The 8th China Cloud Computing Conference
第八届中国云计算大会
技术融合 应用创新

自然语言处理在企业应用领域的实践

畅捷通信息技术股份有限公司

李鲲

自然语言处理在企业应用领域的实践

自然语言处理在企业应用领域面临的挑战

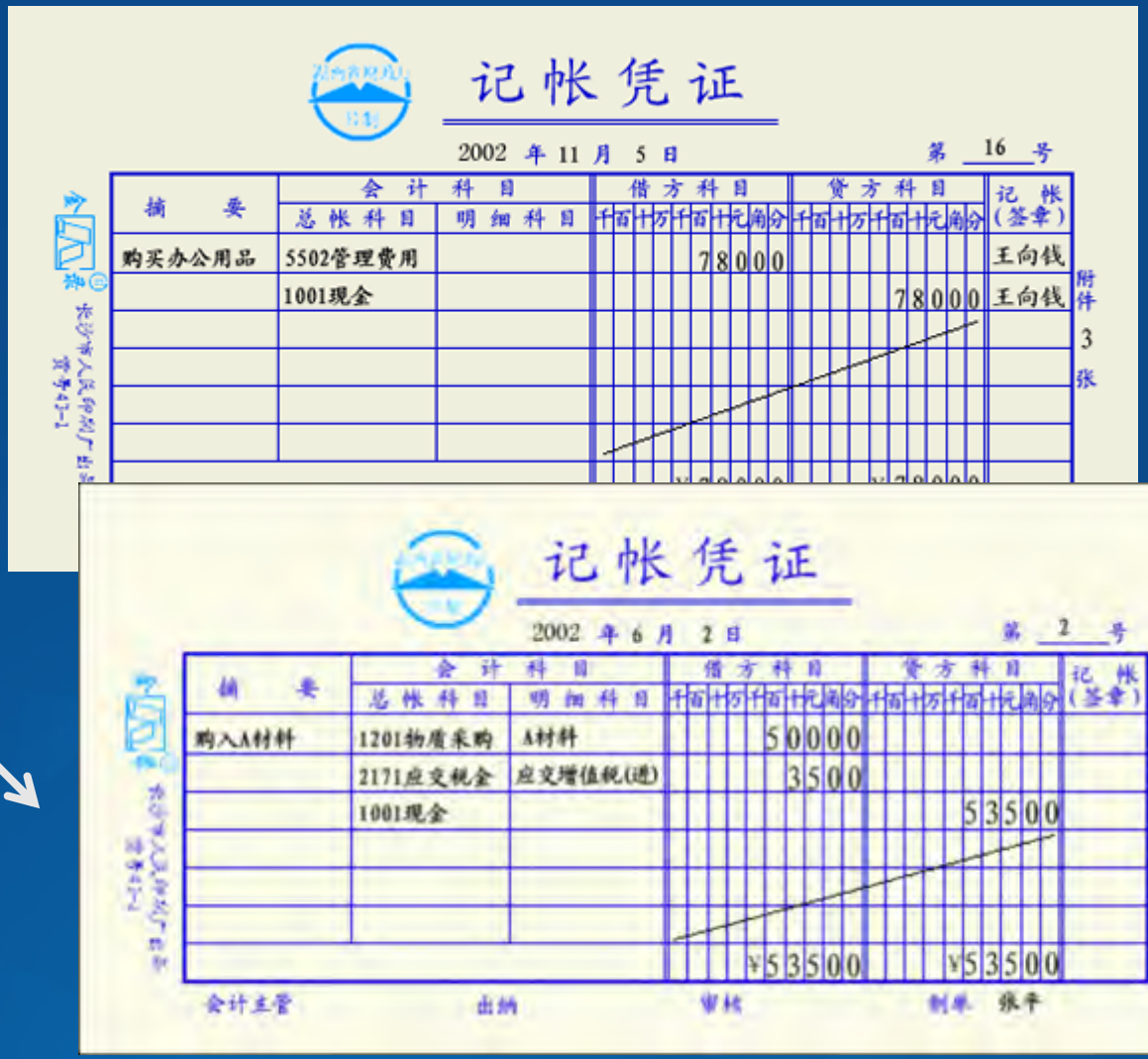
用户语音输入	百度语音识别
花之语杭白菊	话剧行白居
鲜花椒油	新发就咬
海天老抽800ml	黑天脑抽八百毫升
三五麻辣鱼150g	三无码了鱼150克
黎红鲜花椒油330ml	李红仙化交融330好声
黄心猕猴桃	广西猕猴桃
白皮花生	白皮花生

