# 智能合同文本处理应用组件

### 0、命题企业介绍

浙江鸿程计算机系统有限公司（简称：鸿程系统）成立于1996年，目前注册资本14196万元，是由中国电信与浙江省能源集团主要战略投资的信息化服务提供商，通过研究大数据、人工智能等创新技术，专注于面向数据大脑、数据运营等数字智能化场景应用，构建数据应用解决方案和复杂信息系统集成解决方案，致力于为通讯、能源、交通等重点企业提供高效的数字化解决方案和技术服务，为智慧城市建设提供城市数字化转型和城市数据智能运营服务。

鸿程系统以“创新、沟通、追求卓越”为企业文化的核心理念，倡导“鹰一样的个人，雁一样的团队”企业精神，聚集一流人才，紧跟世界信息技术发展潮流，创建卓越的信息化服务提供商。

### 1、背景说明

随着经济主体越来越多，各种服务模式、销售模式等的复杂多样，衍生模式更加复杂，企业对合同的进行智能化管理显得更加迫切，而合同文本是一种典型的非结构数据，需要通过人工的方式将相关信息进行结构化识别和输入。

随着深度学习、自然语言处理、知识图谱等技术的发展，对合同进行智能处理已经成为可能。

### 2、项目说明

**【问题说明】**

* 合同文本因为格式不同，识别准确率偏低
* 提供的开发接口效率不高
* 模型迭代、维护、交接成本高，缺少统一管理和沉淀

**【用户期望】**

搭建一个合同文本处理能力组件，可以通过API的形式提供如下服务：

**（一）合同信息有效抽取和自动化标签：**

（1）信息抽取功能

适用多样化合同文本，通过自然语言处理、深度学习算法技术，对合同文本如客户名称、金额、签约时间、产权主体等信息结构化抽取。

1. 通过扫描合同文本，获取合同相关信息，以参数的形式返回给编辑界面，供编辑填写确认。例如合同基本信息，收付款信息，银行信息，票据信息，验收信息，进度信息，里程碑信息，知识产权信息，税务要求等等。
2. 要能够根据关键条款和关键字自动识别匹配并填写，要能够根据关键字情况做相应的智能调整，要能够进行容错处理，如果重要关键条款缺失，要提供预警和分析功能。
3. 关键条款识别准确率达到100%。

（2）合同标签功能

根据预先定义的标签库智能自动推荐匹配合适的合同标签，以参数形式返回到合同标签管理界面。合同标签是用于不同维度的合同分类画像，例如按行业分类、按区域分类、按客户类型分类、按规模分类，每个合同的分类数不一定相同。

1. 标签库的配置功能，能够维护和改进标签库，提高分析准确性和有效性。
2. 要能够根据历史已经形成的合同标签信息，不断提高标签匹配程度和判断准确度。
3. 标签准确率达到90%。

**（二）合同文本识别与编辑**

（1）合同识别转换

1. 要能够将图片、pdf等不同文件格式的不同字体合同文件识别成纯文本。能够识别WORD、WPS的文字
2. 要能够将不同来源（客户、供应商）不同排版格式的合同识别成文字后，按原排版格式输出成文本合同。
3. 能够支持多种字库，常用办公软件字库需要全部支持。能够支持多语种，能支持中英文混合的语种内容识别。
4. 对于图片转化识别准确率要达到90%以上，PDF转化识别率要达到98%以上，WORD\WPS\EXCEL等文本合同要在不同文件格式、排版格式、字库与语种模式下文字识别准确率达到100%。

（2）合同起草

根据法务部门预定义的工作场景跟合同场景，将关键条款填入，能够系统自动生成合同格式文本导出。工作场景可以形成一定的格式。对部分关键条款能够进行强制性设置，并进行合规性检查。

