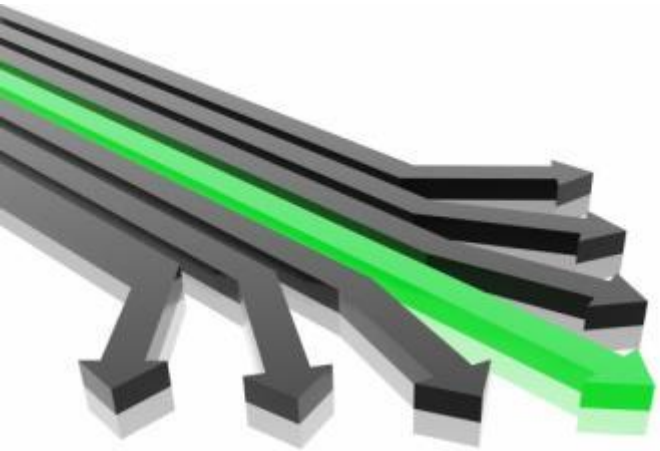
# 小i机器人在中文语义开放平台的研究与进展

陈培华
小i机器人
2016年9月·北京

# 目 录



Part 1

公司介绍

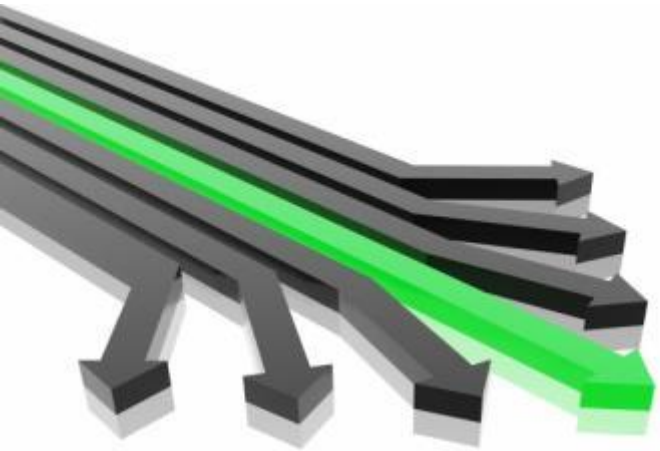
Part 2

背景分析

Part 3

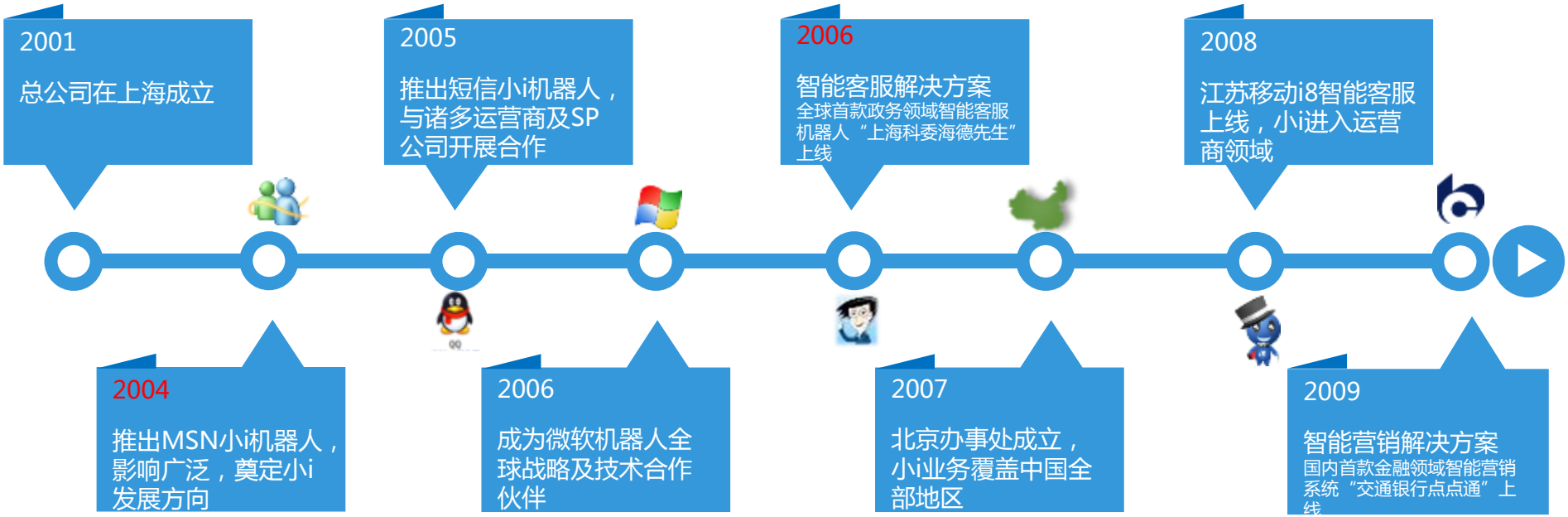
小i语义开放平台

# 目 录



# 历程

## 小i 15年来专注于自然语言理解（NLU）人工智能领域



# 历程

## 小i 从2C迈向2B，为企业量身定制智能平台

**2010**

签约京东商城，小i进入电商领域

**2011**

联通总部智能QQ服务系统上线，小i覆盖三大电信运营

**2011**

推出下一代智能交互平台，引领智能交互技术进入平台化时代

