提供案件中不同角色的模拟接口。用户在使用APP进行预决策分析时，可以选择不同的视角，进行多角度的分析。例如，在用户选择原告视角时，提交了所拥有的证据线索之后，系统在给出惩罚区间的同时，也会给出指导，比如怎样使证据链更加完善等；而用户也可以在此基础上，选择被告视角，系统会给出如何应对原告的诉讼和相关的应对措施。转换身份和视角，使得律师在应对具体案件时可以多角度多方面，有效避免遗漏重要信息，从而保证证据的完整性。在此基础上，该APP还会进行期望性预测，综合分析案情之后，针对当事人的处境，进行有利和不利的预测，以便用户在最终判决结果出炉之前做好准备。而这些决策功能的准确性，则主要来源于已有的经典案例和用户提供的现实案例，系统通过训练、学习，不断提高自己的性能。

2.5 律师训练模式

本APP所针对的主要人群是法律工作服务者和律师，在立项依据中我们有提到，当前，法律行业存在着僧多粥少、良莠不齐的问题，即有法律需求的用户很多，但法律从业者却很少，且其水平往往参差不齐，难以满足众多复杂的案件。

因此，本APP为广大的律师群体提供一个对于自己业务能力的训练平台。初级水平的律师，可以在本APP上得到大量的经典案例和部分的时事案例进行训练。并且，引入人工智能的思想，该APP会在分析律师的训练结果之后，对律师进行分级，提供不同级别的训练案例，做到对不同层次的律师有着不同层次的针对性训练。这些律师训练之后的数据，也会在之后反馈到云服务器，进行数据的统计和分析，系统在学习和训练之后，为律师提供更好的服务，做到良性循环。

2.6 信息安全和维护

身处大数据时代，互联网背景之下，信息的安全问题尤为重要。而对于一个APP来说，用户的数据往往都存储于云服务器之上，这也难免会让用户担心自己的数据泄露。并且，由于法律信息的敏感性，一旦出现用户数据泄露问题，后果更是不堪设想。作为服务用户的APP，如果做不到对用户信息的保护，也会出现严重后果。因此，本APP涉及到的用户信息，将只在云服务器保留7天时间。APP在生成相关结果后，会提醒用户及时下载报告，而在设定时间到达之后，不论用户是否下载了报告，都会在云服务器中将其删除，从源头上保护用户的隐私。

第3章 实施方案

3.1 实现方法与技术

本APP是基于人工智能的法律APP，运用人工智能的理念和方法是本APP的关键。其中，如何体现人工智能的理念，通过自然语言处理，实现对信息的识别和提取，即获取用户的描述和上传的图片，去除其中的噪声干扰，并进行提取摘要等操作，是本系统中的一项关键技术。此外，如何利用机器学习的技术，从大量的用户反馈数据中，学习出有效的模型，并在检索查阅资料和案件预决策中起到作用，也是本系统中的一项关键技术。

### 3.1.1 基于自然语言理解的文字识别技术

在本APP的功能实现过程中，首先要做的是获得用户所提供的信息，包括合同文本、合同照片和语音输入等，去除噪声后提取其中的文本信息，并将之转化为APP可用的规范化格式。目前而言，在自然语言处理方面的研究，主要有自然语言理解和自然语言生成两个方向，分别用于计算机的“听懂”和“回答”。在本系统中，暂时只需要做到能够“听懂”，即可以把用户输入的语音信息和文本图片信息转化为自然语言。

从世界范围来看，绝大多数语种使用的都是拼音类文字，即由各种字母、字符组成，而计算机对于拼音字母类的识别已经比较完善，因此在该研究方向上来说较为简单。但是对于我们的汉语来说，汉字的编码方式和语句构成与其他语种都不同。首先就是分词的问题，对于拼音文字来说，词与词之间有着明显的空隙，每个词代表的意思相对独立，便于计算机的识别。而汉字的词与词之间则没有明确的间隔，且大部分汉字都是组合存在，有其特有的意义。事实上，无论是实现自然语言理解，还是自然语言生成，都是非常复杂的过程，其中涉及语言学、心理学、逻辑学、声学等科学，因此要实现通用的高质量自然语言处理系统极为困难。

在本APP中，我们针对法律法规和相关的合同文本条款，构建数据库和专家系统，并搭建自动文摘系统。首先，合同类文件的规范格式是相对固定的，因此我们会先构建一个包含了规范形式合同的数据库，用于对输入信息的比对和提取。同时，会提供一个专家系统的接口，对于那些格式存疑的合同进行人工的校验，并将最终的数据反馈给数据库进行更新。其次，在进行文字识别时，我们使用OCR技术，识别印刷体文本和手写体的文稿，并进行版面分析、表格识别等。

