**讯飞智元人工智能平台**

**使用指南**

# 产品架构

讯飞智元人工智能平台产品整体架构如下：



图 讯飞智元人工智能平台产品整体架构

根据新一代公安信息网和公安大数据规范要求，系统实现分层解耦、前后端分离，整个系统遵循标准规范体系、运维保障体系和安全保障体系，整体架构如下：

* **平台服务层**

采用灵活统一的分布式计算架构，将语音识别、图像识别、文本识别的算法以及模型封装在该平台，通过消息调度服务向上层应用提供人工智能的相关算力与处理识别服务。

* **讯飞智元人工智能开发平台**

基于平台层服务和数据服务，平台构建了可以进行AI能力生产的系统，方便第三方厂商进行个性化项目交付。

# 产品功能

## 平台服务层

平台服务层是讯飞智元人工智能平台的核心，也是各类引擎可以灵活高效工作的载体。具体由讯飞智元人工智能平台管理系统、感知智能服务系统、认知智能服务系统构成。

### 平台管理系统

#### 基础管理平台

AI引擎基础管理平台是一套基础软件系统，支持基础设施的伸缩，在硬件资源上分配服务节点，自动运行程序，并且无需知道内部详情，不需要手工将应用程序包部署到每台机器。从资源视角看，AI引擎管理将底层基础设施抽象，简化了应用的开发、部署以及对开发和运维团队的管理。集群中包含多个节点。通过AI引擎管理组件部署引擎服务时，部署者可以在节点间随意对服务进行分配。

##### 集群扩容

可以将新节点加入集群，然后将服务分配到这些节点上，根据需要将应用集群规模放大或缩小。

##### 健康检查

在服务器发生故障时，可以迅速进行告警。随着集群大小的增加，需要频繁的处理出现故障的计算机组件。监控应用程序，并且在应用程序出现故障时重新启动他们。这不仅降低了运维团队的工作压力，也减少了系统从故障中恢复的时间。

##### 日志管理

日志管理平台具备的能力包括：

* 日志得到集中，汇聚在一起；
* 灵活的日志格式化、过滤方法；
* 索引日志内容，快速返回查询结果；
* 具有伸缩性，在各个环节都能扩容；
* 具有高可用性，在单一节点失效的情况下不影响使用；
* 具有图形查询工具、报表产出工具。

##### 服务跟踪

平台记录了系统每天上千万次的服务调用信息，来自前端和后端应用，还有数据库、缓存、存储。以图形化的方式提供完善的服务调用跟踪信息，为服务出错后的快速定位、服务链路的性能优化、服务链路的流程优化等提供了非常有价值的参考数据。

##### ﻿配置中心

实现配置项集中统一管理，避免程序分发的时候逐个修改配置，并且提供了一套配置修改后即时生效的机制。

#### 服务开放平台

##### 登录及首页

* 用户登录

用户登录功能采用最基本的登录模块，即输入用户名和密码即可登录，暂不支持注册，密码修改或找回等功能。

* 平台首页

首页作为平台的门户主页，体现平台的AI智能和科技感，用户进入平台即能够清晰了解平台核心能力。

首页包括四大板块：

1. Banner：通过视觉语言传达平台科技感定位。
2. 工具集板块：展示几个基础工具集，以便捷的工具辅助用户完成日常工作，并加深对能力的理解；
3. 顶部菜单：顶部菜单作为平台的导航入口，点击控制台进入到平台主页；
4. 用户信息：显示用户账号及头像信息

##### 基础工具集

* 实时转写工具（选配模块）

主要面向人人对话场景的语音转文字，广泛应用于会议、客服、采访等场景。可辅助用户完成日常工作中需要将实时语音转写为文本，便于后续进行文字编辑等工作。

实时转写工具为选配功能，使用时需要单独部署实时转写引擎。

具体功能如下：

1. 语音输入：提供语音输入入口，用户可在说话时，点击“开始识别”将语音输入入口打开，接收实时语音。
2. 实时转写：引擎接收到语音后，会实时将语音转写成文本，并在结果页面展示出来。点击“结束识别”，将结束收音。

* 图文识别工具（选配模块）

集成图文识别工具，支持pdf、jpg、png等常见类型图片文件识别为文字；主要应用在文书写作、报告撰写等各办公办案场景中，满足卷宗资料识别、公文文件文字提取等需要。

图文识别工具为选配功能，使用时需要单独部署图文识别引擎。

具体功能如下：

1. 支持在语音助手中快速打开。
2. 支持文字段落和EXCEL表格识别。
3. 支持上传图片进行图文识别，支持jpg、JPEG、png、bmp、tif、PDF文件格式。
4. 可复制当前页面识别结果，也可选择部分结果复制。
5. 可将识别结果导出为word，并支持批量导出。

* 离线转写工具（选配模块）

集成离线转写工具，支持常见类型音频文件转写文字，主要应用在会议、报告、调研、走访等场景中，通过上传录音文件快速转写文字，满足文档撰写需要。

离线转写工具为选配功能，使用时需要单独部署离线转写引擎。

具体功能如下：

1. 支持在语音助手中快速打开
2. 用户可将录音文件上传至系统进行转写，支持目前市场主流录音格式，如：mp3、wav、pcm、m4a等。
3. 支持上传视频文件，并对视频中的音频进行提取和转写，支持的视频格式包括：WMA、AVI、FLV、MP4格式。
4. 可将转写的文本导出，支持批量导出。
5. 音频与文本对照，可以边听边看，方便用户进行校对与编辑。

* 语音合成工具（选配模块）

系统提供语音合成工具，可将用户输入的文字快速合成音频文件，可用于警务告知书宣读、智能问答播报、法律法规条文宣读、纪律宣读、信息播报等多场景下的语音播报的需求。

语音合成工具为选配功能，使用时需要单独部署语音合成引擎。

具体功能如下：

1. 支持在语音助手中快速打开
2. 提供在线合成工具，将用户输入的文字播报出来；
3. 可将用户输入的文字合成为音频文件导出，支持批量导出功能；
4. 提供与人工智能平台相匹配的音色服务供选择；
5. 提供与人工智能平台相匹配的语种供选择；
6. 提供语速调整功能，可控制合成语音的说话速度。

* 多语种翻译工具

本系统主要辅助用户完成多语种文本的快速翻译。

多语种工具为选配功能，使用时需要单独部署文本翻译引擎。

具体功能如下：

1. 支持多语种文本之间的互译。
2. 快速翻译：支持手动输入粘贴文本待翻译文本进行翻译，并提供中文原文与外文翻译结果的左右分栏高亮对照，便于用户进行理解与校对。
3. 查看历史记录：点击查看历史记录，可以查看用户最近操作的5次快速翻译记录，点击记录名称，可查看该任务翻译详情。
4. 翻译结果导出：可将翻译的结果导出，并支持批量导出

##### 平台控制台

控制台主要集成平台核心数据，集成多模块数据汇总显示，让管理员通过控制台了解平台整体运营概况，包括平台能力数量、应用数量，能力调用调用历史，应用及能力用排行前后排名TOP5，能力配额等报表：

* 能力数量、应用数量统计

统计平台接入的能力服务和应用的总量。

* 能力调用概况：

可根据时间段、行政区划、单位、用户统计能力的总调用次数，并以图表的方式展示，供建设单位了解能力的总体调用情况。

* 调用排行

统计一段时间内能力调用次数总和的前五名和后五名。

统计一段时间内，应用调用所有能力总和的前五名和后五名。

* 能力配额概况

1. 展示一段时间内，所有能力的的配额情况，包括关联应用数、调用量、集群授权数、并发额度、已分配配额、可分配配额。
2. 展示能力关联的每个应用的调用详情，包括：应用名称、子能力、调用量、调用量限制、QPS使用量限制、截止时间、分配配额。
3. 可查看每个应用的调用报表、调用数据、并发路数。
4. 应用调用报表主要展示根据时间段、行政区划、单位、用户统计能力的总调用次数，并以图表的方式展示，供建设单位了解能力的总体调用情况。
5. 调用数据主要展示每一次调用的具体信息：调用编号、应用名、服务名、开始时间、结束时间。
6. 并发路数主要展示该应用调用当前能力的并发情况，以图表形式展现。

##### AI能力管理

* 能力管理

能力管理主要向平台输出多种能力的功能版块，该模块可对平台能力进行创建、编辑、查看及删除操作，通过该模块的管理及配置，AI能力才能够支持各个应用调用，满足不同应用场景需求。

1. 能力列表

点击“能力管理”进入能力列表页面。通过列表，可以查看现已建立的各种能力，可查看各能力基本信息，能力名称、并发额度、能力编码、所属类型等信息。

能力列表支持以下操作：

1. 筛选：支持能力名称、能力编码、发布状态等多维度条件筛选；
2. 新建能力：点击新建能力进入新建页面；
3. 导入能力：支持从本地选择文件导入能力；
4. 导出能力：平台现有能力支持导出；
5. 查看能力：点击能力卡片可查看能力详情；
6. 编辑能力：可编辑能力基础信息及接口配置信息；
7. 发布：可对未发布的能力进行发布操作；
8. 下线：可对已发布的能力进行上线操作；
9. 删除：支持对已经创建的能力执行删除操作，对能力执行删除时，需提示用户二次确认。
10. 新建能力

平台创建能力有两种方式，一种可以全新创建能力，一种可以通过本地导入能力，完成能力创建。

（1）新建能力，点击新建按钮，进入新建页面，填写能力的基础信息及接口配置信息。

能力基础信息包含能力名称、能力编码、所属类型、能力描述、演示地址及能力图片，API文档为该能力的文档地址，完善基础信息后点击下一步填写能力接口配置。

接口配置主要包括：接口名称、接口地址、路由地址、接口类型、授权环境、能力分组等。

一个能力可配置多个接口，当配置多个接口后，可以调用该能力的不同服务，子能力可根据业务需求进行配置，配置后，可支持用户根据需求选择。

能力支持设置并发总配额，管理员可根据该能力的实际调用情况，设置并发额度，考虑实际并发通常不会达到所设置的并发额度，可酌情设置并发超额。

（2）导入能力，点击导入可以从本次选择文件导入，导入的能力需要同平台现有能力做代码校验，避免相同能力重复导入。

1. 查看能力

点击能力卡片，可查看该能力的详细信息，查看该能力的基础信息、能力配额、API文档、接口配置信息、安全策略、负载策略、示例Demo及子能力信息。

1. 编辑能力

编辑能力可对能力的基础信息进行编辑、接口配置及子能力配置信息进行编辑修改。

1. 删除能力

支持管理员删除平台能力，能力删除后，将不再显示，对于已经调用该能力的应用也将失效，对能力执行删除时，需提示用户二次确认。

1. 发布/下线能力

（1）发布能力：平台能力，需要发布后用户才可调用，未发布的应用不对用户可见。

（2）下线能力：对于已发布的能力，管理员可对应用执行下线操作，下线后的应用不支持用户调用，已调用的应用也将失效

1. 导出能力

平台能力支持导出，可单个导出，也可批量导出，点击导出选择需要导出的能力点击确定，完成导出。

* 分组管理

为了便于管理API，增加分组管理功能。同一个分组下的API使用相同的环境变量。包括：分组管理列表、新增分组、编辑分组、删除分组、管理环境变量。

管理环境变量：可在每个分组中添加或删除环境变量。添加环境变量时，可设置环境变量类型、变量名称、变量位置、变量值。

* 能力报表

能力报表主要可查看每一个能力在平台的调用情况，被哪些应用调用，调用历史及调用数据。

1. 能力应用调用列表

可查看每一个能力在不同应用下的调用情况，今日调用、上月调用、总调用、调用量限制、QPS限制及调用截止时间，每一个应用调用数据支持导出

1. 能力应用调用详情

可从能力的维度查看具体每个应用的调用历史详情，包括调用报表、调用数据，调用报表支持按时间维度查看不同时间范围的数据，调用数据可查看不同时间的调用数据，调用数据支持下载，便于丰富训练数据库。

##### 应用管理

* + - 应用列表

超级管理员应用列表为平台所有应用，超级管理员可在应用列表新建应用，也可删除编辑列表应用，应用管理员的列表仅为自己通过审核的应用。

* 新建应用

新建应用，需定定义应用名称、应用编码、应用描述、调用数据留痕、及该应用调用的能力，超级管理员新建的应用默认为创建人自己，也可点击设置为应用管理员。

调用数据留痕，如果设置为开，该应用数据调用记录保存且可见，支持用户下载，设置为关，该应用数据调用记录将不保存；

选择应用能力，需要对调用的能力QPS、并发配额、调用量、截止时间等参数进行设置。

* 编辑应用

应用管理员应用列表中的数据，不支持编辑操作，超级管理员应用列表中的数据支持编辑操作。

* 删除应用

支持用户对应用进行删除操作，删除应用时，需要二次确认，应用数据删除后，该应用的历史数据做逻辑删除。

* 查看应用

查看应用可查看应用的基本信息、应用统计信息、用户权限信息及IP地址信息。

1. 应用统计信息

可查看该应用下关联的能力调用历史和调用数据情况，支持用户按照时间维度、用户维度、单位维度、区域维度统计查看，支持运营管理员用户对调用数据下载，丰富数据集，同时可查看该应用下不同能力的并发路数历史趋势。

1. 应用基础信息

应用基础信息包括应用名称基本信息及能力信息，支持在查看信息的时候对应用的基础信息和应用能力进行编辑操作。

1. 用户管理

用户管理主要管控当前应用的使用权限，添加的用户可访问应用，调用能力，对已添加的用户可进行删除操作。

* IP地址管理

通过对IP地址进行管理，设置IP地址黑白名单，增强应用访问安全性和稳定性，应用访问及能力调用灵活可控，IP地址管理如果关，则设置的黑白名单不生效，如果开，黑白名单则生效。

##### 授权管理

* 申请管理

1. 申请列表

申请列表分为待办和已办，待办申请包含未提交、未审核和已驳回的申请，已办申请主要为审核通过的能力。

1. 新建申请

新建申请需要填写应用名称、应用描述、应用调用数据留痕，配置该应用的下的能力，选择能力时，平台会提供通用的默认参数，支持用户根据应用需求自定义各能力的参数。

1. 查看申请

用户提交申请后，可查看该应用的具体详情，详情内容主要为应用名称、应用描述及应用下的能力。

* 审核管理

审核管理主要是应用管理员提交的应用申请，需要由超级管理员审核，审核分为待办和已办，该应用下只要有未审核的能力都在待办列表，该应用下所有能力都已审核则在已办列表。

1. 待办列表

超级管理员可对提交的能力申请进行审核通过或驳回操作，支持按应用批量操作，驳回申请时需要填写驳回原因。

1. 已办列表

对于审核通过的能力，支持超级管理员进行删除操作，此处删除，仅删除申请及审核记录，不影响该能力正常调用。

##### OEM管理

平台部署到生产环境后，支持可视化定制平台名称和平台LOGO，形成自己的品牌化产品。

##### 我的应用

1. 应用列表

应用列表仅显示当前账号下可管理的应用名称，显示该应用创建人，APPID和Scret Key。

1. 应用详情

应用详情可查看应用基本信息和能力信息。

通过能力文档可查看该能力的开发文档。

#### 资源监控平台

统一监控平台针对高要求的分布式、动态化以及敏捷环境提供应用性能监控能力。通过监控、故障排查、性能诊断、业务分析等功能，极大的提升应用可视化、缩短根源问题定位时间。

统一监控平台作为基础支撑，为上层业务系统提供监控、配置、运维服务，及时有效跟进处理故障；对下层硬件提供实时资源监控，以预防生产环境问题产生。事前对故障及时预警，事后提供详实的数据用于追查定位问题，为运维服务提供强力有效支撑。

统一监控平台基于应用、业务和系统等不同角度，全面监控、管理复杂业务应用系统，解决应用性能问题，满足业务需要。

##### 设备接入

* 代理管理

如果监控的设备分布在多个不同网域，且彼此不能互通，就可以分别在不同网域的前置机部署代理，分头采集各自网域的设备指标，上报给监控服务。部署的代理节点会自动注册到监控服务中集中管理。

可以查看连接状态，如果代理假死或服务终止，连接状态会显示为“异常”，每个代理承载的服务器监控数量和拨测任务数量都会在列表中展现，以便用户及时关注到每个代理的运行负荷。

* 硬件设备接入

硬件接入包括了硬件设备本身以及硬件相关的一些基础设施的信息接入。硬件设备本身目前支持服务器（包括服务器上安装的操作系统）和网络/安全设备。接入协议集成了TCP、SNMP、IPMI、NDP/LLDP（链路发现协议）。

1. 主机

主机管理模块是对本系统所有监控的主机进行统一接入和集中管理，主机接入方式支持无侵入式Agent探针自主上报、snmp扫描、IPMI方式拉取管理口数据三种方式。在管理界面中可对主机资源可以查看、编辑和删除操作。通过Agent自主上报的情况下，主机资源通常由监控节点自动注册无需运维人员单独新增操作，个别信息可以由此管理界面修改；如果采用SNMP扫描发现，则需要在目标服务器上开启SNMP服务，并确保UDP的161端口正常可用，才能正常扫描发现；IPMI是基于管理口的硬件设备监控，接入时需要确保维护的IPMI连接信息准确无误。

1. 网络/安全设备

网络/安全设备接入支持以SNMP协议接入常规的网络设备（路由器、交换机）、安全设备（物理防火墙、防毒墙）等。接入时指定IP或IP段，采集代理自动扫描，并将设备信息上报到监控平台。

注：如果设备支持NDP/LLDP协议，在扫描时会自动发现本地端口连接外部网络设备的链路连线。

可查看已接入平台的所有网络/安全设备信息，包括当前设备的运行、访问状态。支持按照设备类型、采集代理、设备IP/名称、负责人和厂商查询。

* 软件服务接入

1. 标准应用

系统部署初始化会给定一个默认应用，工程师可以录入业务应用，将现场运维的业务系统都在该模块管理起来。

应用信息新增时会自动生成一个应用编码，通过Agent自发现主机时，会将该编码复制到Agent的配置文件中，主机自发现时，就会自动将主机挂靠到该应用下。

1. 数据库

由于平台支持以SNMP方式接入主机，SNMP协议无法获取操作系统上一层级的组件信息，因此平台提供远程连接数据库的方式，对常用数据库性能进行监控。

选择数据库所在的主机、数据库类型，并输入端口、数据库登录账号密码和数据库实例名称。保存设置后即建立连接，开始按默认周期采集指标数据。

注：目前支持接入的数据库类型只有三种：mysql\oracle\mssql。

1. 中间件

由于平台支持以SNMP方式接入主机，SNMP协议无法获取操作系统上一层级的组件信息，因此平台提供jmx远程连接的方式，对常用的中间件性能进行监控。

选择中间件所在的主机、组件类型，并输入监控端口、远程连接账号和密码。保存设置后即建立连接，开始按默认周期采集指标数据。

注：中间件的指标采集是基于jmx方式，因此需要预先在中间件的配置文件中加入jmx的监控配置才能实现远程监控。目前支持接入的中间件类型有两种：weblogic/tomcat。