**2012**

推出小i 语音助手

进军智能电视、智能家电、智能家居、智能车载等智能终端应用领域

**2012**

小i签约建行总行、招行等金融机构，全面覆盖金融领域

**2013**

推出智能知识库解决方案以及基于iBot Cloud的智能云服务解决方案

**2015**

Ina机器人诞生

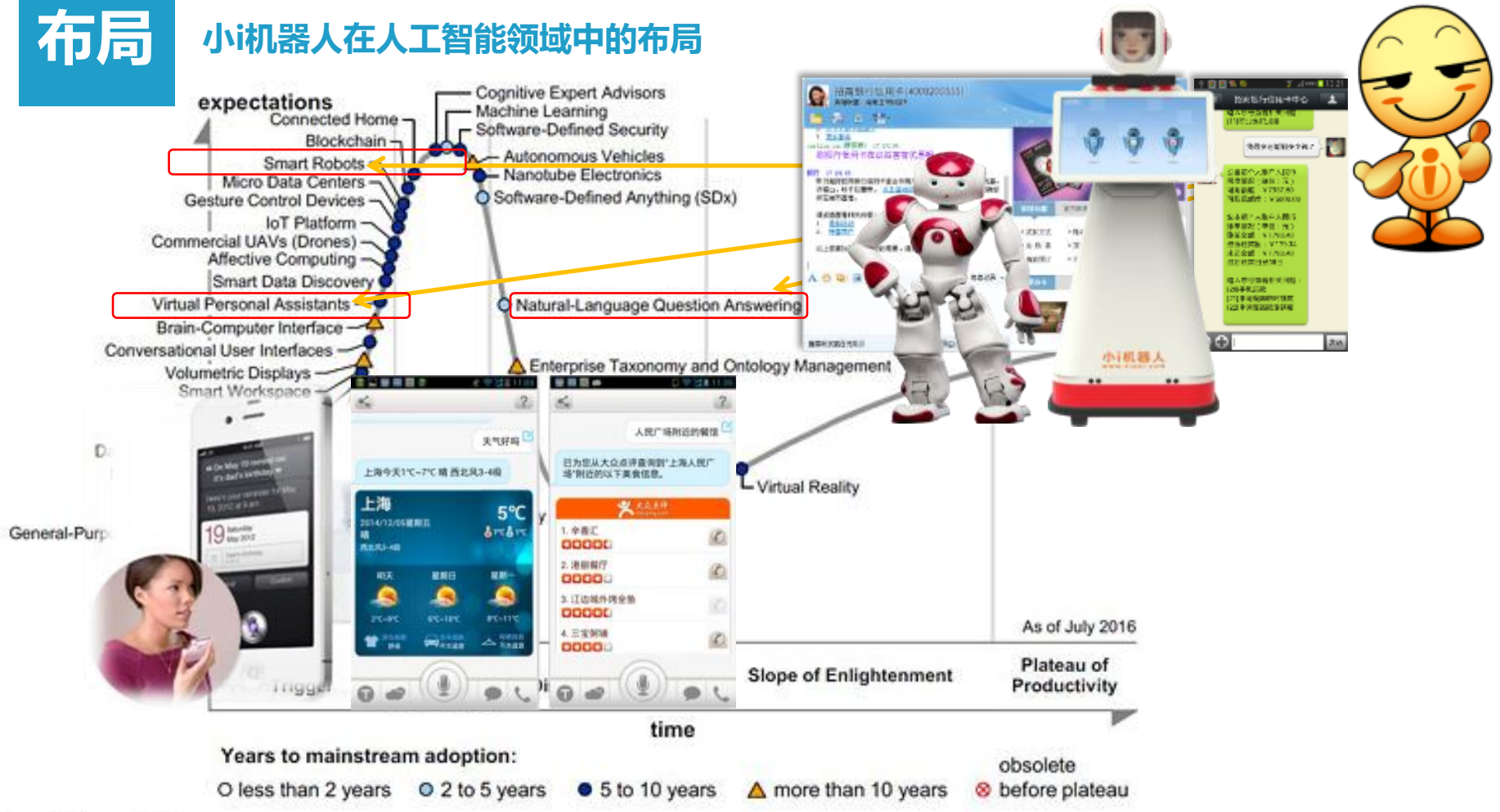
为交行打造国内首款智能实体机器人，成为业界标志性创新事件

**2015**

中国最强大脑云智能平台发布，推动全球智能机器人产业化进程

# 布局

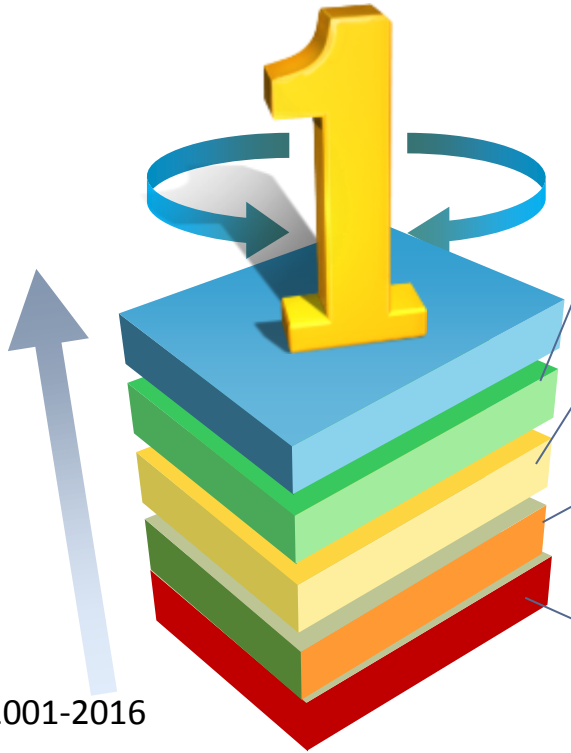
## 小i机器人在人工智能领域中的布局




Source: Gartner (July 2016)

# 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

# 行业标杆

## 小i 机器人 — 助力客户打造各行业智能机器人服务标杆

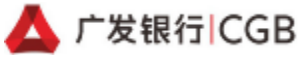
小 i 为每个行业智能交互的先行者打造标杆！

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

# 技术体系

## 系统集成和能力输出层

- 企业级软件产品 / 云平台和行业子云 / 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

## 语义理解和智能交互层

- 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/垂直搜素/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

