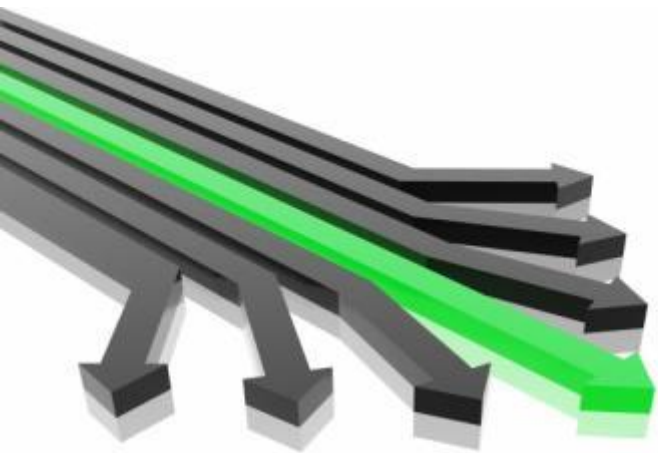


小i机器人在中文语义开放平台的 研究与进展



陈培华
小i机器人
2016年9月·北京

目录



Part 1

公司介绍



Part 2

背景分析

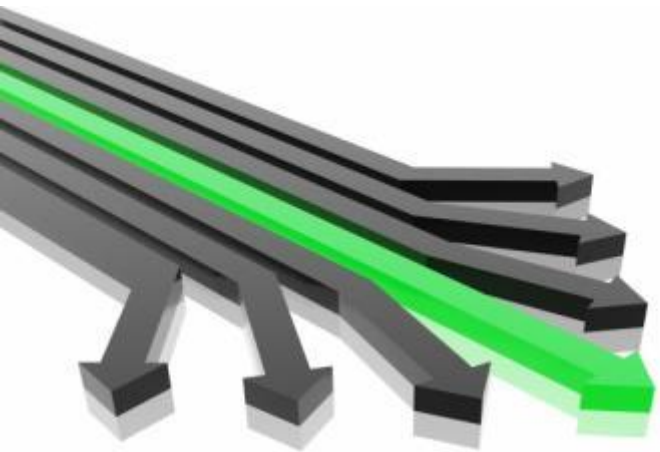


Part 3

小i语义开放平台



目录



Part 1

公司介绍



Part 2

背景分析



Part 3

小i语义开放平台



历程

小i 15年来专注于自然语言理解 (NLU) 人工智能领域

2001

总公司在上海成立



2005

推出短信小机器人，
与诸多运营商及SP
公司开展合作



2006

智能客服解决方案
全球首款政务领域智能客服
机器人“上海科委海德先生”
上线



2008

江苏移动i8智能客服
上线，小i进入运营
商领域



2004

推出MSN小机器人，
影响广泛，奠定小i
发展方向

2006

成为微软机器人全
球战略及技术合作
伙伴

2007

北京办事处成立，
小i业务覆盖中国全
部地区

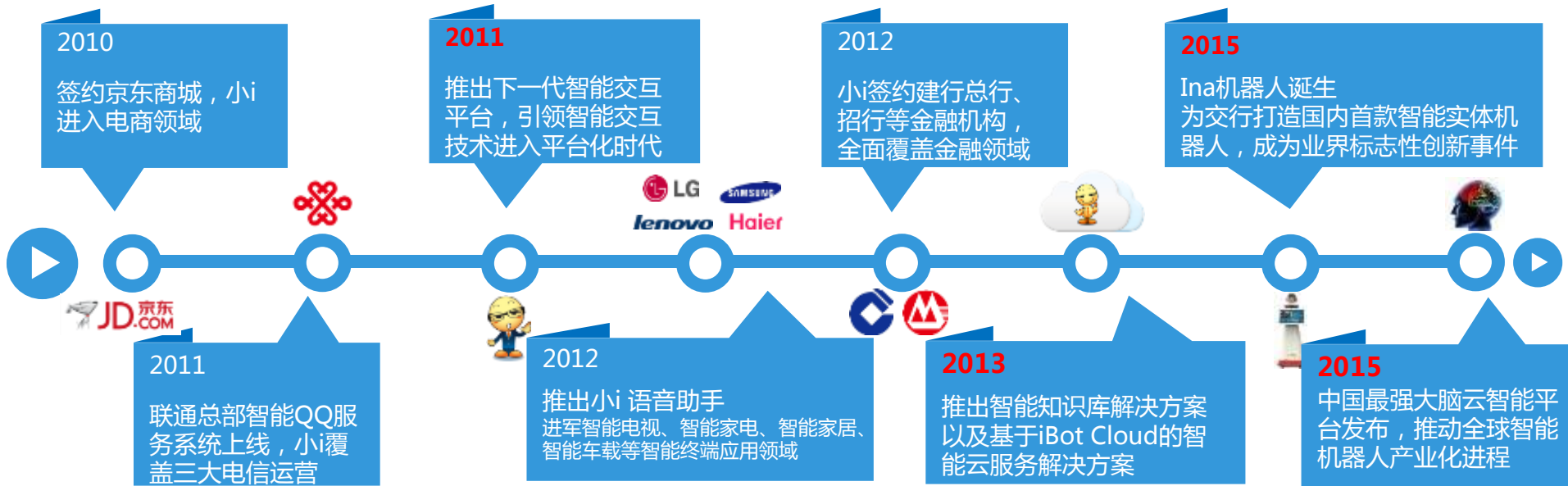
2009

智能营销解决方案
国内首款金融领域智能营销
系统“交通银行点点通”上
线



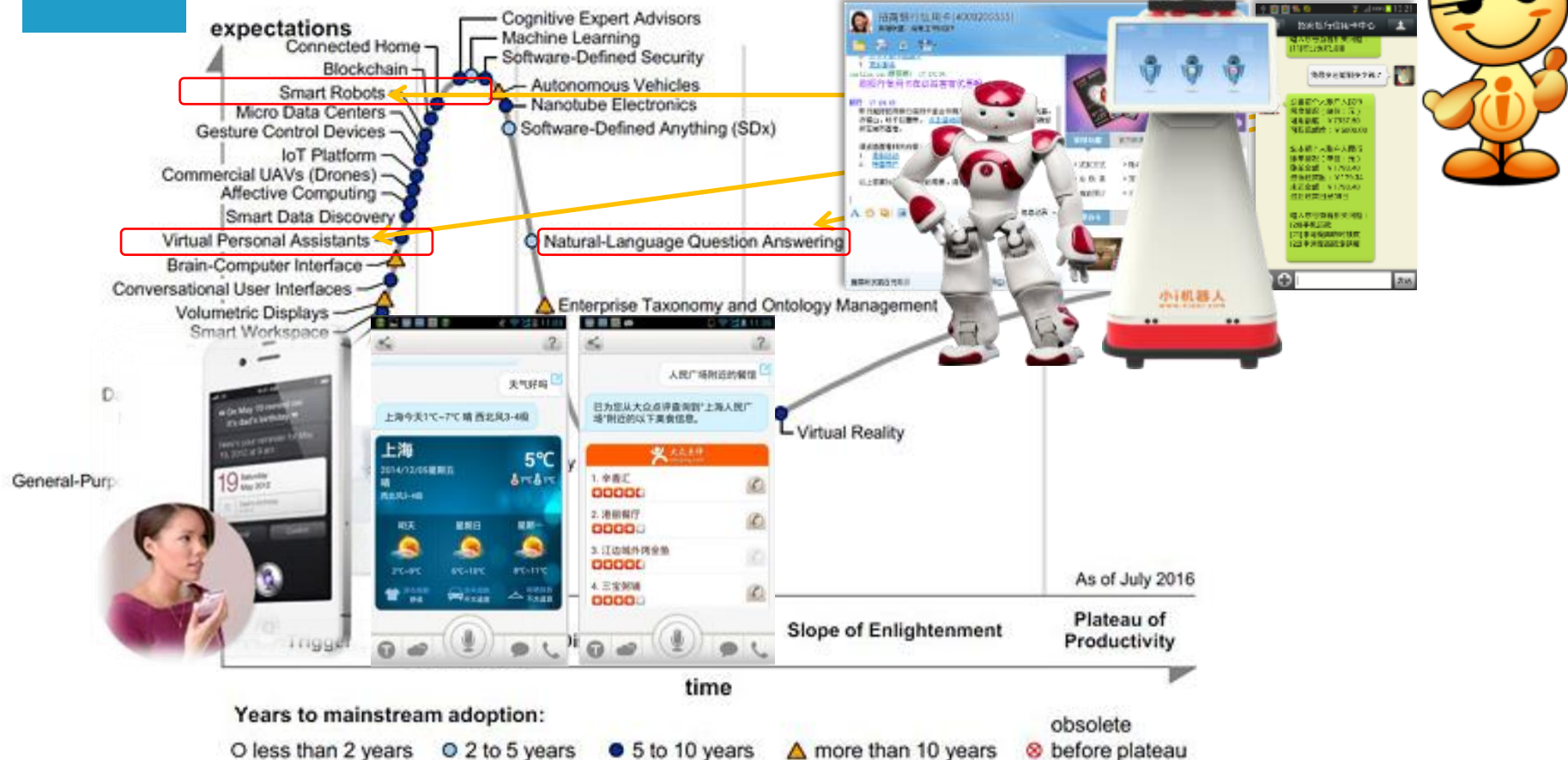
历程

小i 从2C迈向2B，为企业量身定制智能平台



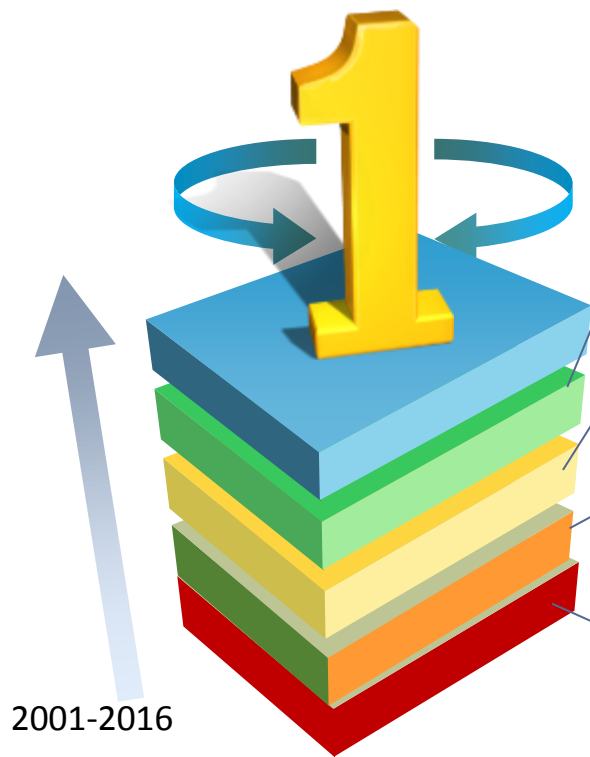
布局

小机器人在人工智能领域中的布局



成就

小i机器人是智能机器人中国第一品牌



全球最大的智能机器人技术提供和平台运营商