##### IT资源监控

* 监控总览

监控概览主要对当前服务器的接入情况、组件接入情况，运行过程中产生的告警进行趋势进行宏观分析和可视化展示，并提供当前正在告警中的实时快照。

用户可以根据【接入主机】情况，了解平台已接入的服务器总接入数量/存活数量，如果是采用agent方式接入的，可以体现agent终端的存活/不存活数量，可关注到不存活的设备详情。

根据【接入组件】情况，可了解平台接入的组件节点总数有多少，其中多少个节点不可用。

通过【服务健康度】情况，可了解当整个运维环境的运行健康状况，计算规则为：发生异常告警的监控对象数/总的监控对象数。通过“问题服务器”链接，点击可以下钻到问题服务器的监控面板页面，通过“告警数”或“当日新增”（告警数），点击可以下钻到待处理告警列表，关注告警数据详情。

通过【异常发生趋势】了解最近一段周期以来，整个局点的异常分布情况；

通过【实时告警快照】视图，可查看当前主机和组件最近正在发生的异常告警快照。

通过【主机异常动态top10】可关注到运行情况比较严重的10台服务器，并且可以通过链接下钻查看监控详情。

可以关注到设备本身的资源占用越界告警，也能关注到主机上部署的数据库/中间件报告的异常。

* 监控面板

1. 仪表盘

仪表盘视图按九宫格形式展示，每一格展示一个或一组的指标趋势，支持两种运维场景视角：同一个指标在不同监控对象上的差异对比；同一个监控对象的不同指标在一个视图展示。并且支持九宫格全屏浏览。

1. 主机监控面板

操作系统的监控面板采用列表视图展示本系统中所监控的所有服务器资源，在主机管理中自动上报的服务器都会在监控面板中展示。通过查询条件可以检索出想要查看的具体服务器监控信息，列表视图信息包含当前主机状态、主机IP、服务器名称、CPU利用率、内存利用率、操作系统、内存、所属机房、更新时间、系统时差等。

1. 物理服务器

物理机监控面板主要用于对机房物理服务器的日常巡检管理，以杜绝设备运行隐患，例如：服务器双电源，其中一个电源坏了，另一个电源持续供电，CPU/GPU风扇有问题，运行温度过高等。

提供所有已接入的物理服务器的巡检视图，按照故障严重程度排序。

在面板中，设备以卡片形式加载呈现，每个卡片包含了温度、CPU、内存、风扇、磁盘、电源6大分类，数字分别体现每个分类下的设备或指标的正常/异常数量，以便巡检工程师能一目了然知道设备运行的总体情况。

按照故障的严重程度，分为严重（红色）、紧急（橙色）、预警（黄色）、正常（绿色）。

1. 数据库/中间件

数据库/中间件监控面板主要对自发现的各类数据库/中间件进程可用性、服务器资源占用情况、异常预警状态的集中展示，方便巡检人员及时识别运行风险，提前规避故障发生。

1. 网络/安全设备

网络/安全设设备监控主要用于监测设备本身的性能以及端口报文的传输流量，并对端口状态变迁、性能瓶颈、流量瓶颈提供预警，避免网络因素引起的业务稳定性问题。

* 指标管理

指标管理是监控平台基础字典的管理模块，用户维护监控项和指标集的基础库。

1. 监控项

主要对监控系统所能监控的各类监控实体类型进行了集中定义，这里按照大的分类且分成三个板块：操作系统、数据库/中间件、网络/安全设备。

“操作系统”类型的监控对象指操作系统层面的CPU、内存、网卡、磁盘存储、操作系统进程等监控目标对象。

“数据库/中间件”类型的监控对象指部署在操作系统上的组件，如tomcat、mysql、nginx、activemq等等。

“网络/安全设备”类型的监控对象指在网络拓扑中的交换机、路由器、防火墙等设备。

注：在监控系统部署时，会初始化一套标准的监控对象数据，目前已经覆盖了操作系统，常规的网络/安全设备和大多数常用的数据库中间件。

1. 指标集

指标集管理功能主要用于维护监控项的指标集，例如：我们需要关注内存这一监控对象的占用率，交换区使用比例，那么这两项就是内存这个对象的指标集。

* 模板管理

1. 阈值方案

阈值方案配置的目的旨在定义共性模板，减少工程师对每个监控对象实例分别配置的繁重工作量。

阈值方案模板按照“操作系统”、“数据库/中间件”、“网络/安全设备”三大类型分开来查询和维护。

阈值方案定义了3个要素，分别是“检查设置”、“监控规则”和“告警设置”，“检查设置”定义了采用什么脚本，多长时间巡检获取一次指标；“监控规则”定义了什么情况下认定为指标异常，触发告警；“告警设置”定义了在发生异常时，告警信息以何种方式推送给谁。

阈值方案定义好之后，新发现的对象实例，会自动采用对应的方案进行默认配置，如此一来在自动扩容增加主机时，就无需人工单独去配置。

1. IPMI模板

由于不同厂商型号的服务器主板结构差异较大，因此需要针对不同厂商和不同型号的服务器分别定义模板，用来映射主板的结构（例如A服务器有5个内存插槽，B服务器有6个内存插槽）。

硬件监控模板默认提供了“CPU”、“内存”、“磁盘”、“电源”、“温度”、“风扇”六个基本的分类维度。此外，支持用户根据不同品牌型号的硬件，增加特别需要关注的对象分组。

* 监控配置

1. 数据库/中间件

无论以哪种方式接入的数据库/中间件，都会采用对应的阈值方案生成实例配置。但在不同的服务器资源和性能场景下，默认阈值并不一定完全适用，这时就需要用户做适应性调整，并补充告警推送的接收人。

数据库/中间件实例配置支持单项配置，也支持批量配置。配置时可以单点调整一个或几个组件实例的采集周期、采集插件、阈值规则和告警通知策略。

1. 网络/安全设备

该模块主要针对网络/安全设备的运行指标、网络流量指标设置监控阈值规则，当设备本身运行过程中资源占用过高，或端口上下行流量占用大量带宽时，能及时发现并告知运维人员。

设备接入时，会根据默认的阈值配置方案，自动生成实例的配置，针对实例配置的更改主要针对单个的设备进行个性化调整。

配置阈值规则时，可以选择开启或禁用（只有开启时，才作相应的阈值规则分析，产生告警）。阈值规则目前支持两种形式，一种是固定阈值，一种是基于一段周期的指标计算的基线进行匹配（基线匹配主要计算这段周期的最大值、最小值和平均值，可配置大于最大值告警，小于最小值告警，和均值上下浮动？%告警）。

1. 物理服务器

该模块主要针对物理服务器各组成硬件的运行性能指标，设定阈值和收敛规则。当通过IPMI采集到设备性能数据时，依据预设的阈值规则进行分析，产生阈值告警。

和其它监控对象不同的是，物理服务器不会根据默认的阈值方案自动生成实例监控配置，需要手动添加。

##### 服务可用性监测

定义基于http(s)\ping\telnet访问方式的业务拨测任务，定期检查业务应用的接口可用性、页面返回内容的正确性。对于网络运维用户而言，可以通过ping定期监测网络节点的连通性、通过telnet定期监测端口的开通情况。

* 健康检查

1. 拨测面板

服务健康检查主要是基于业务拨测的结果反映服务实时可用的结果和既往一段时间内的离线时长。拨测面板即是对拨测任务的执行结果做集中化展示，常用于业务巡检。

拨测任务的结果按照平均可用率从低到高排序，可用率低的排在前面，便于用户快速识别到问题。列表中可用率为任务下多个拨测点可用率的平均值，平均延迟为任务下所有拨测点响应延迟的平均值，最大延迟为任务下所有拨测点响应延迟中的最大值，最小延迟为任务下所有拨测点响应延迟中的最小值。

支持查看每个拨测点的执行细节：其中可用率为一段周期内可用的拨测次数/总拨测次数，平均延迟为这段周期内所有拨测延迟时间的平均值，最大延迟为这段周期内拨测最长的一次耗时，最小延迟为这段周期内拨测最短的一次耗时。

在执行历史中，可以查看拨测执行的优劣化趋势情况：从中可以看到指标劣化趋势，在折线上同时标注了异常发生的时间点，对应到下方的异常列表，也反映了异常发生的详情，包括恢复时间。

1. 故障统计

主要用于统计拨测节点既往一段时间发生的故障数和总离线时长，以便运维巡检人员重点关注相关节点的网络环境，设备运行环境等周边因素。

* 拨测设置

1. 拨测脚本

支持根据业务需要自定义业务拨测脚本，在脚本中定义个性化的实现逻辑和返回结果，例如：将总响应时间拆分成请求准备时间、SSL时间、DNS时间等。

一个脚本可以用到多个不同的拨测任务中，新上传的脚本在调试期间支持在线编辑，即时生效。并能根据脚本查询关联的拨测任务。

1. 拨测任务管理

功能的业务目标为定义拨测任务，设置拨测的目标地址，监控点（支持多点拨测），告警触发条件以及告警推送策略。

任务维护支持单条创建，也支持excel模板批量导入。

##### 数据可视化

* 可视化仪表板

平台默认提供了三套宏观数据分析模板，分别为“资源态势”、“监控告警”、“运维效能”。其中，“资源态势”视图体现服务器资源的总体使用情况，是否均衡，产生的资源超负荷告警情况如何；“监控告警”则是从异常告警的视角，统计告警的来源分布，滞留情况，影响面；“运维效能”重点体现MTTR/MTTA等体现运维执行效能的宏观指标。

* 放映室

放映室模块主要将多个仪表盘排序编排，在投屏展示时，实现自动轮播切换。放映室播放时，能实现多个仪表盘视图的过渡切换。

##### 告警分析

告警分析支持对跨来源的异常事件进行去重、压缩、抑制，然后依据预定的策略可自动转单派发给工程师。

* 接入概览

告警分析组件提供事件接入的标准化接口，监控子平台主动调用接口，将异常事件推送给告警分析组件。接入概览视图对接入的异常事件进行相似性分析，识别哪些故障频繁重复发生，相关性事件占比如何。

【从接入来源维度】：接入了多少个监控子系统，是否存在跨系统发现的同一异常事件。

【从事件的维度】：有多少重复发生的事件，覆盖哪些场景。

【压缩比统计】：原始事件有多少，经过告警分析收敛后，压缩到多少。

视图中，会体现一段统计周期内，所有被告警分析组件接收的异常事件，并且提供相似性事件比对（主要通过设备和异常场景进行匹配），便于运维人员识别频繁发生的重复事件，重点予以应对。

* 告警设置

1. 收敛策略

异常事件接入到告警分析模块之后，需要进行去重、收敛和抑制。“收敛策略”功能就是对事件的分析处理过程的算法策略进行预定义。

过滤策略预定义：定义过滤条件，启用后会按照过滤条件，将不关注的异常事件过滤掉（这些过滤掉的数据将不被拿来作后续分析处理）；

压缩规则预定义：定义事件的收敛策略；

处理规则预定义：定义事件压缩后，是否自动转单。

1. 推送方式

异常事件经过压缩收敛后，信息需要推送给相关的运维工程师，推送方式主要有三种：

1. 通过短信、微信、邮件等第三方通道推送；
2. 通过websocket方式推送给WEB终端，在终端弹框，播放告警音频；

推送策略可以指定多个，逻辑是按照告警级别、内容或IP地址，匹配到不同的运维负责人。

* 告警列表

异常事件通过策略定义的收敛算法合并，产生一条一条的告警数据，在“告警列表”中提供结果查询。

告警信息可以展开查询压缩前的原始异常事件。

#### 数据运营平台

提供各类引擎的运营指标监控和基础报表展示功能。

##### 标准指标集定义

根据引擎能力和业务要求，定义各类引擎的标准指标集，如能力调用次数、转写字数等，另外，根据运营需要，定义运营指标的统计维度，如时间段、单位、区域、用户、应用。

目前定义的标准指标集如下：

| 序号 | 能力 | 指标名称 |
| --- | --- | --- |
|  | 语音预处理 | 总调用次数 |
|  | 方言语种识别 | 总调用次数 |
|  | 离线语音转写 | 总调用次数 |
|  | 处理的文件时长 |
|  | 总转写字数 |
|  | 实时语音转写 | 总调用次数 |
|  | 总使用时长 |
|  | 转写字数 |
|  | 语音合成引擎 | 总调用次数 |
|  | 合成字数 |
|  | 通用OCR | 总调用次数 |
|  | 卷宗OCR | 总调用次数 |
|  | 票据OCR | 总调用次数 |
|  | 中文分词 | 总调用次数 |
|  | 命名实体词识别 | 总调用次数 |
|  | 要素抽取 | 总调用次数 |
|  | 关系抽取 | 总调用次数 |
|  | 文本分类 | 总调用次数 |
|  | 中英互译 | 总调用次数 |
|  | 总翻译字数 |
|  | 维汉互译 | 总调用次数 |
|  | 总翻译字数 |

##### 指标埋点及标准接口

为达到运营指标监控的目的，需对各类引擎进行指标埋点，并通过标准接口传入运营平台，以便在引擎能力调用时，获取相关指标值以及进行统计。

##### 基础报表展示

在获取指标值并根据相关维度进行统计后，会生成基础报表，方便用户更直观的了解平台能力被调用的情况。

### 感知智能服务系统

#### 语音预处理引擎

语音预处理引擎能够对原始语音进行处理，使处理后的信息更能反映语音的本质特征，为后续的识别工作提供准确信息。

针对实际应用中的场景类型，语音预处理预计包含能量四门限算法、基于规则的噪声判断算法及基于模型分类器判决等步骤。这三个步骤的复杂度依次提升，分别针对不同的音频场景类型进行检出，最终实现不同的音频场景分割并检测出其中的纯语音片段。具体算法描述如下：

1. 能量四门限算法，通过“静寂状态”、“语音起始状态”、“语音稳定状态”及“语音衰减状态”四个状态的定义和设置，并分别设置相应的状态间跳转所需四个能量门限值，根据当前音频中每一帧的能量信息，实现四个状态之间的跳转最终实现音频中能量较高的语音片段的检测。
2. 基于规则的场景判断算法，该算法主要对利用音频的频段能量，对通过了能量四门限算法的信号片段，进行是否是音乐、噪音等场景进行初判。
3. 通过基于规则的初步检测，可以将多数的场景检测出来，但是考虑到具体环境的需求，每一套系统的特定应用环境下，场景情况均不同。因此，最终音频场景识别模块还经过模型分类器判决算法得到最终的有效语音。该方法可以根据实际系统的应用环境，训练出与实际应用场景中所出现各种场景相匹配的模型，进而有效的判断。在训练过程中，本系统引入区分性训练（Discrimination Training），并使用最小分类误差准则使训练后的语音模型与其它几种非语音模型更具有区分性，提高场景分辨的效果。

语音预处理引擎主要功能包括如下功能：

1. 语音有效性检测

对语音文件中音频的截幅、背景噪声、有效时长进行检测。

1. 语音活动检测

对输入的音频流进行分析，确定用户说话的起始和终止的处理过程。一旦检测到用户开始说话，语音开始流向识别引擎，直到检测到用户说话结束。这种方式使识别引擎在用户在说话的同时即开始进行识别处理。

1. 按键音检测

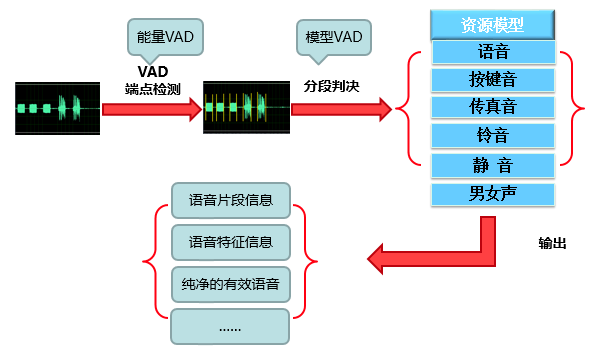
识别语音信号中的多音双频信号(DTMF或者按键音)，并输出对应语音片段。

1. 说话人分离

针对一段音频多人说话语音进行分离，解决多说话人分离难题，实现多角色自动区分，可以准确区分每一位的说话内容。

1. 男女声鉴别

通过对男性和女性声音的语音特征的分析，撷取说话者的声音特征，藉此判断说话者的性别是男性或女性。



Vspp处理过程

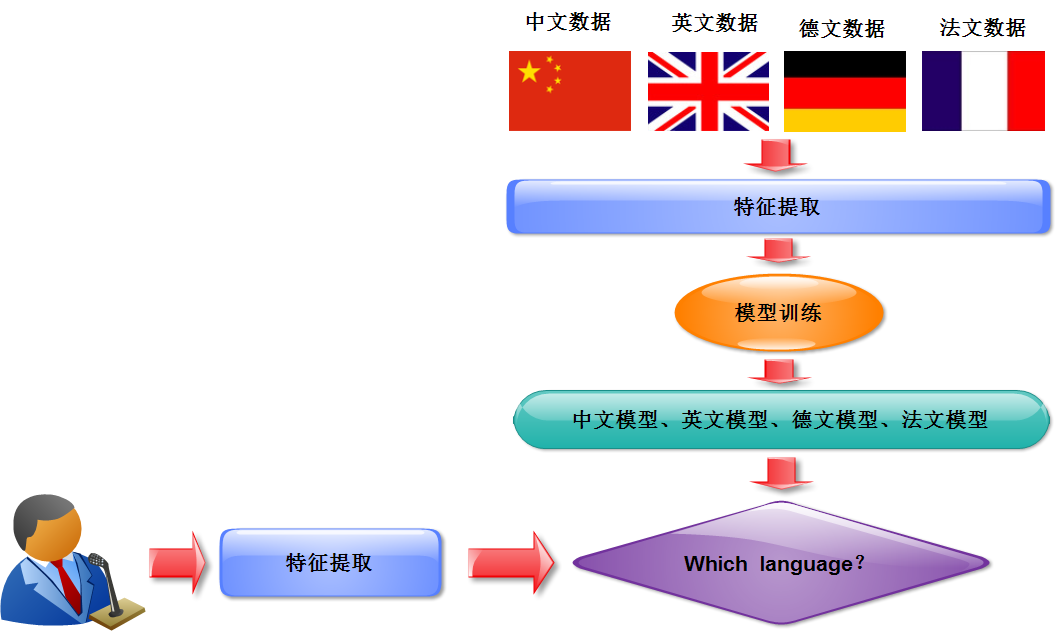
主要应用在一些需要对语音进行各类处理的场景，能够有效地对语音信息进行分析，使用户快速获取有用信息，从而广泛应用于语种识别、声纹识别、语音识别等相关评测系统中。通过提前判断输入语音相关信息，以提升后端系统处理性能。