本基于自然语言理解的文字识别技术有如下优点：

1. 由于针对性的进行了数据库的搭建，也引入了专家系统的接口，因此与传统的自然语言处理系统相比，本系统对于合同类文件的识别能力更强，且效率更高。
2. 利用OCR技术进行纸上字符的识别。随着扫描仪、数码相机等电子设备的普及，限制OCR技术的硬件要求几乎不再存在，因此在进行操作时，会非常简单便捷，且识别率较高。
3. 加入专家系统，并引入机器学习的思想，使得对合同类文本的识别更为专业，且可借助用户的反馈进行不断的优化，从而使系统的工作更为准确。

### 3.1.2 基于机器学习的资料处理、决策技术

本APP的核心功能是基于用户所提供的信息，进行资料的检索查阅，并给出预决策。而在这个过程中，我们用到主动学习的方法。我们知道，人工智能的核心是机器学习，主要研究的是如何利用经验数据提高自身性能，而充分并且高质量的数据则是学习有效性的基础和关键。我们在利用一些传统的监督学习方法进行分类的时候，往往会遇到这样的问题：训练样本的数量越多，质量越高，那么分类的效果就越好，但在现实场景中，高质量的训练样本却非常少。因此，我们引入主动学习的思想。如图3-1，为主动学习的过程示意图。



图3-1 主动学习过程图

主动学习通过利用一定的算法，查询到最有用的未标记的样本，交由专家系统进行标记，然后将这些标记之后的样本加入训练集，系统基于新的训练集进行更新，训练出更新后的模型，进行下一轮的迭代。在主动学习的这一过程中，关键性的任务就在于如何设计出合理的查询方案，也就是如何设计出一定的准则，按照这样的准则来查询实际需要的未标记样本。目前来看，查询的方法可以分为三种策略：基于信息量的查询策略、基于代表性的查询策略以及综合多种准则的查询策略。

在本APP的样本数据池中，主要由用户提供的大量信息组成，这导致了数据池中可能存在着部分的错误信息。同时，由于法律法规和合同类文件的规范性，我们预置的模型往往是较为精确的，需要改动的较少。因此，我们选择基于信息量的查询策略。基于信息量的查询策略在目前而言是最为常见，它可以对那些能最大限度减少当前模型不确定性的样本进行查询。在查询过程中，基于已标记的样本，我们通过预设模型的预测置信度、模型错误率下降期望等指标，对查询样本进行度量，从而对样本的信息量进行查询。

本技术的优点在于：

1. 大数据时代，APP的功能和效率需要不断更新，为更多的用户提供更高质量的服务。通过主动学习的方法，基于用户的实际反馈，对系统的数据库进行更新，满足了用户的实际需求，并完善了系统。
2. 以预先构建的数据库为基础，基于已标记好的样本，对未标记的样本进行查询，通过预设的各参数指标，最大限度减少了用户提供的错误信息对规范样本的影响。
3. 引入专家系统的接口，对那些系统难以辨别，或是辨别可能有误的数据进行人工处理，既保证了效率，又不失准确。

3.2 系统构成分析

3.2.1 云服务器

云服务器是一种简单高效、安全可靠、处理能力可弹性伸缩的计算服务。事实上，云服务器就是一个集成的大型软件平台，包括了云计算中心和云数据中心。

云计算中心基于模块化设计，包括了数据采集模块、数据处理模块和相应的应用模块，各模块之间通过TCP/IP协议进行通信，其功能模块关系如图3-2所示。数据采集模块负责采集用户的反馈信息，并将其送入数据处理模块，经处理后，将所得的结果送入应用模块，并将最终的结果传给用户的终端。

云数据中心则是主要作为系统的存储中心，主要包含了用户个人信息数据库、各种不同类型的法律法规的基础数据库以及合同类文本规范数据库。个人信息数据库主要用于存放用户个人的检索报告和查阅记录等，可提供给用户查看。各种不同类型的法律法规的基础数据库则用于资料的检索，同时，会根据最新的法律法规和用户反馈进行更新。合同类文本规范数据库则用于对合同类文件的审阅，主要在识别用户的输入信息时起到作用。

图3-2 云服务器工作模块

3.2.2 移动终端

移动终端主要是指各类能够与互联网连接，与云服务器通信的现代通讯设备，例如手机、平板电脑等。在这些移动终端上安装我们设计好的应用程序，就可以使用我们所提供的服务。

应用程序作为人机交互的中间桥梁，起着连接作用。本应用程序提供一个图形界面，主要有这么几个功能：系统配置、知识科普、功能工具和联系合作。系统配置不必说，就是对系统的界面等进行选择。知识科普，则是基于最新的法律法规条款和相关解释，推送给用户，并会以具体事例的形式生动的给出。功能工具，则是该应用程序的主要部分，包括了文件上传、资料审阅和预判决的功能选择。联系合作，则是提供了用户与资深律师，与后台工作者联系的渠道。