服务全球用户超过5亿

拥有最先进的智能人机对话引擎
每年数百亿次的对话交互

拥有多项智能技术关键知识产权
智能客服、人机交互、智能问答、即时通信、短信等

在多个行业沉淀了最大的领域知识库/语义库
通信、金融、电子商务、政府等行业

小i 机器人 — 助力客户打造各行业智能机器人服务标杆

小 i 为每个行业智能交互的**先行者**打造**标杆**！

运营商



金融银行



保险、证券行业



电商平台



航空、汽车等行业



功能列表

小i 机器人—特色功能

 笑话 请输入： 给我讲个笑话/笨孩子	 星座配对 请输入： 星座配对/巨蟹座和双子座	 酒店查询 请输入： 200元左右的酒店/酒	 手机归属地查询 请输入： 查询13800009999是哪	 血型测试 请输入： 血型分析/血型分析	 邮政编码查询 请输入： 北京的邮编/100000是	 购买电影票 请输入： 电影票查询	 号码归属地 请输入： 13818888888的手机号的归属地是哪
 豆瓣服务 请输入： 东周列国志好看吗/黄	 文字闯关 请输入： 我要玩文字闯关	 天气查询 请输入： 查询天气/今天冷么/	 新闻播报 请输入： 新闻/热点新闻/读书新闻	 聊天 请输入： 你好/你叫什么/我今	 股票查询 请输入： 大盘查询/上证指数/福	 购物 请输入： 我要购物/我要买电脑	 福利彩票查询功能 请输入： 福利彩票查询
 心理测试 请输入： 我要玩心理测试	 童年测试 请输入： 我要玩童年测试	 菜谱查询 请输入： 菜谱/给我查查菜谱/	 尺码查询 请输入： 尺码对照表/我身高175厘	 娱乐场所查询 请输入： 附近的KTV/上海火车	 快递查询 请输入： 查询快递/查询单号XXX	 时间查询 请输入： 现在是几点/查询纽约9	 常用电话查询 请输入： 顺丰快递的电话是多少