#### 语种识别引擎

语种识别技术用于识别出一段语音是何种语言，包括不同民族语言。如英语、法语、德语、汉语、藏语、维语、蒙语等。也可鉴别汉语中的方言，如普通话、粤语等。

方言语种识别引擎具备较高的准确率及处理速度，取决于采用了如下核心算法：

1. 基于区分性训练的语种识别性能提升：传统GMM(高斯混合模型)技术路线的语种识别性能已基本满足常规语种识别任务，针对网络音视频数据信道和背景环境复杂的情况，研发基于噪声鲁棒性的PNCC (Power-Normalized Cepstral Coefﬁcients)特征的语种识别系统；为进步增加各语种模型之间的区分性，采用MMI(Maximum Mutual Information，最大互信息)准则就是进行区分性训练的准则之一对各语种的GMM(高斯混合模型)模型进行区分性训练；同时，采用SVM（支持向量机）建模，有效地避免经典学习方法中过学习、维数灾难、局部极小等传统分类存在的问题。通过以上技术研发，提升语种识别模型的噪声鲁棒性和模型区分能力，确保复杂数据环境下的语种识别性能稳定。
2. 高性能语种识别引擎。在算法层面上，主要采用高斯聚类策略，在不降低声学模型维度的基础上，有效降低运算的高斯数，从而降低运算；采用灵活的跳帧策略，在充分保证数据量，以及效果的前提下，减少有效帧的使用，从而提高系统的效率；在工程化方面，采用SIMD(Single Instruction Multiple Data，单指令多数据流)技术实现相同之机器时间内多计算指令的完成，合理架构引擎数据结构和特征、模型维度最大化SIMD效果；采用intel MKL（Math Kernel Library）高性能数据库，提高复杂矩阵的求逆运算的效率；采用VTune分析工具分析代码中的热点部分，然后通过数据结构和引擎架构优化，有效提高二级缓存cache L2命中率，从而提高CPU运算效率。通过以上策略，在保证语种识别引擎核心效果的同时，大幅提高引擎的效率，确保引擎的效率满足海量数据处理应用要求。



方言语种识别引擎在实际工作中，用户可基于引擎识别的结果，有针对性的对目标关注的语种数据进行分类或引流或精细化的分析，将会大大提升服务器资源的利用率，缩小业务处理的数据范围。

具体包括如下功能：

1. 支持语种种类

|  |  |
| --- | --- |
| 语种名称 | 英文缩写 |
| 中文 | chin |
| 英语 | engl |
| 日语 | japa |
| 韩语 | kore |
| 法语 | fren |
| 俄语 | russ |
| 藏语 | tibe |
| 维语 | uyhu |
| 越南语 | viet |

1. 支持方言种类

|  |  |
| --- | --- |
| 语种名称 | 英文缩写 |
| 粤语 | cant |

语种识别主要作为转写、翻译的前置处理步骤使用，通过语种分类将语音分发给对应引擎或特定听音人进行处理。

#### 实时语音转写引擎

##### 中文实时语音转写

中文实时语音转写基于深度全序列卷积神经网络框架，通过 WebSocket 协议，建立应用与语言转写核心引擎的长连接，实时输入音频流，将音频流数据实时转换成文字流数据结果。

语音识别是把用户说的任意语音转换成对应的文字信息，包括训练和解码两个过程。



图 语音识别技术

训练：基于语音、语言学、信号处理技术、数据挖掘技术、统计学等领域知识，使用业界领先的RNN技术进行声学建模，并基于海量且话题广泛的文本数据进行超大规模语言模型训练，生成声学模型和语言模型。

解码：针对输入的语音数据，在进行端点检测、降噪后，提取声学，在训练好的声学模型和语言模型资源基础上，使用解码器进行解码，得到最适合的结果。

目前，基于深度神经网络（DNN）声学建模的识别系统构建已经成为大词汇量连续语音识别的主流技术方案。在现有语音识别技术及数据的积累基础上，针对语音识别技术进行大力研发，通过全方位的深度神经网络技术框架完善、超大规模语言模型的快速更新技术、支持个性化词典及超大规模语言模型的解码器技术、以及配套的海量多语种语音和语言数据资源库构建，最终完成覆盖中英等的多语言的语音内容识别系统。主要核心研发工作包括：

（1）基于深度神经网络的声学建模技术：本项目拟首先完成基于串级别区分性准则的DNN训练新方法，并重点关注在大语种数据扩展到海量级别（数万小时级别）时DNN区分性训练的效果，并通过并行运算等方式解决海量数据下DNN区分性训练的效率问题；其次，开展基于DNN框架的前后端鲁棒性技术，通过设计DNN的结构用来表征干净语音、噪声和带噪语音之间的复杂关系，以期达到最佳降噪/变换的效果，充分发挥出DNN的非线性建模能力用于表征干净语音和带噪语音之间的复杂关系，提升语音识别系统的鲁棒性；

（2）DNN框架下针对口音方言的自适应技术：在传统GMM-HMM框架下，基于特定说话人和口音方言的少许自适应数据往往能显著提升识别效果，而在DNN框架下如何实现声学建模技术的自适应是本项目拟开展的技术研发工作。具体的，通过引入可变长度的码字层而改变传统DNN的结构、在自适应更新的过程中引入对模型参数变化的限制、以及将自适应过程中的目标函数从帧级别的信息熵扩展到串级别的区分性准则等策略，实现DNN框架下的自适应技术；

（3）超大规模语言模型的快速更新技术：语言模型的快速更新技术是保障语音识别系统能够及时有效识别出最新热词的关键技术。通过语言模型WFST网络动态并联加载的方式来改善该问题。首先通过实时抓取大量最新网页、微博的方法，完成小语言模型的构建及相应WFST网络构建，并通过将该小语言模型网络和更新相对较慢的通用超大规模语言模型网络直接进行并联并调节相关权重，实现超大规模语言模型的快速更新。

将实时语音转换成文字的在线模式。为了使转换后的文本更易读和理解，提供了文本数字规整、加标点、文本顺滑等功能。具体如下：

1. 支持将中文语音流实时转写为中文文字；
2. 多种格式支持：支持16k16bit pcm语音，支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 说话人分离：支持说话人角色分离
4. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
5. 文本顺滑：主要将识别结果文本中将一些不合理的语气词替换。
6. 智能标点：根据识别结果给文本内容加上标点符号。
7. 数字规整：将识别结果中的中文数字更换成合理的阿拉伯数字。
8. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
9. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预
10. 多语言混说：支持中英混说、中粤混说

##### 英文实时语音转写

英文实时语音转写基于深度全序列卷积神经网络框架，通过 WebSocket 协议，建立应用与语言转写核心引擎的长连接，实时输入音频流，将音频流数据实时转换成文字流数据结果。

具体功能如下：

1. 支持将英文语音流实时转写为英文文字；
2. 多种格式支持：支持16k16bit pcm语音，支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 说话人分离：支持说话人角色分离
4. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
5. 文本顺滑：主要将识别结果文本中将一些不合理的语气词替换。
6. 智能标点：根据识别结果给文本内容加上标点符号。
7. 数字规整：将识别结果中的中文数字更换成合理的阿拉伯数字。
8. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
9. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预

##### 维语实时语音转写

维吾尔语是典型的黏着语言，在构词和语法构成上不同于孤立语言。因此，维语实时语音转写系统在遵循语音识别技术框架的基础上，针对维语的语言特点进行了声学模型和语言模型方面的研究。

在声学模型方面进行了声学模型单元选择、问题集设计、区分性声学特征、声学模型区分性训练等关键技术。

1. 声学模型单元选择：使用维语34个 (没有考虑少见的长元音)音素作为声学模型基本单元，其中包括静音模型和停顿模型。考虑连续语音中的音变现象，本项目使用三音素模型，为了解决三音素在训练集上没有出现的问题，采用基于最大似然决策树的状态共享策略。
2. 问题集设计：决策树的节点的分裂依赖于问题集，问题集设计的关键是利用发音方式的相似性来定义划分特征。维吾尔语是字形与音位相互对应的语言，因此本项目利用维吾尔语中音素在该语言语音系统里的格局对维语音素定义设计问题集。
3. 区分性声学特征：在美尔倒谱系数的特征基础上使用TANDEM后验概率特征，利用神经网络所具有的强大的模式辨识分类能力提高特征之间的区分性。
4. 声学模型区分性训练：在使用极大似然准则训练出声学模型后， 为了提高维语语声学模型之间的区分能力，采用基于最小音素错误准则的区分性训练进一步提高识别系统的性能。

另外，在语言模型方面，研究了维语的语法单元，使用最小描述长度方法将建模单元分解成单词、音节、词干词缀和词干词尾，这些作为语言模型的基本建模单元。研究了维语语言模型的训练方法，首先，对生文本语料进行分句处理；其次，剔除句子中的英文和中文单词，将数字、重量和长度等符号转换成相应的维吾尔语文本；再次，将阿拉伯文式维语文本语料转换成拉丁式的维语；最后，进行语法单元分解，并使用常用的语言模型训练工具进行语言模型训练。

具体功能如下：

1. 支持将维语语音文件转写为文字；
2. 多种格式支持：支持8k16bit pcm 、支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 大语音支持：支持大语音文件进行转写(支持10h)；
4. 说话人分离：支持说话人角色分离（两人）。
5. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
6. 智能标点：根据识别结果给文本内容加上标点符号。
7. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
8. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预

#### 非实时语音转写引擎

##### 中文非实时语音转写

非实时语音识别引擎将预先录制完毕的完整音频文件传输至云端，转写服务处理完成后将输出此音频对应的完整文字结果。

语音转文本识别对语音依次进行语音检出、说话人分离、解码等处理，采用超大规模声学模型和语言模型进行解码。通过多遍解码，提高识别的正确率。解码完毕后，还需要进行重要的后处理，针对识别结果中可能出现的错误通过词后验概率(WPP)等技术给出识别结果对应的可信度得分，对可信度低的识别结果进行智能拒识。最后再进行断句，分段等处理，把识别结果变成可阅读的文字内容。

语音转文本主要流程如下图：



语音转文本技术框图

语音分割模块：针对输入的整段未知电话语音，首先使用基于能量及模型的VAD技术进行分段，再采用基于模型的聚类技术对原始电话语音中的人声、非语音、音乐、背景噪音等现象进行自动判别及归类，再进一步针对不同说话人进行聚类，为后续的语音识别做好准备。

前端特征提取模块：针对分割后的语音提取对语音识别有效的、对噪声鲁棒性较好的声学特征，例如常用的MFCC/PLP特征，对于带调的中文语音识别还需提取Pitch等特征，并针对信道等因素进行初步的CMN/MSVN等规整。

区分性特征提取：除了使用传统的MFCC等谱特征以外，基于神经网络的TANDEM区分性特征是当今语音识别系统中的热点之一，可以对传统仅基于谱特征的语音识别性能起到很好的补充作用。

特征规整模块：为了消除不同说话人对识别效果的影响，我们首先会从特征域尽可能去除说话人的信息，主流的特征域处理方式包括VTLN、HLDA、特征域SAT等技术，通过这些关键技术的加入，我们可以获得冗余信息更少、说话人信息被去除的声学特征。

多遍解码框架：为了将各语音转写关键技术点进行更好的整合，并综合考虑识别效果和识别效率等因素，我们在系统中将使用如上图所示的多编解码框架：首先使用简单声学模型(模型复杂度较小、未去除说话人信息)和低阶语言模型(一般为三阶且模型复杂度较小)，并结合WFST解码器进行一遍的快速解码，并生成对应的词图空间，该词图空间具有覆盖性好、结构紧致的特点。接着，我们将绝大多数对性能提升的核心技术点放到二遍解码中，包括特征规整、区分性特征、模型区分性训练、模型自适应等，并在二遍解码中相应使用更加复杂的声学模型和超大规模高阶语言模型，在词图空间中采用重新打分的方式输出二遍解码结果。虽然二遍解码融入了各关键技术点，且模型复杂度相对于一编解码提升很多，但是由于基于该词图空间的搜索范围小了很多，所以二遍解码的效率非常高。基于上述多编解码框架，我们不仅能将各转写技术点的性能提升发挥出来，在效率上还有巨大的优势。

声学模型训练：为了配合上述多遍解码框架和各特征域技术点的加入，需训练对应的声学模型，并相应加入当前主流语音识别系统中的关键技术点区分性训练以持续提高识别性能，在此过程中对区分性训练的准则和参数优化算法进行研究。除此之外，如果针对海量训练数据(包括有监督的和非监督的)不断提升训练效率，并在声学模型中充分发挥海量数据对说话人、口音、方言等的覆盖效果，也是我们研究的重点。

语言模型训练：基于海量且话题广泛的文本数据进行超大规模语言模型训练，并进行语言模型区分性训练以及基于类别的语言模型自适应研究。

置信度判决：连续语音识别系统中重要的后处理模块，针对识别结果中可能出现的错误通过词后验概率(WPP)等技术给出识别结果对应的可信度得分，对可信度低的识别结果进行智能拒识。

主要功能如下：

1. 支持将中文语音文件转写为文字；
2. 多种格式支持：支持8k16bit pcm 、支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 大语音支持：支持大语音文件进行转写(支持10h)；
4. 说话人分离：支持说话人角色分离（两人）。
5. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
6. 文本顺滑：主要将识别结果文本中将一些不合理的语气词替换。
7. 智能标点：根据识别结果给文本内容加上标点符号。
8. 数字规整：将识别结果中的中文数字更换成合理的阿拉伯数字。
9. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
10. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预

##### 英文非实时语音转写

非实时语音识别引擎将预先录制完毕的完整音频文件传输至云端，转写服务处理完成后将输出此音频对应的完整文字结果。

主要功能如下：

1. 支持将英文语音文件转写为文字；
2. 多种格式支持：支持8k16bit pcm 、支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 大语音支持：支持大语音文件进行转写(支持10h)；
4. 说话人分离：支持说话人角色分离（两人）。
5. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
6. 文本顺滑：主要将识别结果文本中将一些不合理的语气词替换。
7. 智能标点：根据识别结果给文本内容加上标点符号。
8. 数字规整：将识别结果中的中文数字更换成合理的阿拉伯数字。
9. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
10. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预

##### 维语非实时语音转写

非实时语音识别引擎将预先录制完毕的完整音频文件传输至云端，转写服务处理完成后将输出此音频对应的完整文字结果。

主要功能如下：

1. 支持将维语语音文件转写为文字；
2. 多种格式支持：支持8k16bit pcm 、支持带语音头的mp3、wav、wma、m4a语音格式转码；
3. 大语音支持：支持大语音文件进行转写(支持10h)；
4. 说话人分离：支持说话人角色分离（两人）。
5. 智能分句：对转写文本按语义进行子句划分，并在子句之间加注标点
6. 智能标点：根据识别结果给文本内容加上标点符号。
7. 替换列表：将识别结果中的某些指定文字替换成列表中映射的文字。
8. 热词及小包优化：支持短平快的效果优化方案，对人名、领域词汇进行快速的人工干预

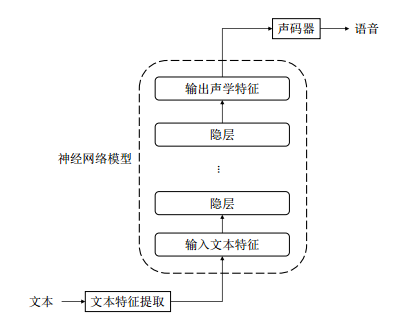
#### 语音合成引擎

##### 中文语音合成引擎

语音合成利用声学、语言学、数字信号处理、计算机科学等多个学科技术，将文字信息转化为可听的声音信息。

目前业界普遍采用基于 HMM 的统计参数语音合成方法进行模型训练及合成，经过二十多年的发展，已经成为一种成熟的语音合成方法。然而，基于HMM的统计参数语音合成方法因声码器对音质的损失较大，声学建模精度不够，合成语音过于平滑，导致听感平淡，音质受损。

基于HMM建模存在的问题，讯飞采用基于神经网络的声学建模方法，抛弃声码器，使用波形的历史信息，逐点预测下采样点，大幅提升音质。基于神经网络的声学建模方法架构如下：



基于神经网络的统计参数语音合成建模方法框架图

图中输入特征为从文本中提取的特征，即用离散或连续的数值特征来描述文本，包括：二值的分类特征，例如音素的 ID、音素的类别、上下文的音素类别、清浊信息等；实数值特征，例如当前音节中音素的个数、当前音段中字的个数、当前帧在音素中的位置等。输出特征为声学参数，包括基频参数，频谱参数，清浊信息。由于基频的离散特性，在建模前，通常需要对其进行插值，得到连续的基频轨迹。

基于神经网络的统计参数语音合成系统训练采用 MMSE 训练准则，采用 BP 算法和 SGD 算法更新模型参数，使预测的声学参数和自然声学参数尽可能接近。在合成阶段，首先对待合成文本提取文本特征，然后使用神经网络预测对应的声学参数，最后经过声码器合成出语音波形。