第4章 完整度

通过查阅现有国内外资料，目前确定了本次基于人工智能的法律app的大致框架。整体app分为两部分，分别为移动终端和服务器。通过终端来获取信息上传到云服务器进行运算后，返回数据给用户终端来实现。如图4-1，为本APP初步构思的开始界面。

图4-1 用户初始界面

在用户终端设计中，将会存在许多关键选项，例如上传文件合同。通过上传文件合同将可以确定本次案件所涉及的具体类型，以及案件中各方处理的经过以及产生的何种纠纷，还包括最终各方处理的结果。其实这是最基础的一步信息收集。信息的准确性和多样性对于本次作品而言是至关重要的，因为以法律角度去审视某件事时，必须要考虑到证据是否确凿，是否符合法律法规，以及是否合乎社会的道德准则等。所以第一步我们会采用图像识别扫描技术，抓取关键词来作为基础。另外，在互联网火热的当下，民众舆论也是可能改变一个案件的走向的，这一关键要素也会被作为信息的一部分。

此外，我们还会设计文本界面，用于记录需要补充的信息。语音界面的设立，除了用于上传语音证据外，还会加入语音识别功能，提取其中的关键字，为第一步的案件信息多样丰富打下基础。

在信息充足有效的情况下，第二步是我们在云服务器所加入的AI学习功能。这一功能目前的进度是处于半完成阶段。将简单的文本信息例如法律条款加入自身的库中用于检索，再将语音识别和文本检索相结合来提供工作效率是目前作品的基础功能。而想要实现AI学习大量的法律案例还是有不少难题需要解决的1）法律条文较多，且会发生更新2）不同地方不同年代对于法律诉讼会做出不同的判罚尺度3）不易将法律案件的全部信息获取，其中会涉及到隐私问题。所以目前提出的解决方案是，1）对于法律条文需要有法律人士来为这个app修订数据库内容2）与地方法院进行合作，筛选出近些年的法律信息作为训练样本，也该类样本不会保存在自己的服务器中。

第三步是分析预测，这一步其实是人工智能的重要体现。现今越来越多的股票预测、彩票预测、胜负预测出现在大众视野里，但是预测这一词很明显是无法保证准确的。而利用人工智能技术，是为了让准确性能提高。这一步目前的进展是，在样本训练学习后，使用者可以选择自己为原告或者是被告，来获取判决的可能结果，判决的可能要点，判决可能涉及的法律条文，以此让使用者能充分了解两方的情况。除此之外，我们还会去寻找资深律师合作，对app的分析预测进行试验和评估，帮助我们优化功能。

第5章 应用推广价值

本次作品主要的应用领域在于法律。法律并不像常人所想，其实法律具有很宽泛的应用层面。随着人们法律意识的增强，对法律服务的需求也越来越大，法律从业者们数量明显呈现出供不应求的情况。本作品的应用主要用于法律从业人士，包括律师、法官等。

在上法庭之前，律师可以通过该app收集申请人或者是辩护人的信息需求，以及私下之间，或者法院下发的文件信息来得到结果分析和预测。根据结果分析，律师可以迅速地寻找到所需要的法律条文，以及对方的驳论点，为申请人或辩护人进行辩论。

除了实际的法律诉讼案件情况，在校园甚至国家人员的法律学习上，可以起到其独到的作用。我们会在app中提供并不断更新一些实例案件，使用者可以在每天浏览中学习到其中的法律知识。根据使用者的浏览习惯以及身份不同，我们也会利用跟踪技术推送不同的案件实例。这不仅能帮助校园中的学生健康发展，也能让国家单位人员秉公执法，加强廉政建设。

推广价值方面，本作品确定的是民间和组织相结合。民间是指法律机构以及法律人士，组织是指各级法院。民间方面，我们会与全国各大法律机构进行商量签约，将其图文信息以及联系方式等录入app中，为使用者提供在线的法律咨询的帮助。对于每个法律机构，我们会根据其知名度、胜诉率等提出不同的签约费。同时，使用者也可以根据app中所提供的许多法律机构，在线去挑选自己想要的法律机构进行线下签订协议，每成立一单，app平台会抽取一定的费用作为收入。而所签约的律师也会通过试验和使用app来给我们做出反馈，相辅相成帮助我们优化不同的功能。组织方面，我们将会寻找法院以及政府进行合作，以法院的大量案件为训练学习样本，不断优化AI系统。我们会将此app引入到法院之中，通过每次法官判决和系统预测分析作对比，来获知该app的缺陷之处，并根据此来改进算法和功能。