 星座查询 请输入： 星座运势/双鱼座的运	 诗词 请输入： 描写七夕的诗句/白日	 团购订座 请输入： 团购网吧/团购SPA/	 机票查询 请输入： 机票/9月2日上午机票/上	 网上冲浪 请输入： 打开新闻/打开小说/	 农历查询 请输入： 查询农历/7月23日是农	 日期查询 请输入： 查询美国日期/查询星	 计算器 请输入： 3+2是多少
 花语 请输入： 三色堇代表什么/紫罗	 老黄历查询 请输入： 查询老黄历/今天老黄	 汇率查询 请输入： 汇率查询/美元兑人	 地图查询 请输入： 查询地图/查找工商银行/	 一站到底 请输入： 我要玩一站到底	 火车票查询 请输入： 查询火车票/火车票查	 搜索 请输入： GOOGLE/搜索新闻/首	 翻译 请输入： 苹果的英文是什么
 历史上的今天 请输入： 查询历史上的今天	 前世今生测试 请输入： 我要玩前世今生测试	 汽车资讯查询 请输入： 我想看小型车/我想	 智能计算 请输入： 32的立方是多少是多少	 百科 请输入： 李开复是谁	 单位换算 请输入： 一米等于多少尺/一吨	 首都查询 请输入： 悉尼是哪个国家的城市	
 心理年龄测试 请输入： 我要玩心理年龄测试	 手机号码测吉凶 请输入： 手机号码测试/138888	 电视节目查询 请输入： 查询频道/湖南卫视	 成语查询 请输入： 坐手不群是什么意思	 生日花 请输入： 生日花/生日秘密	 电话区号查询 请输入： 湖南的区号/021是哪	 节日查询 请输入： 圣诞节是几号	
 成语接龙 请输入： 成语接龙		 餐饮查询 请输入： 附近饭店/朝天门附	 寓言故事 请输入： 讲一个寓言故事		 车销限行 请输入： 北京今天车牌限行情况		