具体功能如下：

1. 支持将中文文字转化为自然流畅的人声。
2. 高质量语音：将输入文本实时转换为流畅、清晰、自然和具有表现力的语音数据。
3. 多音色服务：提供丰富、风格多样化的音色选择，用户可供根据不同应用业务需要，选择最适合应用场景的语音风格。
4. 高精度文本分析技术：保证了对文本中未登录词（如地名）、多音字、特殊符号（如标点、数字）、韵律短语等智能分析和处理。
5. 字符集支持：支持输入UTF-8字符集。
6. 数据输出格式：支持输出Wav格式语音数据。
7. 参数动态调整：开发接口提供了音量、语速、音高（基频）等多种合成参数动态调整功能。
8. 多发音人：标准包为晓燕，晓峰。

##### 英文语音合成引擎

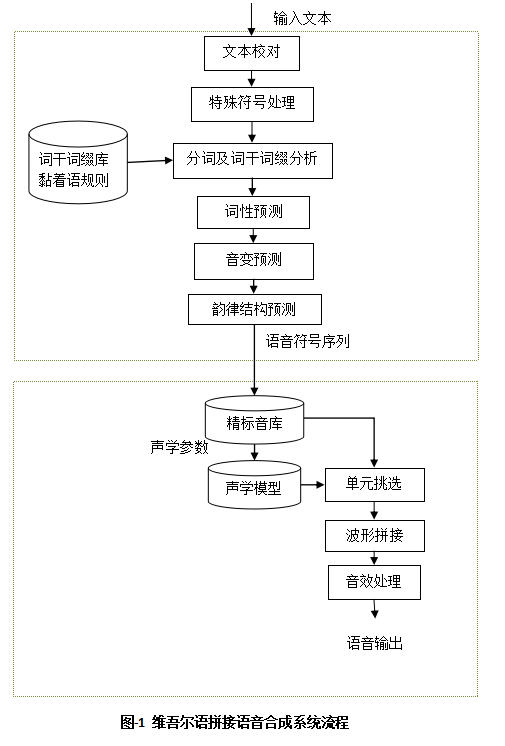
英语语音合成引擎将英语文字转换为自然流畅的人声。

具体功能如下：

1. 支持将英文文字转化为自然流畅的人声。
2. 高质量语音：将输入文本实时转换为流畅、清晰、自然和具有表现力的语音数据。
3. 多音色服务：提供丰富、风格多样化的音色选择，用户可供根据不同应用业务需要，选择最适合应用场景的语音风格。
4. 高精度文本分析技术：保证了对文本中未登录词（如地名）、多音字、特殊符号（如标点、数字）、韵律短语等智能分析和处理。
5. 字符集支持：支持输入UTF-8字符集。
6. 数据输出格式：支持输出Wav格式语音数据。
7. 参数动态调整：开发接口提供了音量、语速、音高（基频）等多种合成参数动态调整功能。

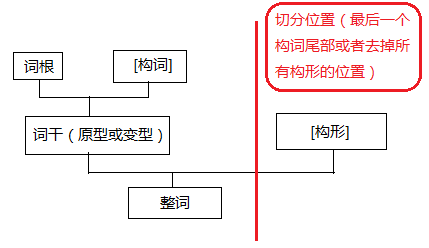
##### 维语语音合成引擎

维语语音合成引擎将维语文字转换为自然流畅的人声，主要包括维吾尔语文本校对、文本规整、特殊符号处理、词干词缀分析、音变预测、韵律结构预测、语料库与音库设计、参数生成、声学模型等方面。



文本校对及规整: 由于维吾尔语文字存在一定程度的不规范，需要将不符合拼写规则的文本内容进行自动校对。维吾尔语文本中还会出现包括阿拉伯数字的特殊符号、年月日等特殊形式，需要根据不同的应用领域，将它们转化成标准维吾尔文。

分词与词性预测:维吾尔分词是以空格为标志分隔的，作为黏着语，维吾尔语单词结构是在词干后黏合多个词缀构成的。词干词缀切分是黏着语词法分析的主要内容，目的是服务于语音合成韵律结构分析、建立识别翻译的语言模型。词干词缀切分是基于词干词缀库自动切分算法是TriGram 的全局解码，并加以规则。词干词缀切分之后，利用决策树算法预测子词词性。



**图 -2 词干词缀定义**

**字变模型:**人们普遍认为维吾尔语是拼音语言，字符和读音一一对应，然而实际语音会发生音变。为了合成自然的语音，需要增加音变处理。通过C45决策树算法形成预测模型。

**韵律层级预测:**确定每个语音单元对应的韵律信息，及停顿的位置及边界调，相关参数包括基频、时长、能量、停顿、重音等。

**维吾尔语语音数据库制作：**维吾尔语语音库是语音合成的数据基础，在专业录音室采集专业发音人的语音数据。对录制切分好的语音句子用语音标注分析软件Praat来进行标注，在自动文本分析的基础上，具有语音学经验的标注人员人工标注录音数据，将录音中的字词边界、韵律层次、重读弱化、发音变调等关键信息标出。形成具有索引标记的基本拼接单元语音数据库。

**声学参数提取：**一般以固定窗长（如25ms）固定帧移（如5ms）提取帧级声学参数，包括频谱参数和基频参数。频谱参数表征语音的音段特征，一般采用倒谱参数（Cepstrum）或者线谱频率（LSF）；基频参数即浊音信号的周期震动频率，简称F0。时长参数作为标记数据的一种，一般通过人工标注或模型自动切分得到。

**声学模型：**统计声学模型库是语音合成的核心资源，用于指导拼接单元选择。以隐马尔可夫模型(HMM)为代表的统计声学建模方法，得到了最为充分的发展，模型训练过程中，首先将带标注的音库进过声学参数提取（包括频谱、基频、时长），形成参数化的数据集合。以帧级声学参数为基础，对基本建模单元（如音素）做上下文属性扩展的，从左到右状态遍历拓扑结构的HMM统计建模， 生成统计声学模型。常用的声学统计模型包括，频谱模型、基频模型、时长模型、拼接模型。声学模型以音素为建模单位，对帧级提取的静态、一阶动态、二阶动态频谱参数（参数维数可根据需求设定）统计建模，构建上下文相关的，五状态从左到右各态历经的拓扑形式HMM模型，每个状态的观测概率密度函数使用单高斯分布来表示。并对其依据所在的上下文环境进行模型扩展，使用基于最大似然准则的Baum-Welch算法进行模型训练。

**单元挑选：**基于挑选算法，搜索最佳的一系列波形单元组合，串联拼接成相应语音。拼接代价模型需要同时考虑分为频谱拼接代价模型和基频拼接代价模型两种，是为了根据衡量单元挑选波形拼接语音合成时两个相邻两个单元之间在拼接处的声学特征的过渡情况决策单元的挑选。使用的特征为训练数据中各拼接点，前一音素语音单元最后一帧的频谱/基频值与后接音素语音单元的第一帧的频谱/基频值的差值。对于基频特征，由于存在清浊音的问题，同样使用多空间概率分布来度量拼接处基频的变化情况。

传统的单元挑选一般基于专家规则和模板匹配技术，而基于统计声学模型的单元挑选技术是讯飞的核心算法。其利用声学统计模型作为单元挑选指导目标，使其具有了统计学意义上的可靠目标模型。实际挑选算法中，首先使用Kullback-Leibler距离（KLD）的进行快速单元预选，将全语音库的全体单元降低到若干备选单元规模，再基于最大似然值准则的，计算备选单元与声学模型的目标代价进行单元挑选，最后通过评价相临单元直接与拼接模型的连接代价，进行Viterbi搜索得到最佳的单元序列串组合。基于统计声学模型的单元挑选技术，降低了传统单元挑选对精标数据和专家调整的依赖，使用统一的概率统计准则指导单元挑选，大大提高了语音合成的稳定性和语音自然度。用该技术的合成系统在连续6届Blizzard Challenge英文语音合成国际比赛中蝉联第一。

**波形拼接合成：**样本单元拼接技术将搜索得到的波形单元序列串拼合起来，通过对前后单元搜索最大相关位置实现相位对齐，通过加窗叠加等平滑算法，保证波形单元之间的顺畅过渡，形成自然流畅的完整合成语音波形信号。为了确保拼接合成的稳定性，拼接单元挑选时将参考声学模型。声学模型的建立依赖于对带有标注信息的语音数据进行声学参数提取，并进行上下文属性扩展和模型训练，由于需要考虑的上下文属性较多，比如前后音素环境、前后声调、韵律位置信息等，从而导致上下文属性的组合数目成指数增长，从而产生模型对数据的过拟合现象。因此需要通过基于决策树的聚类算法，对上下文相关模型进行聚类，使得聚类后的模型有足够的训练数据。在合成预测之时，将通过基于上下文的决策树判断，保证集外文法可以被决策到稳定有效的描述模型。频谱、基频、时长、拼接模型都将通过构建决策树架构保证数据训练的有效和结果预测的稳定。

具体功能如下：

1. 支持将维吾尔文字转化为自然流畅的人声。
2. 高质量语音：将输入文本实时转换为流畅、清晰、自然和具有表现力的语音数据。
3. 多音色服务：提供丰富、风格多样化的音色选择，用户可供根据不同应用业务需要，选择最适合应用场景的语音风格。
4. 高精度文本分析技术：保证了对文本中未登录词（如地名）、多音字、特殊符号（如标点、数字）、韵律短语等智能分析和处理。
5. 字符集支持：支持输入UTF-8字符集。
6. 数据输出格式：支持输出Wav格式语音数据。
7. 参数动态调整：开发接口提供了音量、语速、音高（基频）等多种合成参数动态调整功能。

#### OCR识别引擎

##### 通用OCR识别引擎

OCR文字识别，指利用OCR （Optical Character Recognition，光学字符识别）技术，将图片、照片上的文字内容，直接转换为可编辑的文本。

1. 图像输入、预处理

图像输入：对于不同的图像格式，有着不同的存储格式，不同的压缩方式。预处理：主要包括二值化，噪声去除，倾斜较正等

1. 二值化

对摄像头拍摄的图片，大多数是彩色图像，彩色图像所含信息量巨大，对于图片的内容，我们可以简单的分为前景与背景，为了让计算机更快的，更好的识别文字，我们需要先对彩色图进行处理，使图片只前景信息与背景信息，可以简单的定义前景信息为黑色，背景信息为白色，这就是二值化图了。

1. 噪点去除

对于不同的文档，我们对噪点的定义可以不同，根据噪点的特征进行去噪，就叫做噪点去除。

1. 倾斜较正

由于一般用户，在拍照文档时，都比较随意，因此拍照出来的图片不可避免的产生倾斜，这就需要文字识别软件进行较正。

1. 版面分析

将文档图片分段落，分行的过程就叫做版面分析，由于实际文档的多样性，复杂性，因此，目前还没有一个固定的，最优的切割模型。

1. 字符切割

由于拍照条件的限制，经常造成字符粘连，断笔，因此极大限制了识别系统的性能，这就需要文字识别软件有字符切割功能。

1. 字符识别

这一研究，已经是很早的事情了，比较早有模板匹配，后来以特征提取为主，由于文字的位移，笔画的粗细，断笔，粘连，旋转等因素的影响，极大影响特征的提取的难度。

1. 版面恢复

人们希望识别后的文字，仍然像原文档图片那样排列着，段落不变，位置不变，顺序不变，的输出到word文档,pdf文档等，这一过程就叫做版面恢复。

1. 后处理、校对

根据特定的语言上下文的关系，对识别结果进行较正

OCR识别服务的功能模块包括：

1. 通用手写体识别

支持用户上传的手写体图像，返回识别出的中文、英文、中英文混合等手写文字。

注意：支持文字与水平轴小于±15°夹角偏移，要求文字的最短边最小不低于20像素，最大不超过4096像素，图片大小不超过4MB，暂不支持具有书写时序的字迹识别。

1. 通用印刷体识别

支持多场景、任意版面下印刷体图片或则实拍照片的文字识别，返回图片中文字位置与文字内容，识别的内容包括中文(简体和繁体)、英等语言。

注意：要求文档图片最短边最小不低于20像素，最大不超过4096像素，图片大小不超过4MB，支持文字与水平轴小于±15°夹角偏移。

1. 表格识别功能

通用图文识别支持表格识别，能够将表格识别信息进行格式化展示，并且可以搭配表格恢复引擎将识别的表格结果转换为Excel文件。

1. 单字位置显示

支持识别结果单字位置显示，并且可以对单字位置功能进行开关处理。

1. 支持图片格式

支持多种格式图片识别，目前通用模型ocr识别算法可支持PNG、JPG、JPEG、BMP格式图片识别。

##### 卷宗OCR识别引擎

针对刑事、民事场景下各类卷宗识别，支持整页与单行识别，支持图片分类、证据检出、朝向检测等功能，能够选择单行识别以及单字识别等模式。

卷宗图文识别主要训练材料是卷宗类图片，适合卷宗场景的扫描图片识别（一般为A4纸大小），能够对印刷体、手写体进行识别，并能对签章手印进行位置检出。

主要功能如下：

1. 图像分类

能够对国税地税增值税发票、身份证、邮政EMS快递单、营业执照、律师执业证、户口簿内页、房地产权证、组织机构代码证、结婚证、房屋平面图、法院诉讼费票据、医疗收费收据、机构登记信息、卷宗封面、机动车驾驶证、行驶证进行分类。对于卡证进行详细的识别，需要使用票据图文识别。

1. 中英文识别（含手写体）

支持输出文本行、字符精确位置。

1. 手写体、签章、指印检出

支持手写体、签章、指印检测，输出位置信息。

支持手写体、签章、指印检出检测等功能打开和关闭。

1. 表格识别

支持表格识别，并以结构化的形式进行输出。

##### 票据OCR识别引擎

票据 OCR ，主要是用于票据领域识别图片中文字的服务。和通用 OCR 主要的不同是，票据 OCR 返回图片中信息的结构化数据。最佳的使用场景是，要识别的要素在票据中是容易识别和标注的，即格式比较固定。

票据OCR识别主要基于机器学习相关技术，针对票据场景图像进行建模，并采用大量票据图像数据进行训练，生成检测和识别模型。其中，检测模型用于检测文字所在的区域，识别模型用于对检测的文字区域进行识别。



图 票据识别技术

模型训练：基于光学、信号处理技术、数据挖掘技术、统计学等领域知识，使用业界领先的RNN、CNN技术进行光学建模，并基于多种场景的票据图像进行检测和识别模型训练，生成检测模型和识别模型。

文本识别：输入检测得到的文本行图像，首先经过必要的预处理,例如倾斜校正、归一化等,再送入识别网络提取特征并进行解码，再进行语言模型等后处理得到最终的识别结果。

主要功能如下：

1. 身份证识别

支持对二代居民身份证正反面的关键字段识别，包括：签发机关、出生日期、姓名、住址、性别、有效期、身份证号码、民族。

1. 营业执照

支持法定代表人、住所、注册资本、注册号、实收资本、类型、信用代码、经营范围、名称、组成形式、成立日期、营业期限信息的识别。

1. 出租车票识别

支持对日期、时间、发票号码、单价、发票代码、金额、车号、里程信息识别。

1. 火车票识别

支持对姓名、火车票号码、出发站、座位类型、达到站、出发时间、车次、身份证号、总金额、座位号、检票口、红字代码、出发日期信息的识别。

1. 增值税发票识别

支持对购买方地址和电话、销售方地址和电话、发票代码、价格列表、开票日期、含税金额、合计税额、税额列表、销售方识别号、货物或应税劳务、服务名称列表、单价、购买方识别号、购买方开户行及账号、发票号码、销售方名称、不含税金额、税率列表、购买方名称、大写金额、销售方开户行、账号等信息的识别。

### 认知智能服务系统

#### 文本分析服务

##### 中文分词引擎

中文分词（Word Segmentation, WS）指的是将汉字序列切分成词序列。因为在汉语中，词是承载语义的最基本的单元。分词是信息检索、文本分类、情感分析等多项中文自然语言处理任务的基础。

例如，句子：国务院总理李克强调研上海外高桥时提出，支持上海积极探索新机制。

正确分词的结果是：

国务院/总理/李克强/调研/上海/外高桥/时/提出/，/支持/上海/积极/探索/新/机制/。

中文是一种十分复杂的语言，让计算机理解中文语言更是困难。在中文分词过程中，有两大难题一直没有完全突破。

一是歧义识别：歧义是指同样的一句话，可能有两种或者更多的切分方法。例如：“化妆和服装”可以分成“化妆 和 服装”或者“化妆 和服 装”；在句子“这个门把手坏了”中，“把手”是个词，但在句子“请把手拿开”中，“把手”就不是一个词；“乒乓球拍卖完了”，可以切分成“乒乓 球拍 卖 完 了”、也可切分成“乒乓球 拍卖 完 了”。由于没有人的知识去理解，计算机很难知道到底哪个方案正确。

二是新词识别：命名实体（人名、地名）、新词，专业术语称为未登录词。也就是那些在分词词典中没有收录，但又确实能称为词的那些词。最典型的是人名，人可以很容易理解。句子“王军虎去广州了”中，“王军虎”是个词，因为是一个人的名字，但要是让计算机去识别就困难了。除了人名以外，还有机构名、地名、产品名、商标名、简称、省略语等都是很难处理的问题，而且这些又正好是人们经常使用的词，因此对于搜索引擎来说，分词系统中的新词识别十分重要。新词识别准确率已经成为评价一个分词系统好坏的重要标志之一。

本引擎设计了多种基于篇章级别的分词策略，用户可以灵活配置策略，以满足更多不同用户对文本分词的需求。

引擎支持两种分词策略，分别为通用分词（common）与3gram分词，用户可根据不同场景下的效果选择相应分词策略。

引擎主函数实现分词功能，其具体实现过程参见下图。主函数支持输入长文本，后对输入文本进行内码转换（内部编码为GBK），之后通过内部设置的标点符号对文本进行分割，并使用响应策略进行分词，结果去嵌套后得到分词后的每个词，将每个词拼接并转成外码，输出最终结果。

**内部标点符号分割**

**多条语句**

**文本输出**

**文本**

**分词策略**

**结果去嵌套与外码转换**

**内码转换**

输出结果除分词结果之外，还可以添加每个词的拼音、词性等信息，附加信息配置项可选。

词性标注（Part-of-Speech Tagging, POS）是给句子中每个词一个词性类别的任务，他是自然语言处理中一项非常重要的基础性工作。这里的词性类别可能是名词、动词、形容词或其他。下面的句子是一个词性标注的例子：