平台生态

人机交互

**核心能力**

数据壁垒

学习体系

# 目 录

Part 1

公司介绍

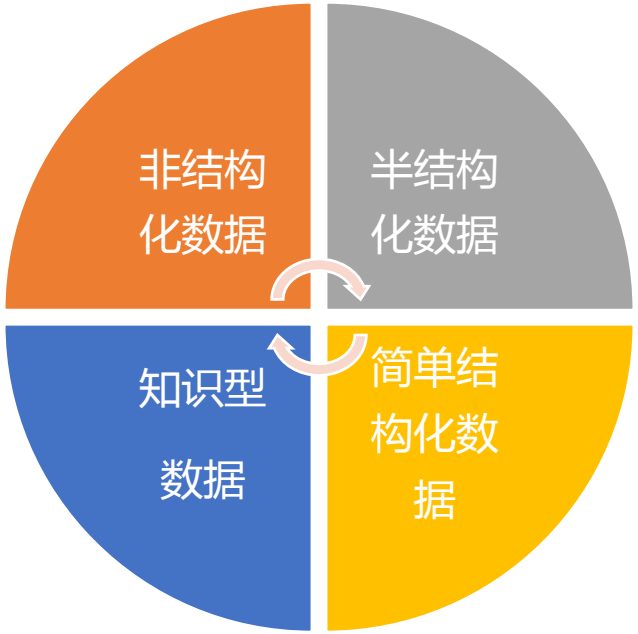
Part 2

背景分析

Part 3

小i语义开放平台

# 语义云概述




自然语言处理

HTTP API

语义云

理解

分享

问答

# 商业化现状

腾讯文智中文语义平台

哈工大语言云

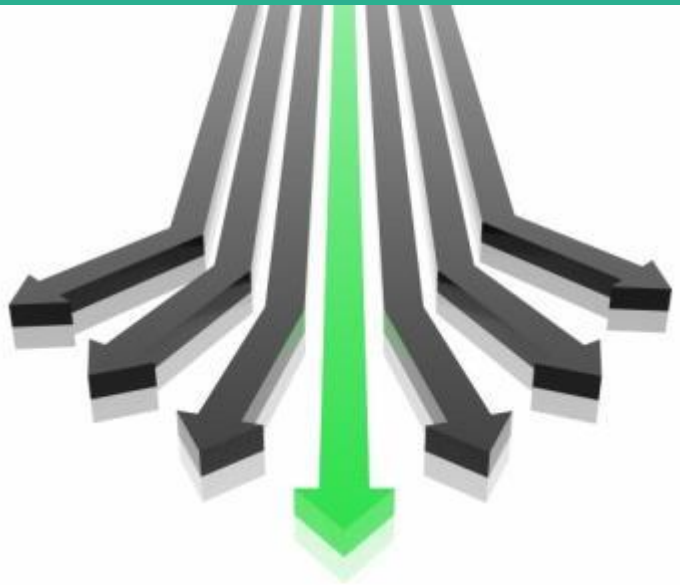
科大讯飞开放语义平台

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