系统集成和能力输出层

- 企业级软件产品 / 云平台 and 行业子云 / 操作系统 / 硬件模块



人机交互和模式识别层

- Omni (互联网/移动互联网/终端) / 听觉/视觉/体感/情感



语义理解 and 智能交互层

- 自然语言处理/语义分析和理解/对话管理/知识推理/上下文/个性化



知识和应用层

- 概念/本体/知识图谱/领域语义库 / 对话库/垂直搜索/应用开发框架



智能大数据和学习体系

- 数据分析/垂直爬虫/机器学习/情感分析/知识挖掘



目录



Part 1

公司介绍



Part 2

背景分析

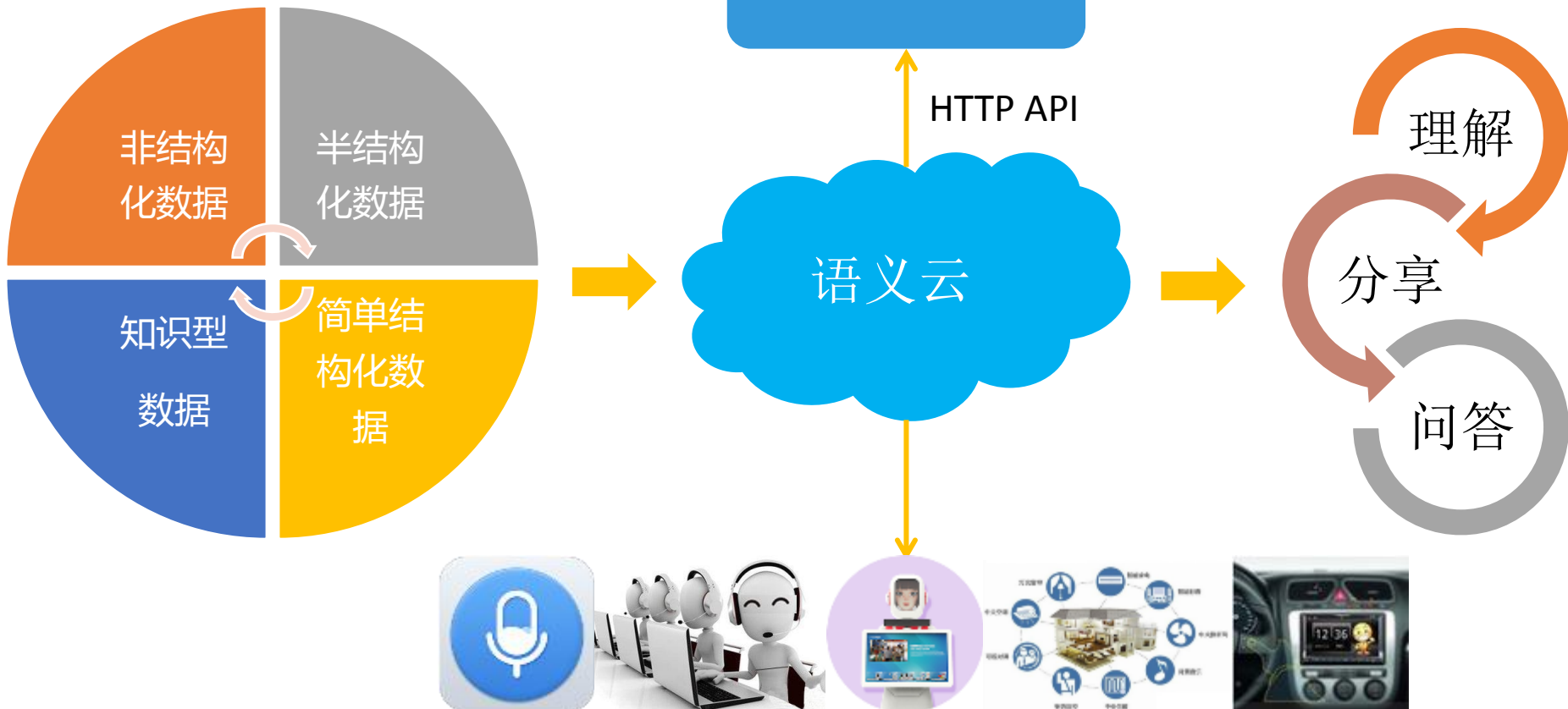


Part 3

小i语义开放平台



语义云概述



商业化现状



腾讯文智中文语义平台



哈工大语言云



科大讯飞开放语义平台

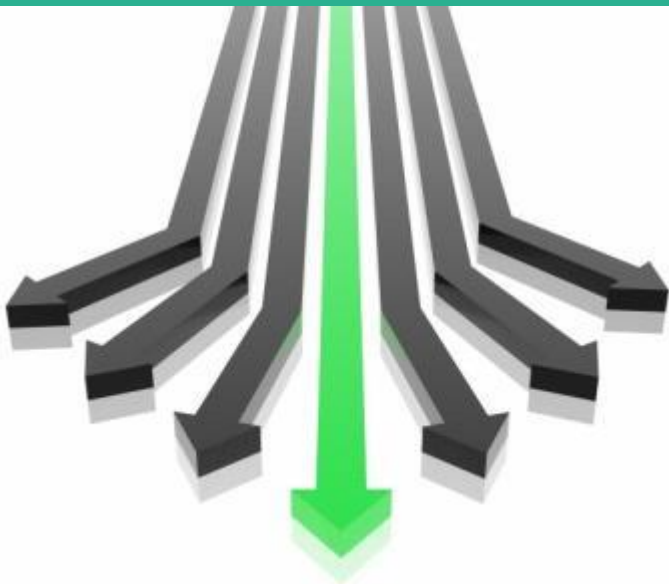


玻森中文语义开放平台

商业化现状

	腾讯文智中文语义平台	哈工大语言云	科大讯飞开放语义平台	玻森中文语义开放平台
分词	√	√	√	√
命名实体识别	√	√	√	√
语义联想	√			√
关键词提取	√			√
自动摘要	√			√
文本纠错	√			√
文本聚类	√			√
文本分类	√			√
情感分析				√
依存句法分析	√	√	√	√
语义角色标注		√	√	
词云				√
网页转码	√			
下载抽取	√			