国务院/ni 总理/n 李克强/nh 调研/v 上海/ns 外高桥/ns 时/n 提出/v ，/wp 支持/v 上海/ns 积极/a 探索/v 新/a 机制/n 。/wp

| **词性编码** | **词性名称** | **说明** | **举例** |
| --- | --- | --- | --- |
| Ag | 形语素 | 形容词性语素。形容词代码为 a ，语素代码g前面置以 A | 喜/v 煞/ag 人/n |
| a | 形容词 | 取英语形容词 adjective 的第 1 个字母。 | 最/d 大/a 的/u |
| ad | 副形词 | 直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。 | 一定/d 能够/v 顺利/ad 实现/v 。/w |
| an | 名形词 | 具有名词功能的形容词。 形容词代码 a 和名词代码 n 并在一起。 | 人民/n 的/u 根本/a 利益/n 和/c 国家/n 的/u 安稳/an 。/w |
| b | 区别词 | 取汉字“别”的声母。 |  |
| c | 连词 | 取英语连词 conjunction 的第 1 个字母。 |  |
| dg | 副语素 | 副词性语素。副词代码为 d ，语素代码g前面置以 D。 |  |
| d | 副词 | 取 adverb 的第 2 个字母，因其第 1 个字母已用于 形容词。 |  |
| e | 叹词 | 取英语叹词 exclamation 的第 1 个字母。 |  |
| f | 方位词 | 取汉字“方” |  |
| g | 语素 | 绝大多数语素都能作为合成词的“词根”，取汉字  “根”的声母。 |  |
| h | 前接成分 | 取英语 head 的第 1 个字母。 |  |
| i | 成语 | 取英语成语 idiom 的第 1 个字母。 |  |
| j | 简称略语 | 取汉字“简”的声母。 |  |
| k | 后接成分 |  |  |
| l | 习用语 | 习用语尚未成为成语，有点“临时性”，取“临”  的声母。 |  |
| m | 数词 | 取英语 numeral 的第 3 个字母， n，u 已有他用。 |  |
| Ng | 名语素 | 名词性语素。名词代码为 n ，语素代码g前面置以  N。 |  |
| n | 名词 | 取英语名词 noun 的第 1 个字母。 |  |
| nr | 人名 | 名词代码 n 和“人 (ren) ”的声母并在一起。 |  |
| ns | 地名 | 名词代码 n 和处所词代码 s 并在一起。 |  |
| nt | 机构团体 | “团”的声母为 t ，名词代码 n 和 t 并在一起。 |  |
| nz | 其他专名 | “专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。 |  |
| o | 拟声词 | 取英语拟声词 onomatopoeia 的第 1 个字母。 |  |
| p | 介词 | 取英语介词 prepositional 的第 1 个字母 |  |
| q | 量词 | 取英语 quantity 的第 1 个字母。 |  |
| r | 代词 | 取英语代词 pronoun 的第 2 个字母 , 因 p 已用于介 词。 |  |
| s | 处所词 | 取英语 space 的第 1 个字母。 |  |
| tg | 时语素 | 时间词性语素。时间词代码为 t, 在语素的代码 g前面置以 T。 |  |
| t | 时间词 | 取英语 time 的第 1 个字母。 |  |
| u | 助词 | 取英语助词 auxiliary |  |
| vg | 动语素 | 动词性语素。动词代码为 v 。在语素的代码 g 前面 置以 V。 |  |
| v | 动词 | 取英语动词 verb 的第一个字母。 |  |
| vd | 副动词 | 直接作状语的动词。动词和副词的代码并在一起。 |  |
| vn | 名动词 | 指具有名词功能的动词。动词和名词的代码并在一起。 |  |
| w | 标点符号 |  |  |
| x | 非语素字 | 非语素字只是一个符号， 字母 x 通常用于代表未知 数、符号。 |  |
| y | 语气词 | 取汉字“语”的声母。 |  |
| z | 状态词 | 取汉字“状”的声母的前一个字母。 |  |
| un | 未知词 | 不可识别词及用户自定义词组。取英文 Unkonwn首 两个字母。 |  |

##### 命名实体识别引擎

命名实体识别（Named Entity Recognition，简称NER）是信息提取、问答系统、句法分析、机器翻译、面向Semantic Web的元数据标注等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。一般来说，命名实体识别的任务就是识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。

通常包括两部分：（1）实体边界识别；（2） 确定实体类别（人名、地名、机构名或其他）。英语中的命名实体具有比较明显的形式标志（即实体中的每个词的第一个字母要大写），所以实体边界识别相对容易，任务的重点是确定实体的类别。和英语相比，汉语命名实体识别任务更加复杂，而且相对于实体类别标注子任务，实体边界的识别更加困难。主要体现在：

（1）汉语文本没有类似英文文本中空格之类的显式标示词的边界标示符，命名实体识别的第一步就是确定词的边界，即分词；

（2）汉语分词和命名实体识别互相影响；

（3）除了英语中定义的实体，外国人名译名和地名译名是存在于汉语中的两类特殊实体类型；

（4）现代汉语文本，尤其是网络汉语文本，常出现中英文交替使用，这时汉语命名实体识别的任务还包括识别其中的英文命名实体；

（5）不同的命名实体具有不同的内部特征，不可能用一个统一的模型来刻画所有的实体内部特征。

与自然语言处理研究的其他任务一样，早期的命名实体识别方法大都是基于规则的。系统的实现代价较高，而且其可移植性受到一定的限制。

自 20 世纪 90 年代后期以来，尤其是进入 21 世纪以后，基于大规模语料库的统计方法逐渐成为自然语言处理的主流，一大批机器学习方法被成功地应用于自然语言处理的各个方面。根据使用的机器学习方法的不同，我们可以粗略地将基于机器学习的命名实体识别方法划分为如下四种：有监督的学习方法、半监督的学习方法、无监督的学习方法和混合方法。

下表对这些方法进行了简要归纳。

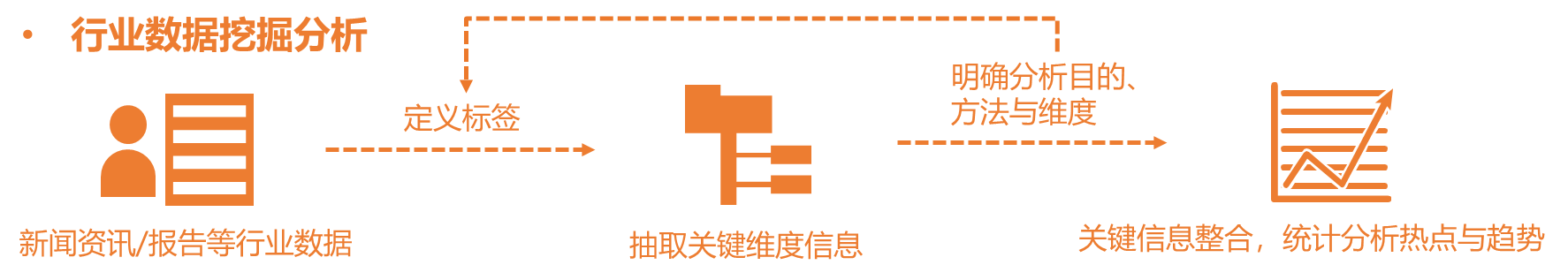


本系统融合关键词、文法、SVM和神经网络策略针对公安数据进行分析，实现人名、机构名、地名、时间、日期、货币和百分比等通用实体的识别。

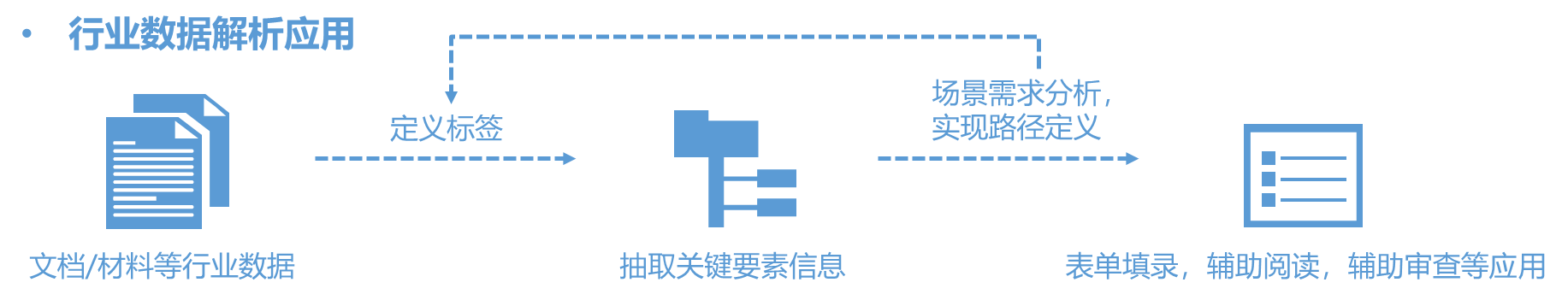
##### 要素抽取引擎

要素抽取技术是从自然语言文本中抽取行业专家预先定义好的标签内容，可以为词、句子或短语，进而对行业文本进行结构化。结构化本身不是目的，只是实现应用的一种方式。通常在两类应用中会用到文本结构化。

第一类是行业数据分析（大致思路如下图）。面对大量行业文本数据，大多数情况下都是自然语言文本的非结构化数据，用户需要利用这些数据做一些关键信息统计或业务分析，就需要对原始文本进行按需结构化，而做哪些结构化则需要按照业务需求进行分析定义，技术上称为“定义标签”，也叫“定标”过程。例如已有全量刑事判决书数据，需要统计犯罪嫌疑人作案时年龄分布。根据对判决书的了解，通常不会直接有犯罪嫌疑人年龄的表述，则此时无法定义“嫌疑人作案年龄”的标签直接抽取。需要间接定义“嫌疑人出生日期”+“嫌疑人作案时间”两个标签（这两类标签内容判决书中都会描述）进行抽取，基于抽取到的结果再进行换算得到嫌疑人作案年龄。



第二类是行业数据解析应用（大致思路如下图）。面对非结构化的行业文本数据，根据业务需求，需要提供一些工具类辅助对这类数据进行结构化应用，包括提取关键信息进行表单填录，辅助阅读，辅助审查（包括规范性、合法性等质检）。

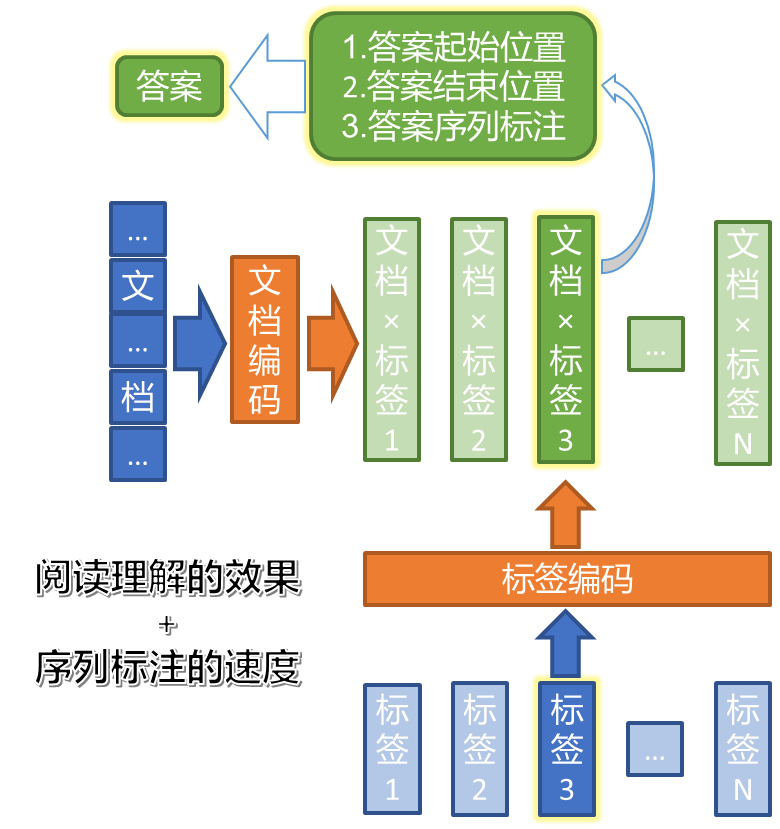


要素抽取是信息抽取领域的一个研究方向，当前主流方式是采用序列标注的方式实现（常见网络结构为CNN-BiLSTM-CRF），任务目标是从一个文本序列中找到某个事先定义好的标签是从第几个字到第几个字。此方法存在以下缺点：

1. 对于一些长标签（即标签内容文本特别长的情况）效果很差，
2. 对于标签重合的问题无法解决（例如“作案经过”标签内容中包含了“作案地点”与“作案时间的标签”）
3. 单个模型通常只能支持一类文本的要素抽取，多文本要素抽取对于硬件资源要求较高

基于以上问题，我们提出了基于阅读理解框架的BERT-CNN-CRF模型。该方案用于解决多标签序列标注问题，支持嵌套标签和重叠标签，通过滑窗支持任意长度的文本。与CRF序列标注模型相比，该方案能较好的解决长标签断开的问题。主要具备以下特性：

1. 对于复杂文本与标签效果好，可以解决包括长标签以及标签覆盖的问题
2. 一次预测所有标签位置，和序列标注一样效率高
3. 多类文本标注数据可以联合训练到一个模型中，部署成本低



要素抽取引擎可以有效解决抽取任务中常见的难点问题：

1. 标签嵌套的问题。例如，“北京大学”中“北京”是地名，“北京大学”是机构名，对于这种情况，“北京大学”和“北京”都能抽取出来；
2. 短的标签抽取效果好，长的标签效果差的问题。对于句子和段落级别的要素，本文的要素抽取引擎也可以很好的抽取出来，例如司法领域的“作案经过”和“辩解内容”等段落级标签，抽取出来的长内容完整，信息不会中断；
3. 标签数量多会显著影响抽取速度的问题。本文的要素抽取可以一次性抽取文书中所有已知的标签类别，即使标签类别超过100个，也不会显著影响抽取的速度。

本产品配套搭配“公安认知生产平台”，非算法人员的一般人员可快速上手完成要素抽取能力服务生产。“公安认知生产平台”以任务流形式引导全流程，设计引擎服务包、标注规范、协议文档自动生成功能，无需线下人工编写整理，降低任务复杂度，实现研究人员0投入，成本大大降低。

任何没有机器学习背景的人员，都可以通过要素抽取交付平台，使用简单地交互获取要素抽取的能力。

##### 关系抽取引擎

关系抽取是在要素抽取的基础上，进一步找到要素信息点之间的关系，建立信息点之间的关联，比如将某个“出生日期”与某个“犯罪嫌疑人”进行关联，我们可以知道其中的所属关系。通过要素抽取+关系抽取，我们将非结构化的自然语言文本转化为结构化的机器可查询处理的信息，迅速获取所关注文本信息点和信息点之间的关系，显著提升文本处理的效率。

关系抽取的应用场景主要有：

（1）行业数据信息的结构化，提升文本处理效率；

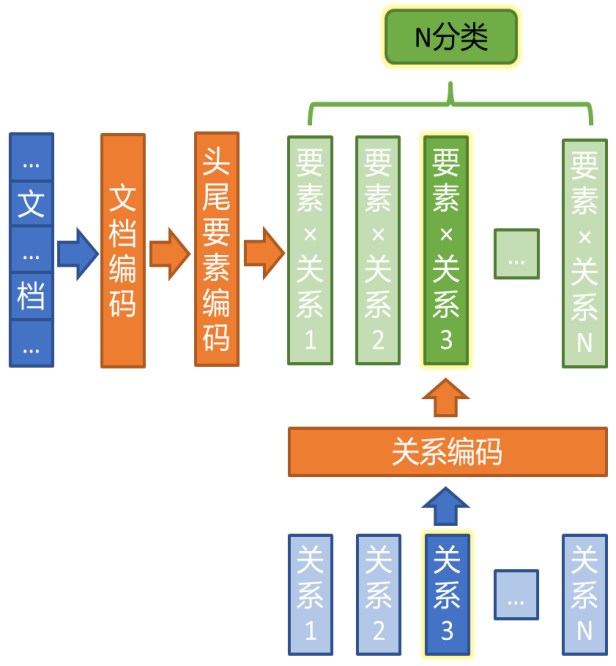
（2）创建新的结构化知识库（knowledge base）并且增强现有知识库；

（3）构建垂直领域知识图谱：医疗、化工、农业、教育；

（4）支持上层应用：问答、搜索、推理等。

关系抽取是信息抽取领域的一个研究方向，常用的关系抽取方法有5类，分别是基于模式匹配、基于词典驱动、基于机器学习、基于本体和混合的方法。基于模式匹配和词典驱动的方法依靠人工制定规则，耗时耗力，而且可移植性较差，基于本体的方法构造比较复杂，理论尚不成熟。基于机器学习的方法以自然语言处理技术为基础，结合统计语言模型进行关系抽取，方法相对简单，并具有不错的性能，是当下关系抽取的主流方法。在深度学习兴起之前，关系抽取的传统机器学习方法依赖于特征工程，而这些特征通常由预先准备的NLP系统得到，这容易在构造特征的过程中造成误差累积，阻碍系统性能。

我们采用了要素-关系交互式的神经网络模型，如下图所示：

使用BERT编码器对文本进行编码，从文本表示中分别取出Head标签和Tail标签信息最后一个字符的隐层表示hHead和hTail；然后将Head标签、Tail标签、关系标签进行合并，获取新标签的BERT embedding，经CNN及Max pooling后得到标签隐层表示hsub和hmul；最后计算头尾要素和关系之间的相似度，选择概率最大的关系为最终预测的要素之间关系。

该方案主要具备以下特性：

1. 端到端的技术方案，具有良好的通用性和可扩展性。
2. 基于要素抽取结果，预测任何两个要素间的业务关系，根据不同业务需求，要素间的关系可人为预先指定，无关系也可以当做一种特殊的关系来处理，我们通过logsumexp loss缓解了关系有无非常不平衡的问题，保障了关系抽取的效果。
3. 充分利用了BERT、CNN强大的特征抽取能力，对Head标签、Tail标签和关系三元组的相似度关系进行建模。
4. 将文本和关系标签分别进行编码，一次预测所有要素间的关系，效率大大提升。
5. 通过滑窗解决了长文本编码的问题。