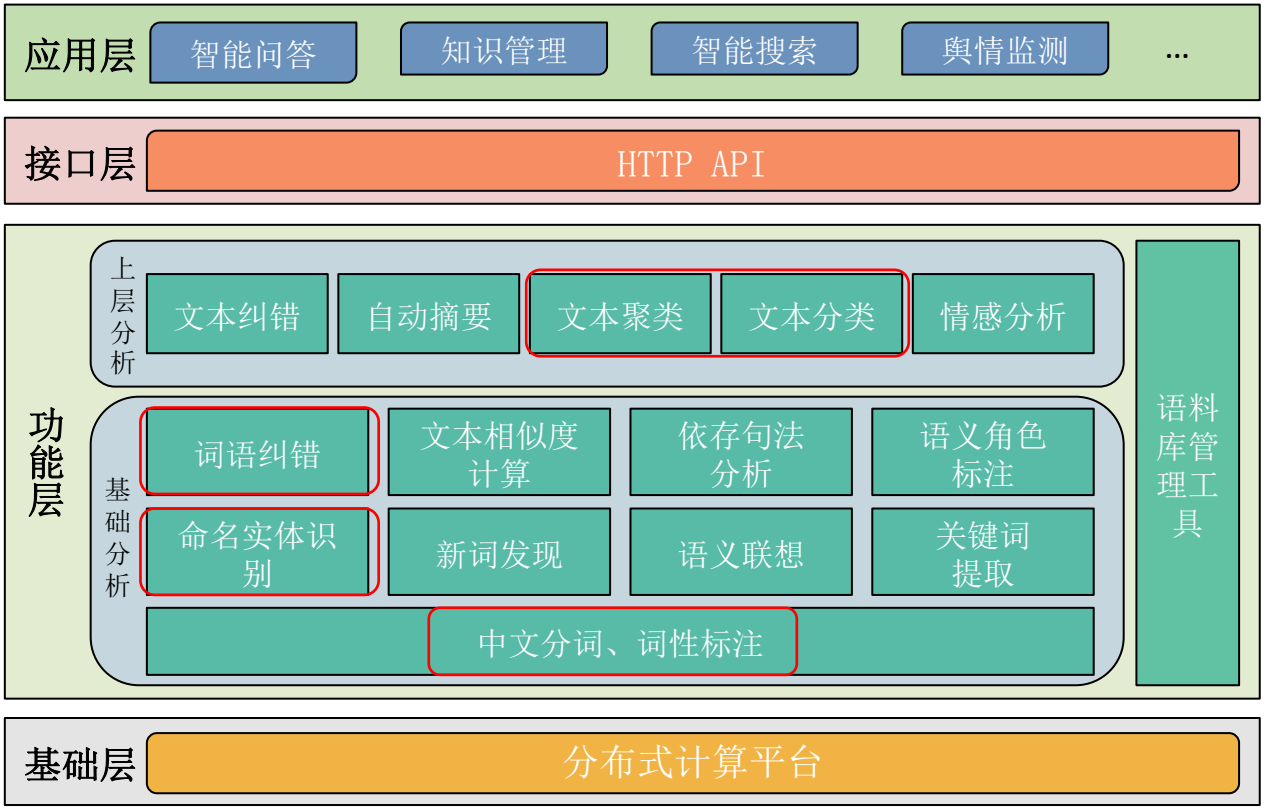
Part 3

小i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



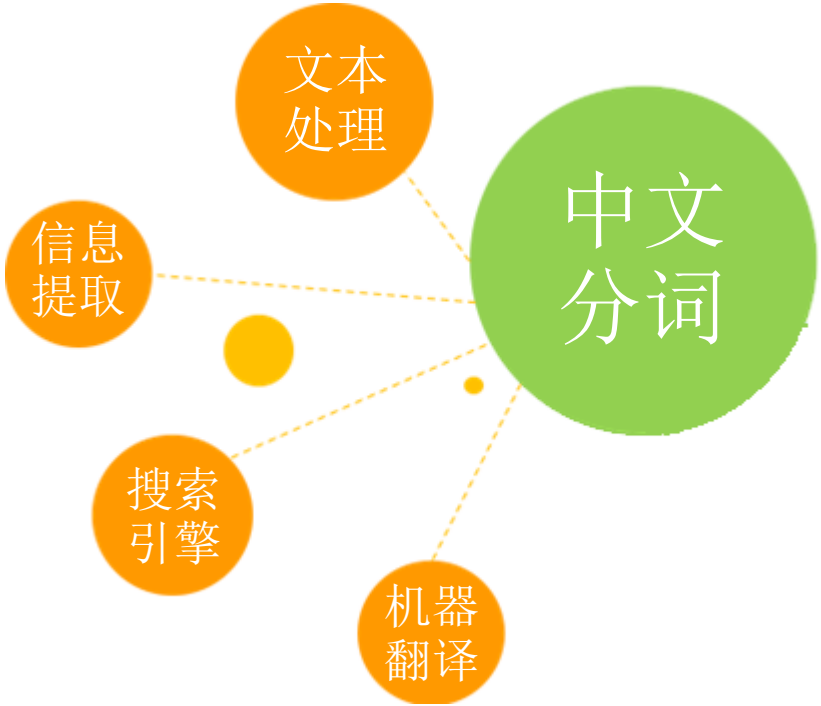

**应用层：**小i中文语义开放平台可以面向多个应用系统使用。

**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



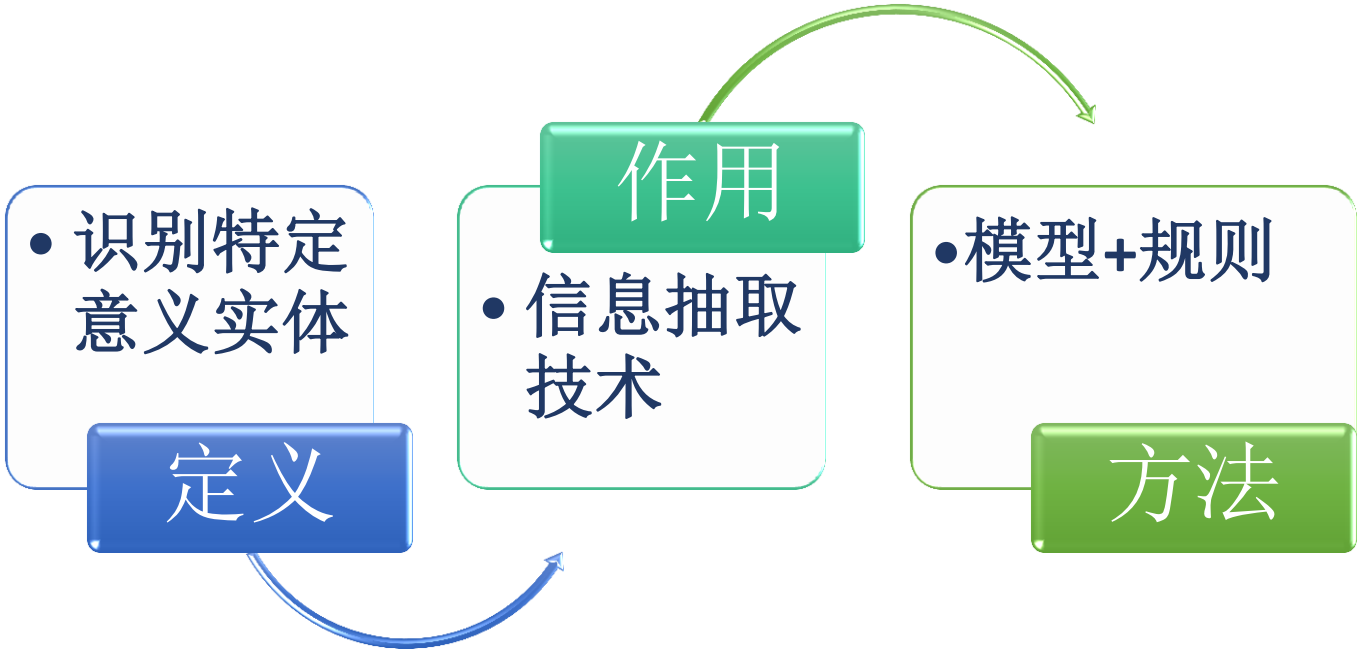