目录



Part 1

公司介绍



Part 2

背景分析



Part 3

小语义开放平台

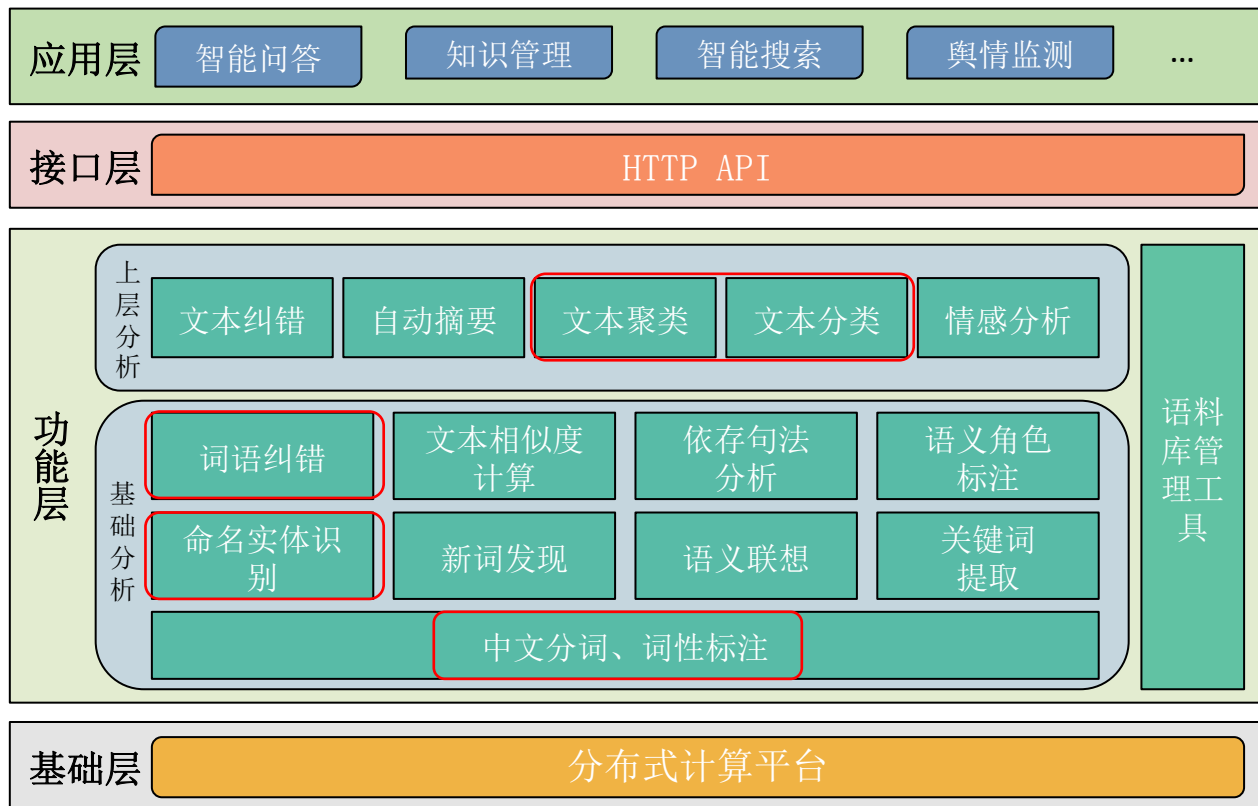


概述

- 小i中文语义开放平台是小i机器人自主设计、研发的一套能够为用户提供多种自然语言处理能力的云服务平台。
- 主要包括：中文分词、命名实体识别、新词发现、词语纠错、关键词提取、自动摘要、文本聚类、文本分类、情感分析等模块，对外提供HTTP API调用接口。
- 采用分布式计算平台和多种机器学习方法。