本产品配套搭配“公安认知生产平台”，非算法人员的一般人员可快速上手完成要素抽取能力服务生产。“公安认知生产平台”以任务流形式引导全流程，设计引擎服务包、标注规范、协议文档自动生成功能，无需线下人工编写整理，降低任务复杂度，实现研究人员0投入，成本大大降低。

任何没有机器学习背景的人员，都可以通过信息抽取交付平台，使用简单地交互获取关系抽取的能力。

##### 文本分类引擎

文本分类引擎的主要功能是对文本所包含的内容进行自动的分析，识别出文本在各个标签体系下的类别，通过对大量的文本数据进行批量处理就能快速判定对应类别从而准确的将文本数据进行归类处理。

目前，文本分类技术不仅广泛应用于教育、信访、医疗等行业场景，在政法业务各个产品或项目中起着重要作用，并取得良好效果。如，基于文本分类的主次件识别可以辅助信访人员快速从大量的信件中挑选出主要信件。

文本分类技术是将自然语言文本按照专家预先定义的标签体系进行自动分类标记。文本分类不仅可以协助人们管理和解决复杂的海量数据，还可以帮助人们从海量的数据中快速地获取用户需要的信息。文本分类技术的应用场景非常丰富，并已经成熟应用于众多领域，如文本内容审核、广告过滤、情感分析、舆论分析、人机对话、意图识别和推荐等NLP领域。

我们提出了基于预训练语言模型的文本分类模型，支持单标签分类和多标签分类，对同一份数据支持多个标签体系下的分类，可分别给出数据在各个标签体系下的分类结果。多个标签体系既可包含单标签标签体系也可包含多标签体系。

模型首先通过预训练的语言模型得到文本的语义表达，然后通过一个全连接层得到文本的分类表示。最后根据不同的分类任务，将文本的分类表示送入不同的分类层。如果是单标签分类任务，则将文本的分类表示送入softmax层，如果是多标签分类任务，则将文本的分类表示送入sigmoid层。

本产品配套搭配“公安认知生产平台”，非算法人员的一般人员可快速上手完成文本分类能力服务生产。“公安认知生产平台”以任务流形式引导全流程，设计引擎服务包、标注规范、协议文档自动生成功能，无需线下人工编写整理，降低任务复杂度，实现研究人员0投入，成本大大降低。

任何没有机器学习背景的人员，都可以通过文本分类交付平台，使用简单地交互获取文本分类的能力。

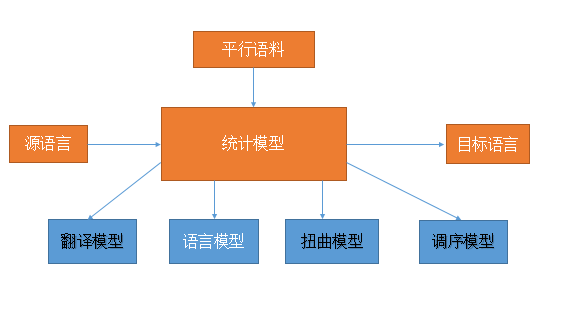
#### 机器翻译服务

##### 中英翻译引擎

本引擎实现英文、汉语文本之间的互译。

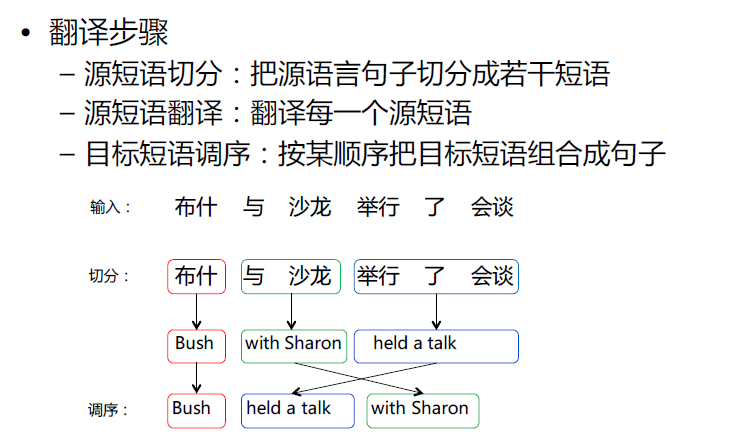
机器翻译主要流程为分词、规则匹配、专有名词匹配、解码、后处理等，其中解码是翻译流程中的核心功能，目前翻译引擎支持两种解码器解码，传统的基于层次短语的解码器PSMT和基于神经网络的解码器NMT。

1）PSMT解码器：



基于统计的短语翻译模型图

PSMT的模型主要包含翻译模型、语言模型、扭曲模型、调序模型等四个模型，主要流程大概如下：



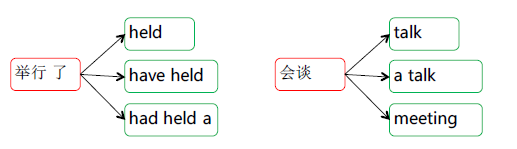
PSMT模型示意图

可以看到翻译的主要过程是将句子按照短语进行切分，每个短句分别进行翻译，然后再进行调序。这样翻译的搜索空间是：

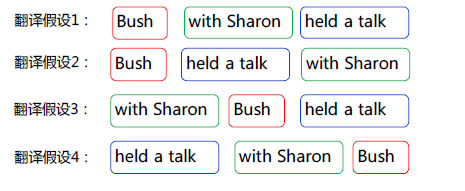
穷举所有的短语切分可能：



* 穷举所有的翻译候选



穷举所有的调序可能

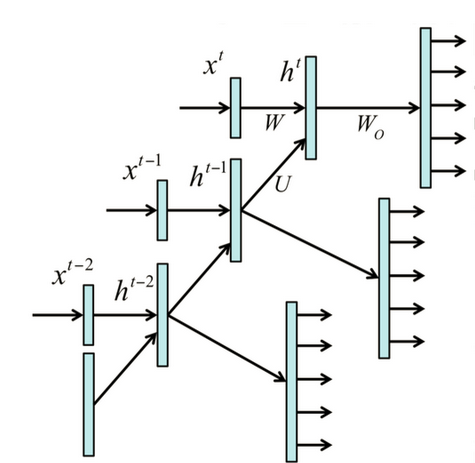


这样在搜索空间中就获得所有的翻译假设，这个时候再对所有的翻译假设利用对数线性模型打分，选择得分最高的翻译假设即为翻译结果。

2）NMT解码器

在自然语言处理，乃至人工智能领域的很多问题中，输入序列的长度和输出序列的长度都是不固定的，比如机器翻译、手写字体识别、语音识别等。要处理这种变长的输入和输出我们通常使用循环神经网络（Recurrent Neural Network, RNN）来处理。对于简单的多层前馈神经网络，每一次输入都要重新计算网络的中间状态（即网络隐层的激活值），它不受之前样本计算得到的中间结果的影响，因为网络不能保存每一次得到的中间状态信息，而RNN就可以保存这一信息。RNN循环读取一个序列（如词序列），根据当前的输入跟历史状态计算当前的状态信息，因此RNN可以处理任意长度的输入，而中间状态也被保存下来，可以一起用于预测输出序列。

RNN的主要思想就是循环地将输入的序列压缩成一个固定维度的向量，也就是网络的中间状态，这一过程是通过不断将当前时刻的输入和历史状态组合实现的。如图1所示：对于某一个时刻的输入（就翻译而言，是一个词一个词的输入），其输出都同时考虑了这一次的输入（x·t）和上一时刻的输出（ht-1）。



NMT模型示意图

但是RNN存在很严重的问题，当RNN的层次太多的时候会出现梯度爆炸或者梯度消失的问题，为了解决这个问题有人在RNN的基础上使用门（Gates）来改进网络结构，这种改进的网络结构叫做长短时记忆（Long short-term memory），这里不做详细展开。

从机器学习的角度看，机器翻译其实就是一个有监督的学习过程，学习一个任意长度的序列到另一个任意长度的序列的过程。前面提到的RNN有两个好处：

可以将一个序列压缩成一个表示（representaion）

以概率的方式对一个序列进行建模。

很多序列化标注问题都可以利用encoder-decoder框架解决，NMT同样采用了这一框架。以“中国经济发展迅速”为例，首先编码（encode）阶段，由一个编码器（Encoder）将一个输入词序列转换为一个固定维度的向量（hi），解码（decode）阶段将根据编码出的向量（hi）生产目标译文，比如“The Chinese economy develops rapidly”。Encode-Decoder框架非常直观，也是NMT背后最基本的思想。这里的编码阶段和解码阶段都用RNN来实现。

##### 维汉翻译引擎

为了构建维汉翻译引擎，需着重研究面向维吾尔语语言特性的自动语言分析处理算法；同时面向口语特性，研究基于深度神经网络的翻译模型和基于大规模循环神经网络的语言模型用于维汉机器翻译，构建高性能可实用的维汉机器翻译系统。整个构建工作包括：

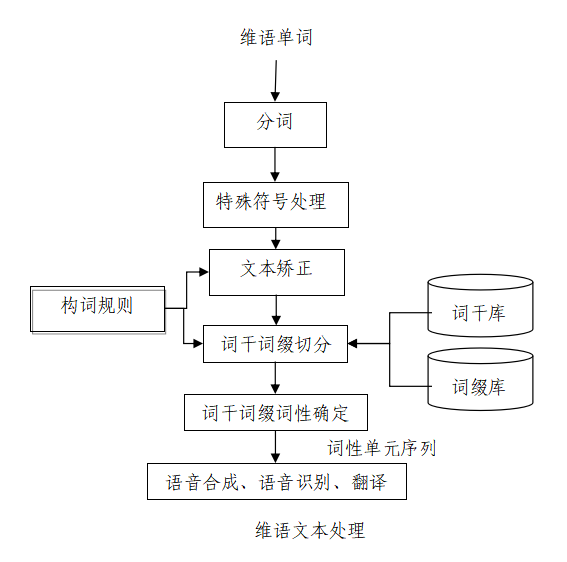
1. 大规模语料资源建设

随着ISO 10646-1:2000版的颁布，蒙、藏、维(哈、柯)、朝、彝等民族文字有了统一的字符集编码，大大促进了民族文字处理软件的发展。统计机器翻译需要积累大量的民语文语料资源，并对民族文翻译语料库的加工处理有丰富的经验，从而为海量的民族文句对齐库的建成提供了有力的数据保证。

在系统研发中，我们从现有的资源中整理出一定规模的汉、维单语数据及汉维双语平行数据。并使用自动段落对齐、句对齐技术，对双语数据进行对齐，并得到句子一级对应的双语平行语料为机器翻译系统的训练提供基础。此外，还使用数据质量检测技术排除掉低质量数据，减少噪声数据对后续机器翻译系统训练的负面影响。

1. 维吾尔语语言自动分析系统构建

维吾尔语语言分析是本系统的重点需要突破的技术之一。维吾尔语是典型的黏着语言，在构词和语法构成上不同于孤立语言。黏着语言大部分单词是由词干或词根后缀加若干个词缀构成，它需要很大的词典才能达到与孤立语言相当的效果。其次自然词词性变化复杂，变化后的词在句法上相当于孤立语言中的短语，黏着后的词序连接变化多样，因此维语单词的词干词缀切分、词法分析等模块至关重要，它与文本校对，整词功能词性确定，音变处理等都有一定的影响，该功能在语音合成和机器翻译中、不可或缺，并且对于语音识别中的语言模型都是十分重要的。下图是维吾尔语语言文本处理框图:



1. 基于深度神经网络翻译模型的维汉机器翻译系统构建

维汉自动翻译系统采用统计机器翻译技术，该技术具有系统健壮性强、对数据依赖小、系统构建全自动化等特点。简单来说，统计机器翻译可以被解释为：对于输入的句子s，我们在所有可能的目标语译文中找到最可能的译文t\*输出：



其中Pr(t|s)表示将输入句子s翻译为目标语译文t的概率，argmax表示搜索最佳译文的过程。

对于模型训练和解码，现在有相对成熟的技术可以实现，比如，最大似然估计+自底向上解码，而模型的选择是翻译系统成功的关键因素之一。本项目拟采用深度神经网络方式的翻译模型建模技术和基于神经网络的大规模语言模型建模技术，将会使得机器翻译效果比原来有巨大的提升。

神经网络联合翻译模型用于计算一个源短语S翻译成一个目标短语T的概率P(T|S)。该概率可以分解为目标短语中各个词ti的概率之积。每个词的概率既依赖于前面已经翻译出来的目标词，也受相应的源短语影响。公式表示为：



其中Si是跟当前目标词Si关系较强的源语言上下文。固定源语言上下文的长度为m。同时通过翻转翻译方向和语言模型的方向，得到一些列模型变种，如：

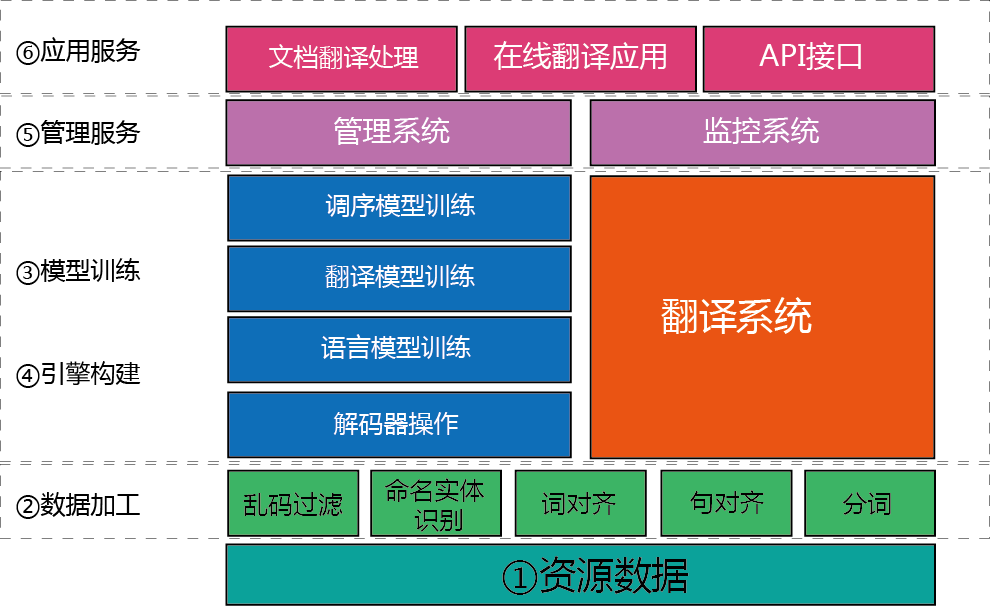




等，这些变种的加入是的模型性能更加稳定。

语言模型的建模也尤为重要，本系统中将会对海量文本语料进行超大规模神经网络语言模型训练，并进行基于自动分类的语言模型组合、基于话题的语言模型自适应等研究工作。

在本系统中，我们使用基于短语的统计机器翻译方法，翻译模型采用神经网络建模方式，语言模型采用循环神经网络建模方式。具体框架如下：



## 开发平台

### 人机交互建模系统

以私有化、可视化的方式进行语义的定制开发和使用。通过可视化的方式将语义的开发和测试简单化，降低语义的开发门槛，能快速适配满足多种场景下的语义定制化需求。

#### 语义管理

1. 可视化编辑

系统可以对技能、意图、说法、实体及辅助词等各种项目进行可视化编辑，相比之前写ABNF文法文件的形式，大大增强了易用性。

1. 技能管理

技能即语义理解系统支持的不同垂直领域的分类，例如weather(天气)、train(火车)、flight(航班)、musicX(音乐)、movie(电影)等；技能管理可以对技能进行增删改等操作。

1. 意图管理

意图即本次语义请求想要实现的操作类型，例如QUERY(查询)、PLAY(播放)、LAUNCH(启动)、TRANSLATION(翻译)、CREATE(创建)、SET(设置)、SEND(发送)、DIAL(拨打)、INSTRUCTION(指令)、OPEN(打开)等。意图管理页面可以对技能中的意图进行增删改等操作。

1. 说法(语料)编辑

语料即用户在完成某一意图时，可能会有很多种表述方式，表达方式的例句即为语料；说法编辑可以对同一个意图的各种不同说法进行增删改等操作，并可为说法中不同部分指定其对应的实体或者辅助词。

智能贴弧。智能贴弧功能可以在输入语料时自动分析语料各部分所匹配的系统内置实体和辅助词，能明显简化录入语料的工作量。

1. 通用(开放)实体

用户的语料中，同一类型的名词将会被总结为实体；为了让用户编辑说法时能更加轻松快捷的设置其对应实体，系统内置了以下45种常用开放实体，开箱即用，可大大简化日常说法的配置工作量。其中包含了比较特殊的5类实体，这些实体取值范围是无穷的，不能通过枚举其取值范围来定义，具体如下：