小i中文语义开放平台采用**CRF算法**对标注语料进行训练学习，将所得到的模型应用于中文分词和词性标注中。

# 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。" + "又或者是：abc-123@163.com。";

实体类别图示:

组织名 电话号码 时间 电子邮箱 地点 人名

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"

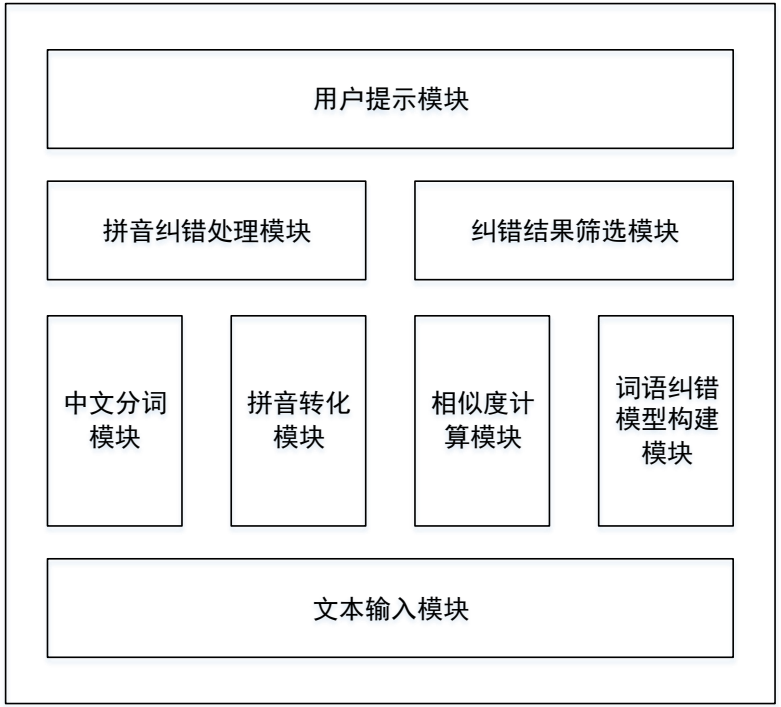
机构名(Ni) 人名(Nh) 地名(Ns)