系统架构



应用层： 小i中文语义开放平台可以面向多个应用系统使用。

接口层： 提供外部调用接口。

功能层： 提供核心的自然语言处理算法和模块。

基础层： 为系统提供分布式存储和计算环境。

核心模块 – 分词与词性标注



小i中文语义开放平台
采用**CRF算法**对标注语料
进行训练学习，将所得到
的模型应用于中文分词和
词性标注中。

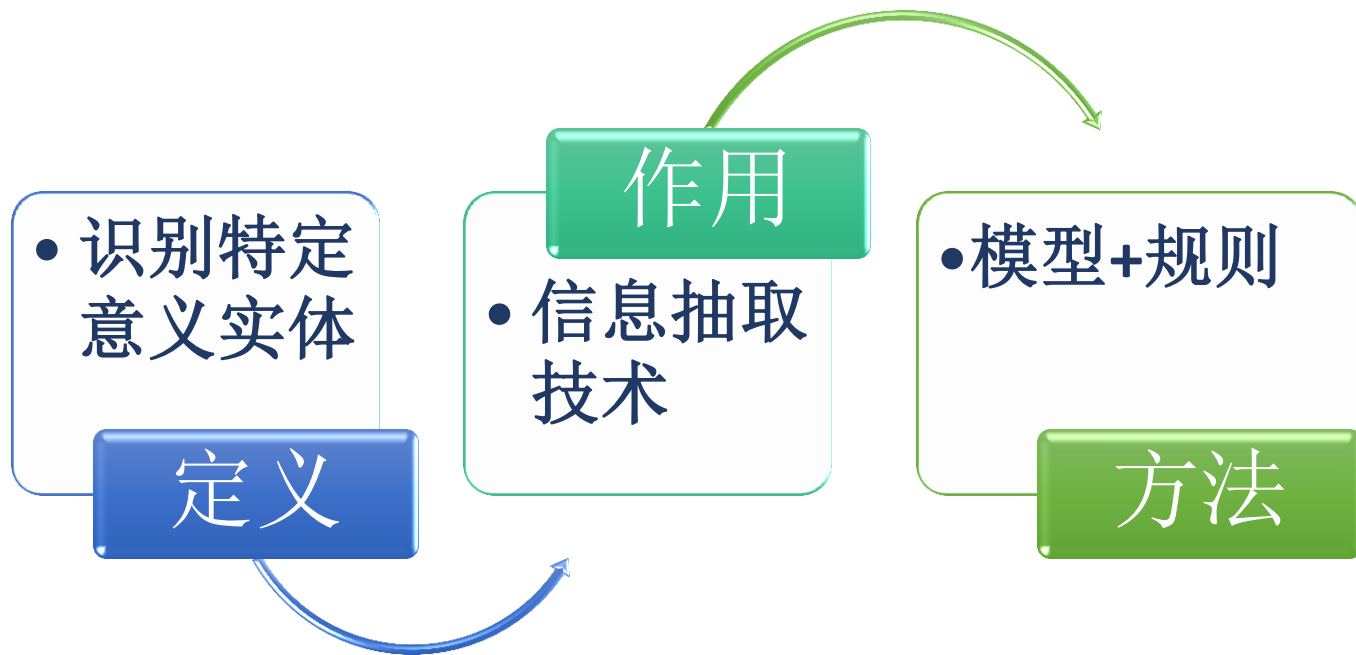
核心模块 – 分词与词性标注

小i中文 分词系统 功能特点

- 
- 1 支持歧义切分处理
 - 2 支持中文词性自动标注
 - 3 支持未登录词识别
 - 4 良好的多编码支持能力
 - 5 提供丰富的知识词典



核心模块 – 命名实体识别

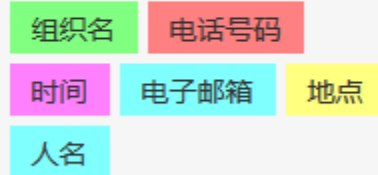


核心模块 – 命名实体识别

实体识别:

"上海贾三思的手机号码是：+8613800000000，固定电话是：021-60000000。"现在是2016年9月2日。"交通银行的免费客服热线为：4008009888。移动的客服电话为：10086";"我的email是：abc@jtu.edu.cn，或者是：abc.1234@tom.com。" + "又或者者是：abc-123@163.com。";

实体类别图示:



☐ 原句 ☐ 分词 ☐ 词性标注 ☒ 命名实体

"上海贾三思的手机号码是：8613800000000，固定电话是：021-60000000。"现在是2016年9月2日。

"交通银行的免费客服热线为：4008009888。移动的客服电话为：10086"

☒ 机构名(Ni) ☒ 人名(Nh) ☒ 地名(Ns)

核心模块 – 词语纠错

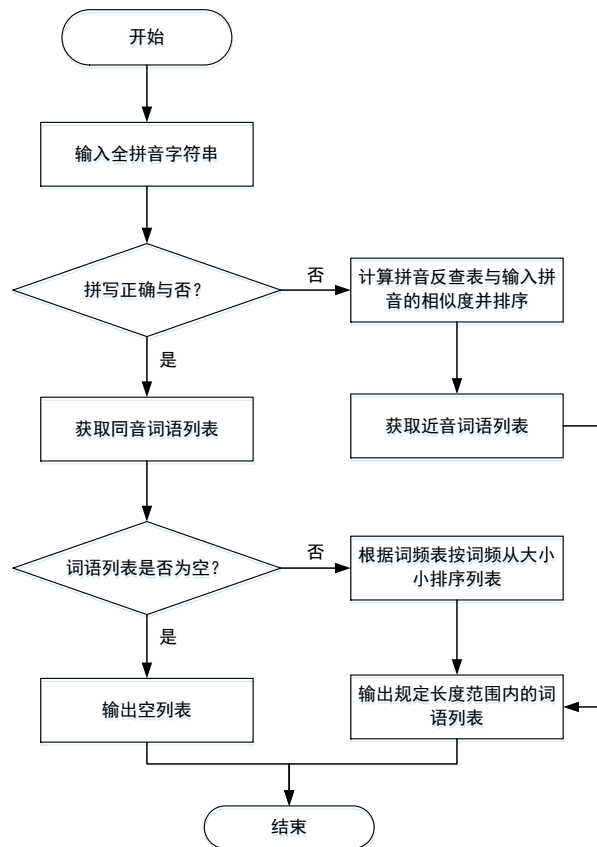
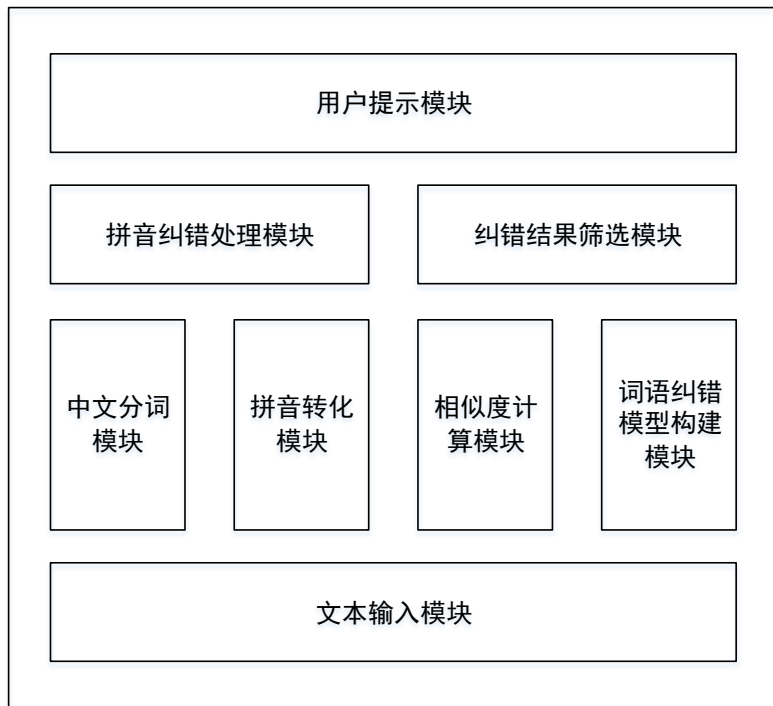