* 计算实体

计算实体用于匹配说法中的简单计算，例如“13乘16等于多少”中，“13乘16”可对应计算实体。

* 人民币实体

人民币实体用于匹配各种常用人民币金额，例如“您一共消费19块3毛3”，其中“19块3毛3”可对应人民币实体。

* 通配实体

通配实体用于匹配没有明显规则的实体信息，并可指定匹配长度。

例如“查询张三的交通违章记录”，其中“张三”为人名，其没有可枚举的取值范围，且没有明显规则，此时可以将“张三”对应通配实体。

* 数字实体

数字实体用于匹配数字的各种常用说法，例如“共五千一百二十二参会”，“五千一百二十二”可对应数字实体

* 时间实体

时间实体用于匹配有关时间的各种常用说法，例如“查询上周三的交通违章信息”，“上周三”可对应时间实体。

* 常用开放实体

|  |  |  |  |
| --- | --- | --- | --- |
| **姓氏** | IFLYTEK.FamilyName | 姓氏 | 赵;钱;孙;李;周;吴;郑;王 |
| **计算** | IFLYTEK.Calculator | 算式 | 9的平方根是多少 |
| **人民币** | IFLYTEK.Rmb | 常见的人民币说法：元，角，分，块，毛，支持各种大小写，一，二，壹，贰等 | 1元;五角 |
| **搜索** | IFLYTEK.Search | 操作词搜索、查看 | 搜索 |
| **网页** | IFLYTEK.Website | 网页 | 新浪娱乐电影网;网易娱乐电视 |
| **语气词** | IFLYTEK.Yq | 语气词 | 了;啊 |
| **前缀词** | IFLYTEK.ActionPrefix | 前缀词 | 请问;请你帮我 |
| **中国高校** | IFLYTEK.ChinaCollege | 中国高校实体 | 清华大学;中国科学技术大学;北京大学;中央民族大学;北京航空航天大学 |
| **通配** | IFLYTEK.Wildcard | 匹配任意1-20个字符 | 1407会议室;红色的叮咚音响;一副黑框眼镜 |
| **时间** | IFLYTEK.Datetime | 日期和时间 | 11月18号;三点;上午;端午节;情人节;星期一 |
| **广播电台** | IFLYTEK.RadioChannel | 广播电台 | 南京体育电台;成都新闻广播电台 |
| **小说** | IFLYTEK.Novel | 小说 | 钢铁是怎样炼成的;斗破苍穹 |
| **数字** | IFLYTEK.Number | 1-20位数字 支持汉字数字和阿拉伯数字 | 1101;一仟二百一拾五 |
| **应用** | IFLYTEK.App | iOS、Android手机应用程序 | 直升机飞跃障碍;美团;微听FM;中级经济师星题库;9.9包邮汇 |
| **中国股票** | IFLYTEK.ChinaStock | 中国国内主要A股股票 | 嘉实恒生;绿地控股;生物股份;精伦电子;卡姆丹克太阳能 |
| **中国热线** | IFLYTEK.ChinaHotline | 常用服务号码 | 119;10086;10010;10000;110 |
| **中国诗人** | IFLYTEK.ChinaPoet | 诗经、唐诗、宋词等诗词作者 | 王安石 ;欧阳修;岳飞;李清照;晏几道 |
| **中国诗歌** | IFLYTEK.ChinaPoem | 诗经、唐诗、宋词等诗词 | 春晓;征人怨;题西林壁;黄鹤楼送孟浩然之广陵;长歌行 |
| **故事** | IFLYTEK.Story | 故事名称 | 三个和尚;画龙点睛 ;青蛙王子;掩耳盗铃童话版;青蛙王子 |
| **音乐专辑** | IFLYTEK.MusicAlbum | 中、日、韩、欧美音乐专辑 | 捉迷藏;MerryChristmas;十送红军 |
| **中国电影院** | IFLYTEK.ChinaCinema | 中国各地区电影院 | 金鸡百花电影院;中环城电影院;余姚万达电影院 |
| **演员** | IFLYTEK.Actor | 中、日、韩、欧美影视演员 | 阿诺德施瓦辛格;angelababy;崔智友;张柏芝;周慧敏 |
| **综艺节目** | IFLYTEK.TVShow | 综艺节目 | 中国好声音澳门海选;KpopStar3;KBS歌谣大祝祭;奔跑吧兄弟;AKB48神TV13 |
| **电视剧** | IFLYTEK.TVSeries | 电视剧 | 康熙王朝;西游记 |
| **视频** | IFLYTEK.Video | 电影，电视剧，动漫，纪录片，微电影，演唱会，大型比赛，动画片，传记等视频 | 美人私房菜之玉蝶传奇;深爱食堂;索命飞刀;双面少女 |
| **电影** | IFLYTEK.Film | 电影院上映的主要院线电影 | 湄公河行动;绝地逃亡;迷失课后 |
| **中国电视频道** | IFLYTEK.ChinaTVchannel | 中国国内能收到的电视频道 | 亳州都市频道|亳州都市;安徽经济生活|安徽经济|安徽经济频道|安徽经视;IPTV理财;家家购物频道|家家购物;CCTV-1|CCTV-1综合|综合频道|CCTV1综合频道|中央一台|CCTV 1|CCTV-1 综合|中央综合 |
| **导演** | IFLYTEK.Director | 中、日、韩、欧美影视导演 | 冯小刚;李安;张艺谋;阿尔弗雷德希区柯克;爱德华兹威克 |
| **中国菜** | IFLYTEK.ChinaFood | 中文特色菜 | 红烧肉;宫保鸡丁 |
| **中国餐馆** | IFLYTEK.ChinaRestaurant | 中国国内餐厅 | 比尔烤肉;宝燕一号;福记大饭店;好苑建国酒店 |
| **中国酒店** | IFLYTEK.ChinaHotel | 中国国内各地区酒店、商务宾馆、会议中心、招待所 | 蜀山国际大酒店;新铁牛商务宾馆;克莱雅商务宾馆;爱情海商务宾馆;喜来登商务宾馆 |
| **中国飞机场** | IFLYTEK.ChinaAirport | 中国国内飞机场 | 天水机场;夏河机场;芒市机场;南郊机场;新桥机场 |
| **中国火车站** | IFLYTEK.ChinaTrainStation | 中国国内动车火车站以及城际铁路火车站 | 北京南;重庆北;南京南;上海虹桥;月山 |
| **中国火车车次** | IFLYTEK.ChinaTrainCode | 中国国内城市间火车车次 | k4695;c2010;c2080;z16;k1137 |
| **中国航班号** | IFLYTEK.ChinaFlightCode | 中国国内、国际航班班次 | mu2348;2b605;lv521;3u8987;ca6232 |
| **航空公司** | IFLYTEK.Airline | 航空公司 | 天星航空公司;中国北方航空公司;中国南方航空股份有限公司;马来西亚亚洲航空司;马丁航空 |
| **中国景区** | IFLYTEK.ChinaSightSpot | 中国风景区地点 | 黄山风景区;三门峡大坝;木兰古门;晴川阁;蝴蝶岛 |
| **中国区县** | IFLYTEK.ChinaArea | 中国国内所有区县名称 | 包河区;虹口区;新兴区;朝阳区;静安区 |
| **非中国城市** | IFLYTEK.NonChinaCity | 全球主要城市名称 | 东京;纽约;伦敦 |
| **中国城市** | IFLYTEK.ChinaCity | 中国所有城市名称 | 合肥市;东莞市;北京市 |
| **非中国省份** | IFLYTEK.NonChinaProvince | 国外省州名称 | 加州;阿拉斯加 |
| **中国省份** | IFLYTEK.ChinaProvince | 中国省份 | 安徽省;广东省;江苏省 |
| **国家** | IFLYTEK.Country | 全球主要国家名称 | 中国;法国;美国 |
| **艺术家** | IFLYTEK.Artist | 中、日、韩、欧美歌手、乐队 | 阿桑;周云蓬;米希亚;周杰伦;小沈阳 |
| **歌曲** | IFLYTEK.Song | 中文、日文、英文三类音乐歌曲 | 双截棍;旅行的意义;十三号星期舞;千纸鹤 |

1. 自定义实体

除了系统提供的通用(开放)实体，用户还可以基于自己的实际业务需要自行定义实体。

* 自定义普通实体

在创建“预订会议室”技能时，“我想订301会议室”，其中301为会议室号码，利用自定义普通实体功能，可以为所有的会议室号码定义一个实体“会议室”，其中的每条记录都是一个会议室号码。

* 自定义组合实体

自定义组合实体可以解决条目不能穷举但是有规则的实体。普通实体用在实体条目可穷举的场景中，例如上面的“会议室”号码。而对于条目不可穷举的实体，例如车牌，则无能为力。而组合实体可以将一个车牌“皖A1111B”看作如下实体的组合：

省简称(京，沪，皖等)

行政区(A，B，C等)

车牌号码(数字及字母)〈重复5-6次〉

通过定义以上三种子实体并组合成车牌实体，可以精确匹配语料中出现的各种车牌。

1. 通用(开放)辅助词

**辅助词：**用于帮助技能编写人员编写语料，但在最终的语义解析结果中不会展现；

系统包含25个通用(开放)辅助词。

辅助词一般用于匹配常见的辅助语，其定义与实体极其类似，只是在使用上有区别。

在说法中设置某一部分为辅助词时，最终的语义结果中不会包含此部分；设置某一部分为实体时，最终的语义结果中会包含此部分；

|  |  |  |  |
| --- | --- | --- | --- |
| **怎样** | IFLYTEK.How | 怎么样 | 怎样|如何|怎么样|咋样 |
| **开启** | IFLYTEK.Launch | 开启 | 打开|开启|开机|登录|进入 |
| **预订** | IFLYTEK.Book | 预订 | 定|订|预订|预定|购 |
| **播放** | IFLYTEK.Play | 播放 | 播放|播|点播|放|放下|播下|播放一下 |
| **查询** | IFLYTEK.Query | 查询 | 查询|查寻|查找|咨询|了解|找找 |
| **非常** | IFLYTEK.Very | 非常 | 非常|比较|十分|特别|相当 |
| **是否** | IFLYTEK.Whether | 是否 | 是|是否|是不是|似|似不似 |
| **随机** | IFLYTEK.Random | 随机 | 随便|任意|自由|随机 |
| **最近** | IFLYTEK.Recently | 最近 | 新|刚|刚刚|最新|最近 |
| **附近** | IFLYTEK.Around | 附近 | 附近|周围|周边|最近|旁边 |
| **那里** | IFLYTEK.There | 那里 | 那里|那|那儿|那个地方|那个位置 |
| **这里** | IFLYTEK.Here | 这里 | 这里|这|这儿|此地|本地|当前位置 |
| **我的** | IFLYTEK.My | 我的 | 我的|我们的|俺的 |
| **你** | IFLYTEK.You | 你 | 你|你们 |
| **我** | IFLYTEK.Me | 我 | 我|我们|俺|俺们|额|额们|咱|咱们 |
| **能不能** | IFLYTEK.Can | 能不能 | 能不能|好不好|行不行 |
| **语气词** | IFLYTEK.Modal | 语气词 | 呐|嘞|了|吧|吗|么|啊|呀|咯 |
| **请你** | IFLYTEK.Please | 请你 | 请|请您|请你|麻烦|麻烦你|麻烦您 |
| **帮我** | IFLYTEK.HelpMe | 帮我 | 帮忙|帮个忙|帮我|给我|替我 |
| **我想要** | IFLYTEK.IWant | 我想要 | 我打算|我想|我要|我想要|要|想 |
| **有没有** | IFLYTEK.Have | 有没有 | 有|可有|有没有|有木有|有么有 |
| **什么** | IFLYTEK.What | 什么 | 什么|啥|哪个 |
| **地点** | IFLYTEK.Where | 地点 | 什么地方|哪地方|哪个地方|啥地方 |
| **告诉** | IFLYTEK.Inform | 告诉 | 告诉|联系|通知|告知 |
| **边缘词** | IFLYTEK.Margin | 边缘词 | 这个|那个|呃|嗯|啊|这|那 |

1. 自定义辅助词

辅助词在定义时跟实体几乎一样，只不过它不出现在最终的语义分析结果中。一般是跟实际业务无关联但是在用户的常用表述中又经常会出现的一类词汇，如「我要」、「查询」、「帮我查」等。

1. 模糊匹配

模糊匹配功能可以在用户所说的话与技能语料有出入的情况下进行智能匹配，可以大大减小为技能定义语料的工作量。

* 句式模糊

句式模糊可以匹配实际说法与定义说法上有稍许出入的场景，常见情况是实际说法中添加了敬语、语气助词之类。

技能定义语料：我想听{刘德华}的歌

用户实际表述：我想听一首{刘德华}的歌曲

系统会自动将用户表述与预先添加的语料进行匹配，并在最终结果中输出歌手槽位的信息为“刘德华”

* 实体模糊

实体模糊可以智能匹配实际说法中与业务密切相关的槽位短语中出现错误的情况，可用于同音，近音，多字少字及错误等情况。

技能定义语料：我想听{刘德华}的歌

用户实际表述：我想听{刘得华}的歌

系统会自动将用户表述中的“刘得华”通过实体模糊功能匹配上歌手“刘德华”，并在最终结果中输出歌手槽位的信息(value)为“刘得华”，槽位的规整信息(normValue)为“刘德华”。

1. 动态实体

动态实体属于语义个性化，我们创建的一个技能常用于不同的场景、不同的应用、不同的使用者，在不同的情况下可能对应不同的实体内容。比如，若技能中有一个“联系人”实体，那么不同的使用者，纳闷的联系人列表是不一样的；再如，若一个技能有“道路名称”的实体，那么对于不同城市，这个道路名称的列表也是不同的。为了能够应对这种情况，我们实现了处理每个请求时，动态加载对应实体的功能，同时提供实体上传编译功能，我们称这种实体为动态实体。

1. 改槽

**槽：**即语义请求中的内容实体，例如时间、地点、歌名、电影名、小说名、电视频道、电话号码等；

多轮对话完善信息的过程中，当希望对之前填过的槽位信息进行变更时，就会用到改槽功能。

比如：

user：“查询下明天的天气”

sys：“请问查询哪里的天气？”

user：“查后天的吧”

（系统将已填日期改成后天，然后继续询问）

sys：“请问查询哪里的天气？”

user：“合肥”

sys：“查询后天合肥的天气，确认码？”

...

1. 跳转

在多轮对话中，插入新的话题，比如：

user：“帮我订张去北京的机票”（进入“订票”的多轮对话）

sys：“哪天出发？”

user：“北京后天天气怎么样？”（进入“天气”的对话）

sys：“查询北京后天天气，确认吗？”

user：“是的”

（回到订票业务）

sys：“哪天出发？”

...

1. 指代

在上面的技能跳转示例中，跳转的时候可能会说“查询那里的天气”，这里的“那里”就属于指代。为了支持这种说法，系统实现了指代消解。

比如：

user：“帮我订张去北京的机票”（进入“订票”的多轮对话）

sys：“哪天出发？”

user：“明天”

sys：“预定明天去北京的机票，确认吗？”

user：“先帮我查下那天北京的天气”

sys：“查询明天北京的天气，确认码？”（系统通过指代消解，识别出“那天”是指“明天”）

...

1. 意图确认

在前面示例中，系统返回“查询北京后天天气，确认吗？”，属于意图确认。在编写意图确认话术时包含槽位，可以生成包含槽位信息的确认话术。

#### 知识库管理

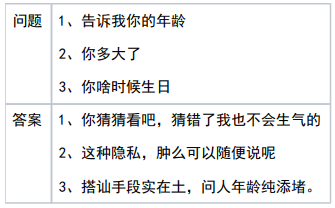
问答是语义理解的一种形式，针对指定问题给出指定答案，问题和答案需要预先设置好。 问答支持如下功能：

1. 问题模糊匹配

例如设定“你多大了”，提问“多大了”、“不知道你多大了”，也能触发。

1. 多对多回复

支持一个问题多种回复，或多个问题一个回复，或多个问题多种回复，但是回复是随机的。比如设置下面一组问题和答案。



提问其中任一问题，系统会随机选择三个答案中的一个给出回复。

##### 内置通用问答

提供9类已有问答库：

| 问答库 | 简介 | 示例 |
| --- | --- | --- |
| 抱怨 | 涵盖心中怀有不满、责怪、批评等内容 | 输入：房价又涨了  回复：我现在还没办法解决这个问题 |
| 十万个为什么 | 涵盖十万个为什么、生活、科学等小知识科普 | 输入：螃蟹为什么会横着走  回复：螃蟹的脚与头胸部是连在一起的，它们的脚关节只能左右移动，所以它只 能横着走了 |
| 情绪 | 涵盖不同心情状态，如伤心、心情好等 | 输入：不开心了怎么办 回复：爱笑的孩子运气都不会太差哦。 |
| 感情 | 涵盖道德感和价值感两个方面，具体表现为爱情、幸 福、美感等 | 输入：你最爱的人是谁 回复：当然有~就不告诉你是谁 |
| 热点 | 涵盖时下最新热点话题、热门信息等 | 输入：苏神是谁  回复：苏亚雷斯嘛，乌拉圭足球运动员，就是在巴西世界杯当中咬人的那个。 |
| 个性化 | 涵盖年龄、性别等基本个人属性介绍等 | 输入：你喜欢网购吗  回复：恩，我比较喜欢网购，一般喜欢不但可以饱眼福，同时也享受逛着大扫荡 的乐趣。 |
| 夸奖 | 涵盖赞美、称赞等内容 | 输入：你太给力了  回复：天生的，没办法，嘻嘻。 |
| 常识 | 涵盖生存技能、生活常识、以及人际交往等各方面 | 输入：古代四大美女是谁  回复：西施、貂蝉、王昭君、杨贵妃。 |
| 问候 | 涵盖常见问候方式、打招呼方式等 | 输入：你好  回复：你好，又见面了真开心。 |

##### 定制问答库

当9类已有问答库不能满足您的个性化需求时，可以通过页面添加自定义的问答库，支持以下两种方式添加问答对：

1、上传excel文件批量导入；

2、输入框编辑逐条添加。

用户可根据日常业务需要，梳理出业务相关的问答知识库，通过定制问答库功能，实现业务问答，并在日常工作和学习中是不断补充完善问答知识库。

#### 警务智能指令集

系统内置多项警务指令集，方便第三方进行集成，实现警务APP控制及警务信息检索。

1. 人员核查指令集

|  |  |
| --- | --- |
| 1 | 查一个人身份证是（号码）姓名叫（名字） |
| 2 | 我要查个人身份证是（号码）名字叫（名字） |
| 3 | 帮我查个人身份证是（号码）叫（名字） |
| 4 | 查个人身份证是（号码）叫（名字） |
| 5 | 给我查个人身份证号码是（号码）姓名叫（名字） |
| 6 | 帮我查一个人身份证号码是（号码）名字叫（名字） |
| 7 | 查人身份证号码是（号码）名字叫（名字） |
| 8 | 查人身份证号码是（号码）叫（名字） |
| 9 | 我要查个人姓名叫（名字）身份证是（号码） |
| 10 | 查人姓名叫（名字）身份证是（号码） |
| 11 | 查人名字叫（名字）身份证是（号码） |
| 12 | 查个人叫（名字）身份证是（号码） |
| 13 | 查个人姓名叫（名字）身份证号码是（号码） |
| 14 | 查个人名字叫（名字）身份证号码是（号码） |
| 15 | 查个人叫（名字）身份证号码是（号码） |
| 16 | 查个人身份证是（号码） |
| 17 | 我要查个人身份证号码是（号码） |
| 18 | 叫（名字）身份证是（号码） |
| 19 | 查一个姓名叫（名字）的人身份证是（号码） |
| 20 | 给我查个名字叫（名字）的人身份证是（号码） |
| 21 | 我要查一个名字叫（名字）的人身份证是 |
| 22 | 我要查一个名字叫（名字）的人身份证号码是 |
| 23 | 帮我查个名字叫（名字）的人身份证 |
| 24 | 查一下姓名叫（名字）的人 |
| 25 | 我需要查个叫（名字）的人 |
| 26 | 帮我查下名字叫（名字）的人 |
| 27 | 姓名叫（名字） |