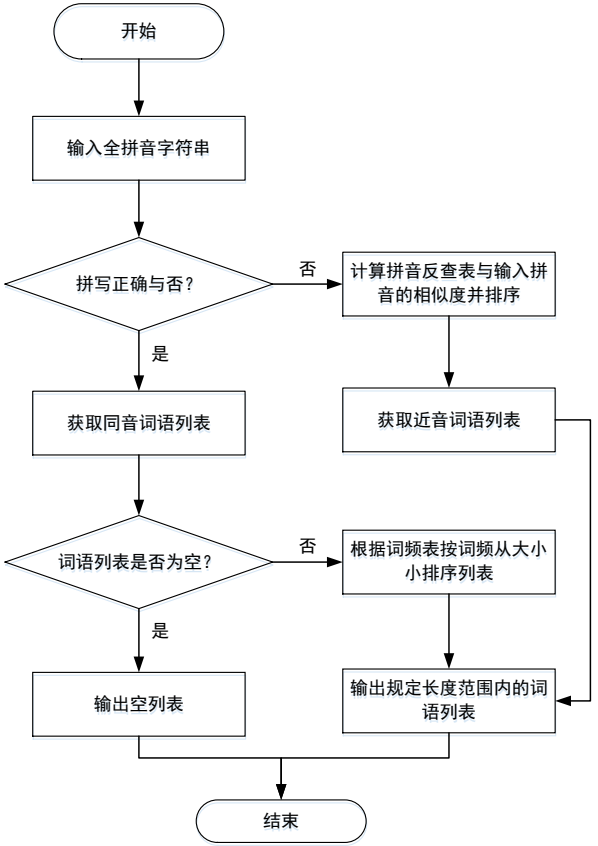
# 核心模块 – 词语纠错

| 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 | 火中取粟 | 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

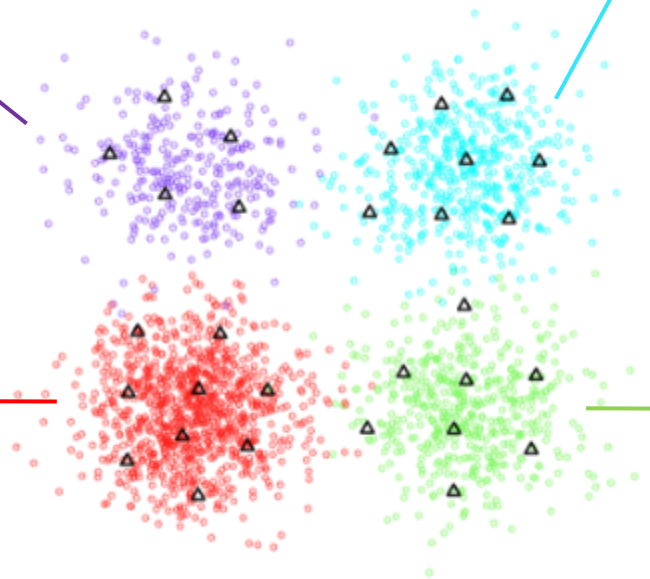


# 核心模块 – 文本聚类

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。

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。

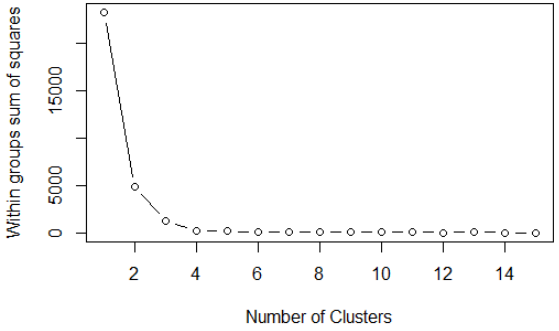


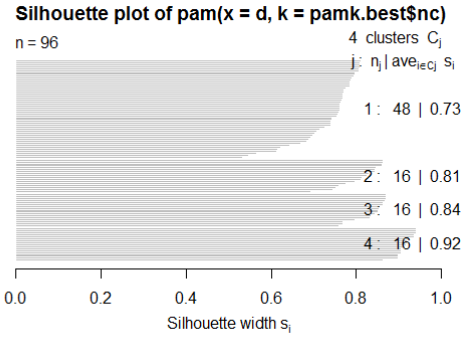
在很多现有聚类应用中，K-means方法常被当做“黑盒子”来应用，而未考虑对其聚类结果的验证，即K值通常是由用户凭经验或随机给定。

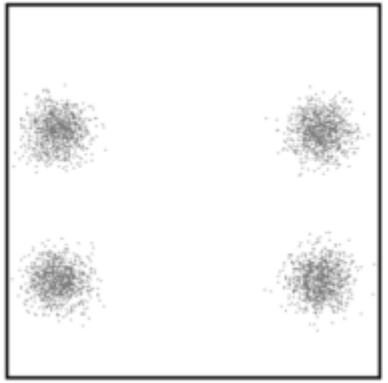
小i机器人采用了{Elkan C., 2003}中所提出的快速K-means聚类方法，使用{Arthur D., Vassilvitskii S., 2007}中的种子选择策略，使用{Pham D T., 2005}中所提出的确定K值的方法，来实现文本聚类系统。