常见错误类型	您输入的词语	您是否要输入
同音别字	事儿生效	十二生肖
近音别字	十而僧小	十二生肖
形近别字	火中取栗	火中取栗
拼音	huozhongquli	火中取栗



核心模块 – 词语纠错

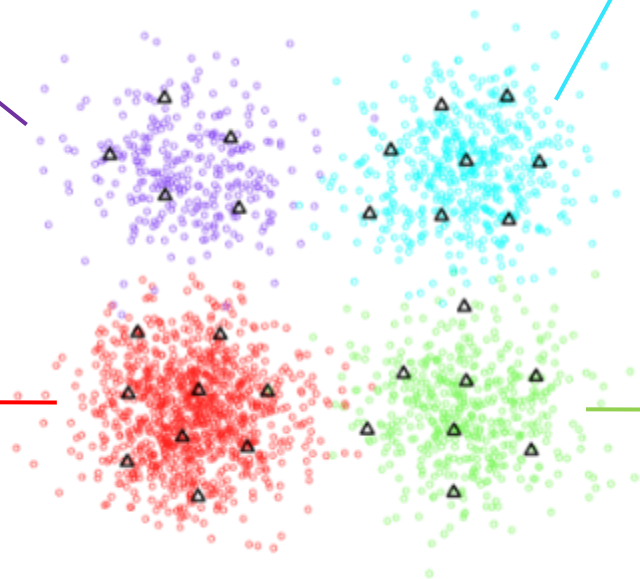
• 词语纠错系统框架及拼音纠错处理方法



核心模块 – 文本聚类

聚类分析是按照一定的规律和要求对事物进行簇划分的过程，是一种无监督分类，没有预定义的先验知识。

在很多现有聚类应用中，K-means方法常被当做“黑盒子”来应用，而未考虑对其聚类结果的验证，即K值通常是由用户凭经验或随机给定。

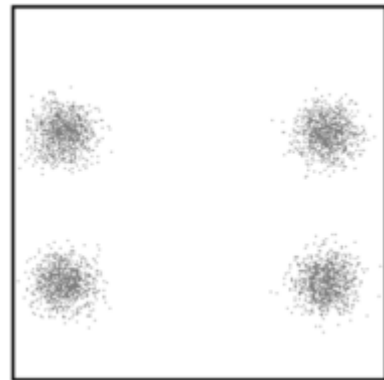
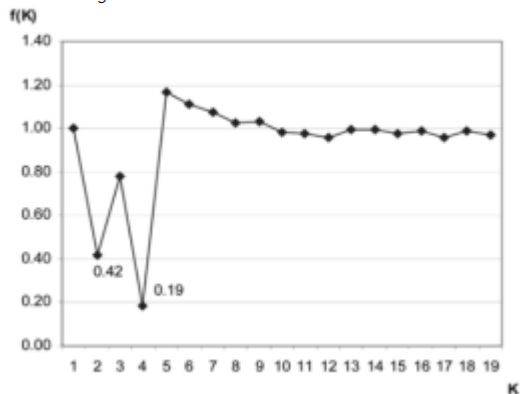
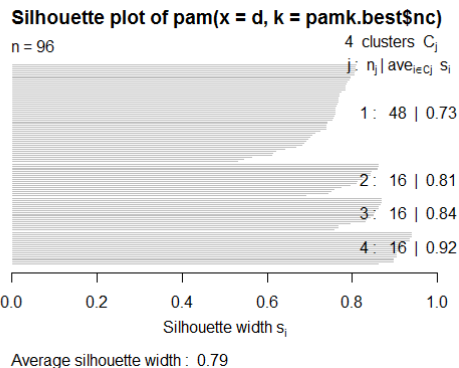
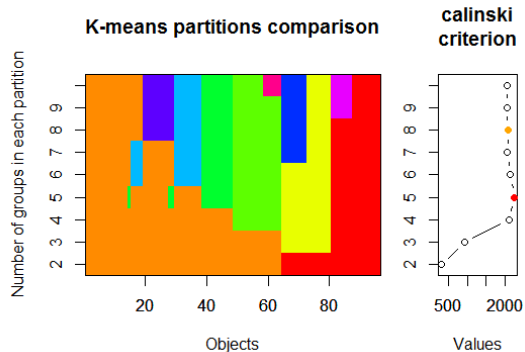
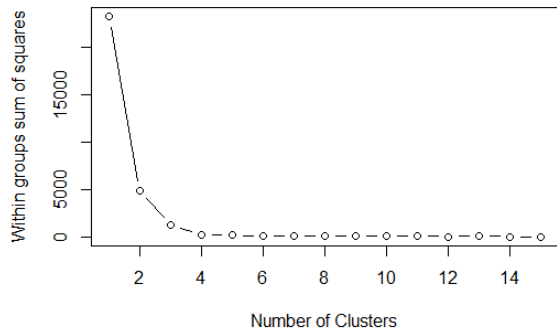