**（三）合同审核和风险预警**

（1）在合同审校流程，快速找出不同版本合同修改区域与版本差异，在合同盖章归档场景，有效识别实际签署纸质合同和电子版合同差异

（2）有效识别文书存在纰漏并给出纠错建议和提示，如错别字、敏感信息等；提高文书表达准确度，避免低级错误、保障文书质量

（3）依据规则和要求，可对企业内部合同的合规性进行自动审核；支持根据定义规则，提示风险内容和法条推荐，辅助法务部门提升合规性管理和效率

（4）根据法务、财务、商务部门预先定义的合同场景和合同规范，对合同进行基本的合规性检查，并输出检查结果和预警建议。

（5）合同关键条款和基本信息的配置功能，能够动态的新增、修改和删除关键条款，并配置基本信息和相应的字段。

（6）能够维护合规检查场景和规范。同时需要提供合适的合规检查的格式和规范定义管理功能，方便自动合规检查维护。

### 3、任务要求

采用B/S架构，后端开发使用Java技术栈和当前流行的中间件、大数据平台及组件、数据库知识以及数据库模型设计知识，满足开发当前系统的需要；前端开发使用VUE框架和相关技术栈。

参赛者需设计一套由建模流程各操作封装而成的小组件，要求如下：

1. 配置合同文本识别小组件，功能包括但不限于：通过图片、PDF、WORD等方式自动导入合同文本，识别成标准格式的合同文本。支持图片的倾斜、翻转等情况下的文字识别。文件大小支持>50M，文字识别响应时间<3s，并发>5QPS。
2. 配置合同文本信息抽取组件，能够对合同文本的关键信息进行抽取，提供给调用接口编辑信息，能够跟进定义的字段进行关键信息抽取。
3. 配置合同分类标准组件，通过多种分类标签方法，满足对合同标签的分类功能，可以根据标签库自动匹配，能够通过人工标签进行智能学习，完善标签库。
4. 配置合同审核预警组件，能满足合同审核预警的功能，提供审核结果人工稽核，并根据人工稽核功能学习优化审核规则和机制。
5. 配置合同编写组件，能够预先定义合同模板，根据甲乙方不同选择，提供关键信息接口，利用关键信息填写自动生成合同，根据人工对生成合同的稽核调整自动学习改进合同模板或者合同编制规则库。
6. 服务运行稳定，可用性达到95%以上。形成提供完整松耦合的中间件产品，能够方便集成进各种应用系统中。形成可以独立部署的接口系统，能够对外提供接口调用服务。
7. 提供源代码，系统结构合理，容错性强，代码编写规范，文档说明清晰。文档应该包括功能设计说明书、库表结构说明书、代码包括明确的注释、接口说明书、关键逻辑说明书等。
8. 参赛者自行从网络上选取各种不同的合同文本进行训练和测试。
9. 代码完整、部署使用准确率达到95%以上的，通过测试后，择最优团队给予2万元奖励。

### 其他数据参考信息：

1. 合同基本信息：

合同名称、合同模式、合同框架、合同类别、采购类别、所属项目、合同所属部门、项目经理、合同性质、对方合同编号、对方合同负责人、（供应商、供应商统一社会信用代码、供应商地址）、（签约客户、客户统一社会信用代码、客户地址）、开户银行、开户银行账号、支行信息、开户地址、（原厂商、原厂商统一社会信用代码、原厂商地址）、（最终用户、最终用户统一社会信用代码、业务板块、所属行业）、对方传真、对方联系电话

1. 合同商务信息：

商务人员、市场销售人员、合同开始日期、合同结束日期、自主技术服务金额、外包技术服务金额、硬件及设施金额、合同总金额、税额

1. 物资清单：

材料编号、材料名称、产品类别、产品品牌、税务编码、数量、单价、规格描述

1. 收/付款计划：

节点名称、计划日期、计划金额、计划比例、实际到账金额、实际到账金额比例、已开票金额、已开票比例

1. 合同里程碑：

节点名称、计划日期、计划比例、标志事件/条件

1. 税票要求：

税票种类、税率

1. 分类标签要求：

行业分类、地域分类、技术种类、客户分类