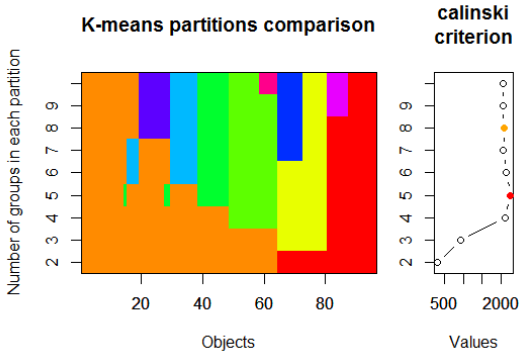
# 核心模块 – 文本聚类

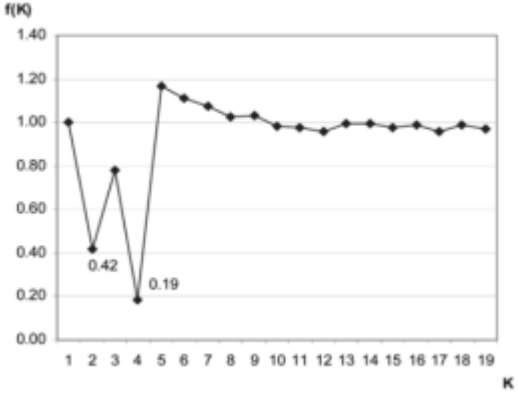
- K值的确定






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

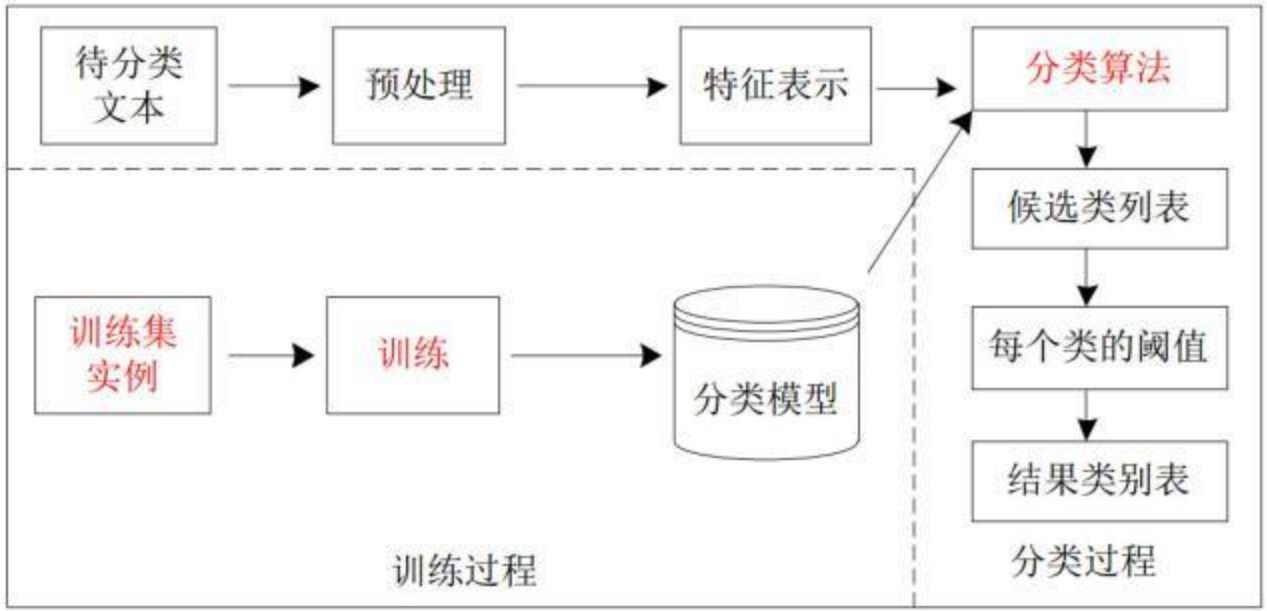
$$\alpha_K = \begin{cases} 1 - \dfrac{3}{4N_d} & \text{if } K = 2 \text{ and } N_d > 1 \quad (3a) \\ \alpha_{K-1} + \dfrac{1 - \alpha_{K-1}}{6} & \text{if } K > 2 \text{ and } N_d > 1 \quad (3b) \end{cases}$$

Ref: http://t.cn/RcyKLU4，{Pham D T., 2005}

Ref: {Rodriguez A, Laio A., 2014}

# 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，为文档集合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！