聚类算法有多种，如基于划分、层次、密度、网格、模型等聚类算法。其中应用最为广泛的则为基于划分的K-means算法。

小i机器人采用了{Elkan C., 2003}中所提出的快速K-means聚类方法，使用{Arthur D., Vassilvitskii S., 2007}中的种子选择策略，使用{Pham D T., 2005}中所提出的确定K值的方法，来实现文本聚类系统。

核心模块 – 文本聚类

• K值的确定



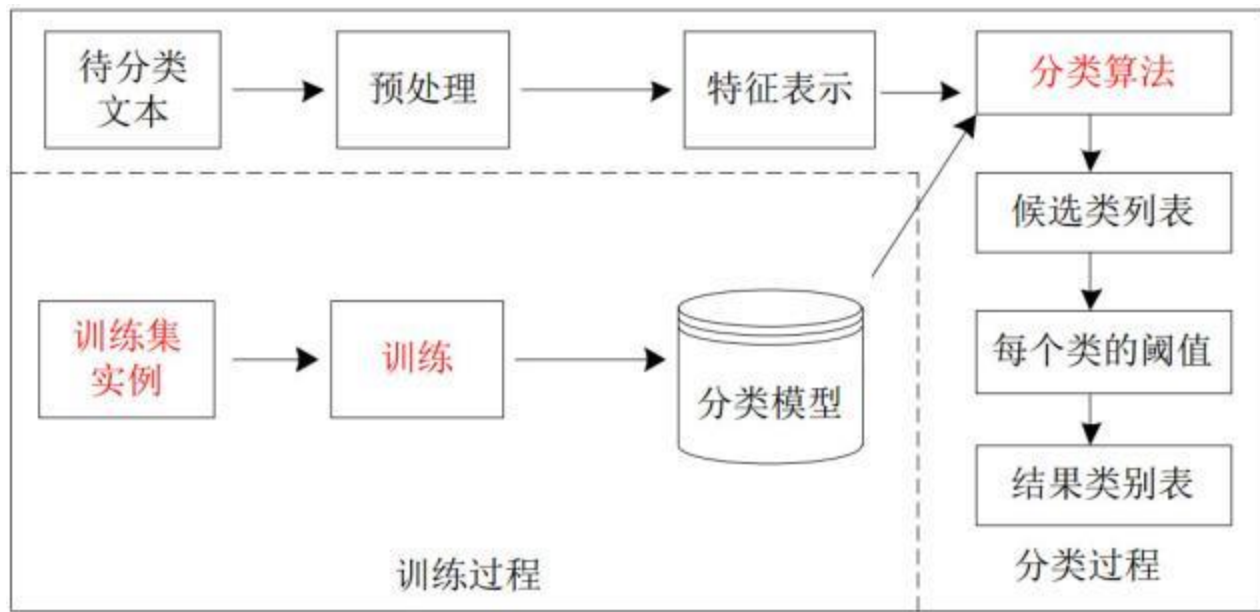
$$f(K) = \begin{cases} 1 & \text{if } K = 1 \\ \frac{S_K}{\alpha_K S_{K-1}} & \text{if } S_{K-1} \neq 0, \forall K > 1 \\ 1 & \text{if } S_{K-1} = 0, \forall K > 1 \end{cases} \quad (2)$$

$$\alpha_K = \begin{cases} 1 - \frac{3}{4N_d} & \text{if } K = 2 \text{ and } N_d > 1 \\ \alpha_{K-1} + \frac{1 - \alpha_{K-1}}{6} & \text{if } K > 2 \text{ and } N_d > 1 \end{cases} \quad (3a)$$

(3b)

核心模块 – 文本分类

- 文本分类是指按照预先定义的主题类别，为文档集中的每篇文档确定一个类别。



核心模块 – 文本分类



基于非线性SVM
的文本分类方法



参考文献

- Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage. <http://www.gartner.com/newsroom/id/3412017>
- 腾讯文智中文语义平台. <http://nlp.qq.com/>
- 哈工大语言云. <http://www.ltp-cloud.com/>
- 科大讯飞开放语义平台. <http://osp.voicecloud.cn/index.html>
- 玻森中文语义开放平台. <http://bosonnlp.com/>
- Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the eighteenth international conference on machine learning, ICML. 2001, 1: 282-289.
- Elkan C. Using the triangle inequality to accelerate k-means[C]//ICML. 2003, 3: 147-153.
- Pham D T, Dimov S S, Nguyen C D. Selection of K in K-means clustering[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.
- Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.
- Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

谢谢！
欢迎联系我们：

地址：上海市金沙江西路1555弄383号1楼
邮政编码：201803
Tel: 8621-39518811

更多精彩内容，敬请关注2016年
世界机器人大会！