1. 路线导航指令集

|  |  |
| --- | --- |
| 1 | 我要去（地址） |
| 2 | 我要到（地址） |
| 3 | 带我去（地址） |
| 4 | 接警地址是（地址） |
| 5 | 怎么去（地址） |
| 6 | 我要去（地址）接警 |
| 7 | 我要到（地址）接警 |
| 8 | 带我去（地址）接警 |
| 9 | 给我指条去（地址）的路线 |
| 10 | 帮我规划一下去（地址）的路线 |
| 11 | 帮我规划一下到（地址）的路线 |
| 12 | 规划一下去（地址）的路线 |
| 13 | 规划一下到（地址）的路线 |
| 14 | 查一下去（地址）的路线 |
| 15 | 查一下到（地址）的路线 |
| 16 | 导航去（地址）的路线 |
| 17 | 去（地址）怎么走 |
| 18 | 去（地址）怎么走最快 |
| 19 | 到（地址）怎么走 |
| 20 | 到（地址）怎么走最快 |
| 21 | 打开路径规划 |
| 22 | （地点）路况 |
| 23 | （地点）的路况 |
| 24 | （地点）周边的路况 |
| 25 | （地点）周边路况 |
| 26 | （地点）附近的路况 |
| 27 | （地点）附近路况 |

1. 车辆核查指令集

|  |  |
| --- | --- |
| 1 | 拍照查车 |
| 2 | 给我查下眼前这辆车 |
| 3 | 帮我查辆车 |
| 4 | 帮我查一辆车 |
| 5 | 我要查辆车 |
| 6 | 我要查一辆车 |
| 7 | 给我查辆车 |
| 8 | 查辆车 |
| 9 | 车辆查询 |
| 10 | 查车辆信息 |
| 11 | 查车 |
| 12 | 帮我查辆车车牌号是（车牌号） |
| 13 | 帮我查一辆车车牌号是（车牌号） |
| 14 | 我要查辆车车牌号是（车牌号） |
| 15 | 我要查一辆车车牌号是（车牌号） |
| 16 | 给我查辆车车牌号是（车牌号） |
| 17 | 查辆车车牌号是（车牌号） |
| 18 | 车辆查询车牌号是（车牌号） |
| 19 | 查车辆信息车牌号是（车牌号） |
| 20 | 查车车牌号是（车牌号） |
| 21 | 帮我查辆车车牌号码是（车牌号） |
| 22 | 帮我查一辆车车牌号码是（车牌号） |
| 23 | 我要查辆车车牌号码是（车牌号） |
| 24 | 我要查一辆车车牌号码是（车牌号） |
| 25 | 给我查辆车车牌号码是（车牌号） |
| 26 | 查辆车车牌号码是（车牌号） |
| 27 | 车辆查询车牌号码是（车牌号） |
| 28 | 查车辆信息车牌号码是（车牌号） |
| 29 | 查车车牌号码是（车牌号） |

1. 打开应用指令集

|  |  |
| --- | --- |
| 1 | 帮我找一下（如：智搜） |
| 2 | 找一下（如：智搜） |
| 3 | 搜索（如：智搜） |
| 4 | 我要用（如：智搜） |
| 5 | 我想用（如：智搜） |
| 6 | 我想使用（如：智搜） |
| 7 | 我要打开（如：智搜） |
| 8 | 帮我打开（如：智搜） |
| 9 | 打开（如：智搜） |
| 10 | （如：民宿查询） |
| 11 | 打开（如：民宿查询） |

1. 人像应用指令集

|  |  |
| --- | --- |
| 1 | 用人像比对 |
| 2 | 打开摄像头 |
| 3 | 用人像比对,查个人 |
| 4 | 用人像比对查下眼前这个人 |
| 5 | 打开摄像头查个人 |
| 6 | 打开摄像头查下眼前这个人 |
| 7 | 人像识别查人 |
| 8 | 人像查人 |
| 9 | 帮我看下这人是谁 |
| 10 | 帮我看下这个人是谁 |
| 11 | 帮我识别下这人是谁 |
| 12 | 帮我识别下这个人是谁 |
| 13 | 帮我识别下这个人是谁 |
| 14 | 帮我用人像识别查人 |
| 15 | 帮我用人像识别查个人 |
| 16 | 调用人像识别查个人 |
| 17 | 我要用人像识别查人 |
| 18 | 我要用人像识别查个人 |
| 19 | 我想查下眼前这个人 |
| 20 | 帮我查下眼前这个人 |

1. 地图控制指令集

支持打开/关闭地图、地图定位、地图移动、地图缩放等指令集。

1. 打开/关闭地图：

|  |  |
| --- | --- |
| 1 | XX市地图 |
| 2 | 打开xx市的地图 |
| 3 | 打开地图 |
| 4 | 重新打开xx市地图 |
| 5 | 再次打开xx市地图 |
| 6 | 那个xx市地图 |
| 7 | 看下xx市地图 |
| 8 | 直接进入地图界面 |
| 9 | 搞到xx市地图 |
| 10 | 转到xx市地图 |
| 11 | 把地图调出来 |
| 12 | 给我把地图搞出来 |
| 13 | 打开xx地图 |

1. 地图定位：

|  |  |
| --- | --- |
| 1 | 定位到XX分局 |
| 2 | 定位到XXX派出所 |
| 3 | 定位到XXXX(重点部位,如外滩\人广\陆家嘴\豫园\迪斯尼\虹桥枢纽\浦东机场) |

1. 地图移动：

|  |  |
| --- | --- |
| 1 | 上/下/左/右移100米 |
| 2 | 向上/下/左/右 |
| 3 | 往上/下/左/右移 |
| 4 | 向上/下/左/右移动 |
| 5 | 向上/下/左/右移一点 |
| 6 | 往上/下/左/右一点 |
| 7 | 向上/下/左/右500米 |
| 8 | 往上/下/左/右边移动 |

1. 地图缩放：

|  |  |
| --- | --- |
| 1 | 放大 |
| 2 | 放大一点 |
| 3 | 搞大一点 |
| 4 | 放大三倍 |
| 5 | 放大五倍 |
| 6 | 缩小 |
| 7 | 搞小一点 |
| 8 | 缩小一倍 |
| 9 | 缩小到最小 |
| 10 | 放大到最大 |
| 11 | 缩小五倍 |
| 12 | 放大两倍查看 |
| 13 | 把当前地图放大2倍 |
| 14 | 继续发大 |
| 15 | 接着放大 |
| 16 | 再放大2倍 |
| 17 | 放大一点点 |
| 18 | 放大一点呗 |
| 19 | 放大一些 |
| 20 | 放大到街道 |
| 21 | 缩小到行政区划 |

1. 摄像头搜索指令集

支持打开/关闭某地摄像头、摄像头查找等指令集

1. 打开关闭摄像头：

|  |  |
| --- | --- |
| 1 | 打开XXXX（地址地名、路名、街道名）监控 |
| 2 | 我想看看XXX监控 |
| 3 | 把XXX监控打开 |
| 4 | 把XXX监控给我打开 |
| 5 | 看看XXX监控 |
| 6 | 给我打开XXX监控 |

摄像头查找：

|  |  |
| --- | --- |
| 1 | 查找XXXX最近的摄像头 |
| 2 | 查找XXXX最近的四路摄像头 |
| 3 | 查找XXXX最近的九路摄像头 |
| 4 | (当前视野内中心点)附近的监控是什么个情况 |
| 5 | 看看周边有多少个摄像头 |
| 6 | 看看周边有几个摄像头 |
| 7 | 看一下百花井（XXX，地址地名）附近的摄像头 |
| 8 | 搜索步行街（XXX，地址地名）的摄像头 |
| 9 | 搜索杏花公园（XXX，地址地名）附近的摄像头 |
| 10 | 长江路（XXX，路名）和青阳路（XXX，路名）交叉口的摄像头 |
| 11 | 看一下三孝口（XXX，地址地名）周围的监控 |
| 12 | 搜一下徽州大道（XXX，路名）和紫云路（XXX，路名）路口的监控 |
| 13 | 查一下广西路（XXX，路名）和南京路（XXX，路名）交叉口周围的监控 |
| 14 | 调一下上海市南门小学（XXX，地址地名）的监控 |
| 15 | 宿松路（XXX，路名）和锦绣大道（XXX，路名）路口东南角的监控 |
| 16 | 蒙城路（XXX，路名）和阜南路（XXX，路名）交口东北角的摄像头 |
| 17 | 包河区政府（XXX，地址地名）周边一百米内的摄像头 |
| 18 | 上海市公安局（XXX，地址地名）周边五十米内的摄像头 |
| 19 | 芜湖路万达（XXX，地址地名）周边两百米内的摄像头 |
| 20 | 天鹅湖畔（XXX，地址地名）南门附近一公里的摄像头 |
| 21 | 看一下五十中（XXX，地址地名）向西五十米（XXX米）处的摄像头 |
| 22 | 查看龙川路（XXX，路名）和宿松路（XXX，路名）路口西北角的摄像头 |
| 23 | 长江路（XXX，路名）和青阳路（XXX，路名）交叉口右边的摄像头 |
| 24 | 跳到彩虹路（XXX，路名）与文曲路（XXX，路名）交叉口的摄像头 |
| 25 | 快速打开祁门路（XXX，地址地名）附近的摄像头 |
| 26 | 调出淮河路步行街（XXX，地址地名）的实时画面 |
| 27 | 给我看下四牌楼（XXX，地址地名）附近的实时路况 |
| 28 | 给我调取合工大南门（XXX，地址地名）的实时情况 |
| 29 | 给我展示上海南站（XXX，地址地名）的情况 |
| 30 | 调取市政府方向的摄像头 |
| 31 | 那个火车站（XXX，地址地名）的监控是怎么样的 |
| 32 | 调一下上海市政府（XXX，地址地名）周围的监控 |

1. 警力搜索指令集

支持打开/关闭警力页面、查看警力等指令集。

|  |  |  |
| --- | --- | --- |
| 1 | 一级操作 | 二级操作 |
| 2 | 查看警力概览 | 打开/关闭警力概览 |
| 3 | 查看全市全部警力 |
| 4 | 查看全市800M警员 |
| 5 | 查看全市运钞车 |
| 6 | 查看全市GPS巡逻车 |
| 7 | 查看全市直升机 |
| 8 | 查看全市警力分布 | 查看XX区的警力分布（各区的） |
| 9 | 查看全市警力统计 | 查看全市在线警力 |
| 10 | 查看全市离线警力 |
| 11 | 查看全市警力 |
| 12 | 查看警员/运钞车/巡逻车/直升机详细信息 |  |

### 文本分析建模系统

公安认知生产平台以要素抽取、关系抽取、文本分类能力服务为输出，提供一套面向非算法人员使用的可私有化部署的交互平台，涵盖项目管理、标签构建、数据标注、模型训练与测试等功能，以任务流形式引导用户便捷地完成能力服务生产。

#### 项目管理

公安认知生产平台以工作流的形式来展示整个项目的创建过程，创建过程共包含：基本信息、标签定义、数据标注和模型训练四个步骤。

1. 基本信息

创建项目时首先需要填写项目基本信息。基本信息部分需要填写项目标题、项目类型、行业需求和需求描述，四项内容均为必填项。

1. 标签定义

标签定义是创建项目的第二个步骤，该模块主要建立不同的标签体系，针对不同的标签体系添加相应的数据以及调整标签的层级。

（1）标签体系

标签体系是标签的集合，内含标注时所需要的标签。

管理员在新建标签体系页面支持按照标签中文名称的普通搜索，同时支持高级搜索。

在新建标签体系页面，左侧的标签为标签池，右侧为标签体系中的标签。标签体系中的标签默认为空，选中标签池中的标签后可以移入到标签体系中。也可以将标签体系中的标签移出到标签池中。此外，可通过搜索框快速搜索指定标签，并且支持同时从中文名称、英文名称、归类等多个维度的高级搜索。

标签池中的标签展示中文名称、标签定义和规律总结三个维度的信息，可点击编辑进行标签修改或查看更多标签信息，若对标签进行修改，修改内容回同步到标签池。

标签体系中的标签展示中文名称、英文名称、协议名称、输出类型、统计类型和备注。协议名称为下拉输入的形式，默认填充常用的历史协议名，下拉展示的内容为该标签历史协议名称。输出类型和统计类型为下拉框形式，默认有填充值，可通过下拉选择。模糊类型的标签可设置统一阈值。

（2）添加数据

标签体系建立完成之后需要添加相应的数据。本地上传数据格式支持.zip、.xml和.txt三种格式的文件，并需要给上传的数据命名。同一标签体系支持多次添加数据，已上传的数据支持删除。

（3）层级调整

层级调整是将标签体系中的标签调整到其所属的层级，以便标注员在进行数据标注时，能够快速找到对应的标签，提高标注效率。

标签体系创建完成后首次打开层级调整时，默认按照标签定义时各标签的所属层级展示。用户可通过拖动标签或拖动层级进行调整，也可以添加新的层级或对已有层级进行修改或删除。

（4）导出

在标签定义模块支持导出标签体系、标注规范和协议文档等文档。

1. 数据标注

标签定义模块完成之后可进入数据标注模块。

（1）是否需标注

数据标注模块中以数据为中心，并以列表的形式展示数据所属的标签体系、数据包含的文档总数和已标注文档数量。对于每条数据需要选择是否需要标注。若选择“否”，则表示该条数据不需要标注，可直接用于模型训练。若选择“是”，则表示该条数据需要标注，不可以直接用于模型训练，需要经过平台分配标注，标注好的数据才可用于模型训练。

（2）分配标注

若数据在“是否需标注”列选择了“是”，即表示该数据需要平台标注。管理员可将该数据对应的标注任务分配给指定的标注员去标注。

在分配标注任务页面，左侧列表包含了管理员创建的所有标注员，右侧列表则是展示该数据对应的标注任务对哪些标注员可见（即，对于单条数据只有选中的标注员才可以进行标注）。右侧已选中标注员列表除了展示标注员名称，还展示对应标注员的已完成标注数量，管理员可查看对应标注员的已提交数据。

此外，在分配标注任务页面还展示了该数据下的总任务数量、已完成标注数量和已跳过标注数量。同时，在分配标注时支持按名称搜索标注员，可快速找到指定的标注员。

（3）查看标注

在选择了是否需要标注之后，可查看数据的标注情况。点击“查看标注”后可查看数据标注详情。标注详情弹窗展示各个标签被标注的份数。

1. 模型训练

（1）添加模型

进行模型训练首先要添加模型。在添加模型时用户需要选择模型训练方案和测试集。测试集可采用系统划分的测试集，采用系统测试集时可修改测试集比例。此外，用户也可以自定义测试集。

（2）查看训练

添加模型并开始运行之后，可点击“查看训练”查看训练状态、训练进度、训练时长和训练日志等信息。

（3）测试

对于任务状态为“已完成”的模型，可以进行测试操作。测试分为两类，一类是效果测试，另一类是体验测试。

1）效果测试

在进行效果测试时，需要进行测试任务的创建。在创建测试任务时，对于所有的标签体系，系统默认全部采用系统测试集（系统测试集下面的开关状态为“开”），用户也可以从本地上传自己的数据作为测试集进行效果测试。

效果测试完成之后可直接查看测试结果（html），也可以将测试结果文件打包下载。

2）体验测试

在进行体验测试时，首先选择需要测试的标签体系，然后直接在左侧的输入框中输入测试文本，点击“确定”之后开始进行测试。最终的测试结果以json的形式展示在右侧的json结果展示区。

（4）修改规则

首次建立模型并训练时，模型处理规则采用的是系统内置的规则。用户可根据测试结果来决定是否需要修改模型处理规则，以达到更满意的效果。需要说明的是，修改规则选项仅限于训练状态为已完成的模型。

（5）下载

当测试结果能够达到预期时，用户可以选择打包下载。下载的内容包含服务包、引擎和资源包。

#### 数据中心

数据中心用于存储用户在创建项目的标签定义模块上传的数据，并以列表的形式展示。通过列表可以看到数据所属的项目、所属标签体系以及数据是原始数据还是用于模型训练的数据等信息。用户可以删除这些数据，也可以将这些数据下载下来并使用这些数据。

#### 配置管理

1. 标签池

标签池内部租户特有的模块，是展示用户建立的所有标签，展示维度包括标签的中英文名称、标签定义、归类、层级、使用记录和可执行操作（编辑、删除）。

（1）新建标签

用户可以在标签池中新建标签。点击“添加标签”打开新建标签弹窗，需要输入标签中英文名称、标签定义、标签层级、规律总结、归类和示例。其中中英文名称和标签定义为必填项，其他内容为选填。

（2）高级搜索

标签池不仅支持按中文名称搜索，同时支持从中文名称、英文名称、归类、标签层级、使用记录、历史协议和规律总结等维度进行高级搜索，可帮助用户快速搜索到想要的标签。

1. 系统配置项

系统配置项中包含在创建项目的过程中所需的各个相关配置项（下拉框）的内容，包括在创建项目时需要用到的项目类型、行业需求、知识场景和数据类型，以及在添加标签时需要用到的归类等。

管理员可以对每个系统配置项包含的具体内容进行编辑。

#### 用户管理

用户管理模块用于对标注员进行管理（增/删/改）。

1. 添加用户（标注员）

管理员可添加用户。管理员需要填写用户名、密码和备注，其中用户名和密码为必填项，且添加用户时有默认的初始密码。

1. 批量导入用户（标注员）

为了能够一次性添加多个用户，公安认知生产平台支持批量导入用户。按照模板要求上传填写好的Excel表格即可。当上传的表格中的存在用户名或密码为空时，会导致整批用户导入失败。

### 应用集成接口

平台提供了标准的接口以及对接文档服务，可以实现对生态合作厂商的应用赋能。