通用语音识别对商品的识别结果示例

自然语言处理在企业应用领域面临的挑战

买办公用品780元

买A材料535元



记帐凭证

2002 年 11 月 5 日 第 16 号

摘要	会计科目 总帐科目	会计科目 明细科目	借方科目	贷方科目	记帐(签章)
购买办公用品	5502管理费用		¥78000		王向钱
	1001现金			¥78000	王向钱

附件 3 张

记帐凭证

2002 年 6 月 2 日 第 2 号

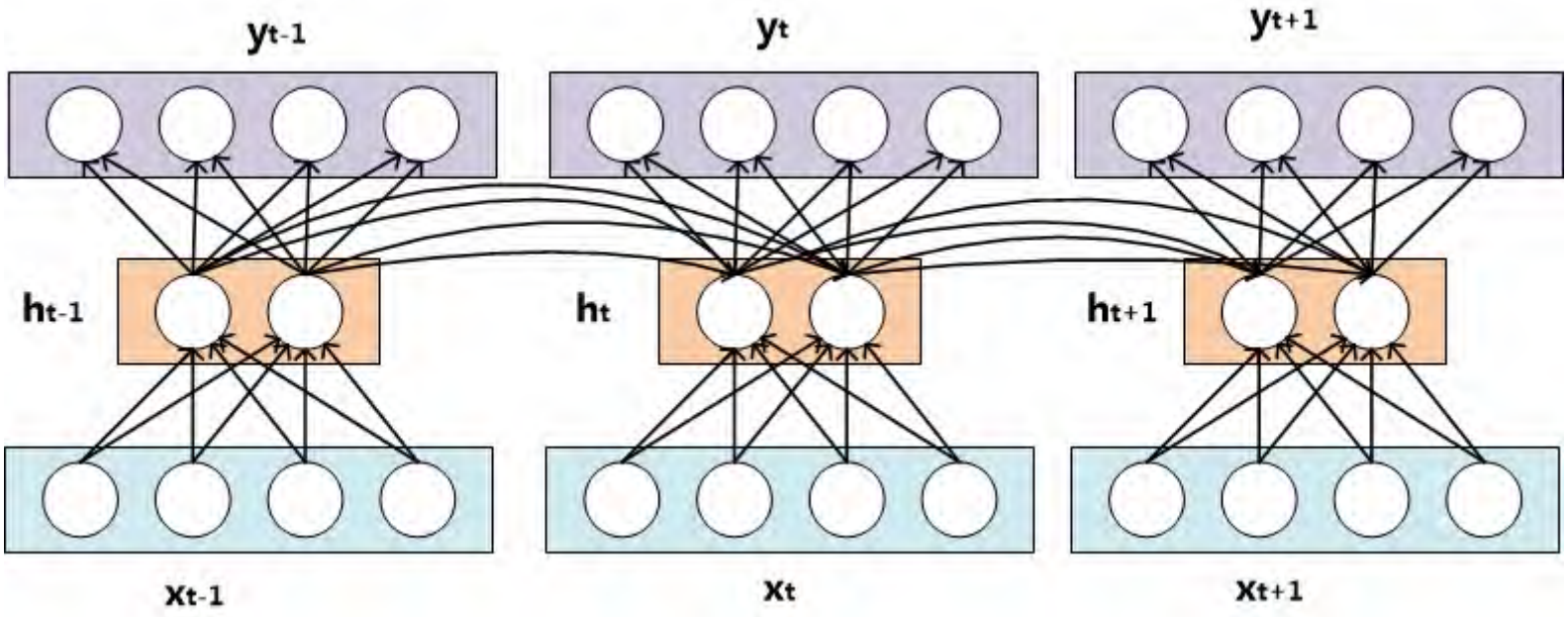
摘要	会计科目 总帐科目	会计科目 明细科目	借方科目	贷方科目	记帐(签章)
购入A材料	1201物质采购	A材料	¥50000		
	2171应交税金	应交增值税(进)	¥3500		
	1001现金			¥53500	
			¥53500	¥53500	

附件 张

会计主管 出纳 审核 制单 张平

自然语言处理在企业应用领域的实践

基于深度学习的财务领域语言模型



$$t_i = W_{hx}x_i + W_{hh}h_{i-1} + b_h$$

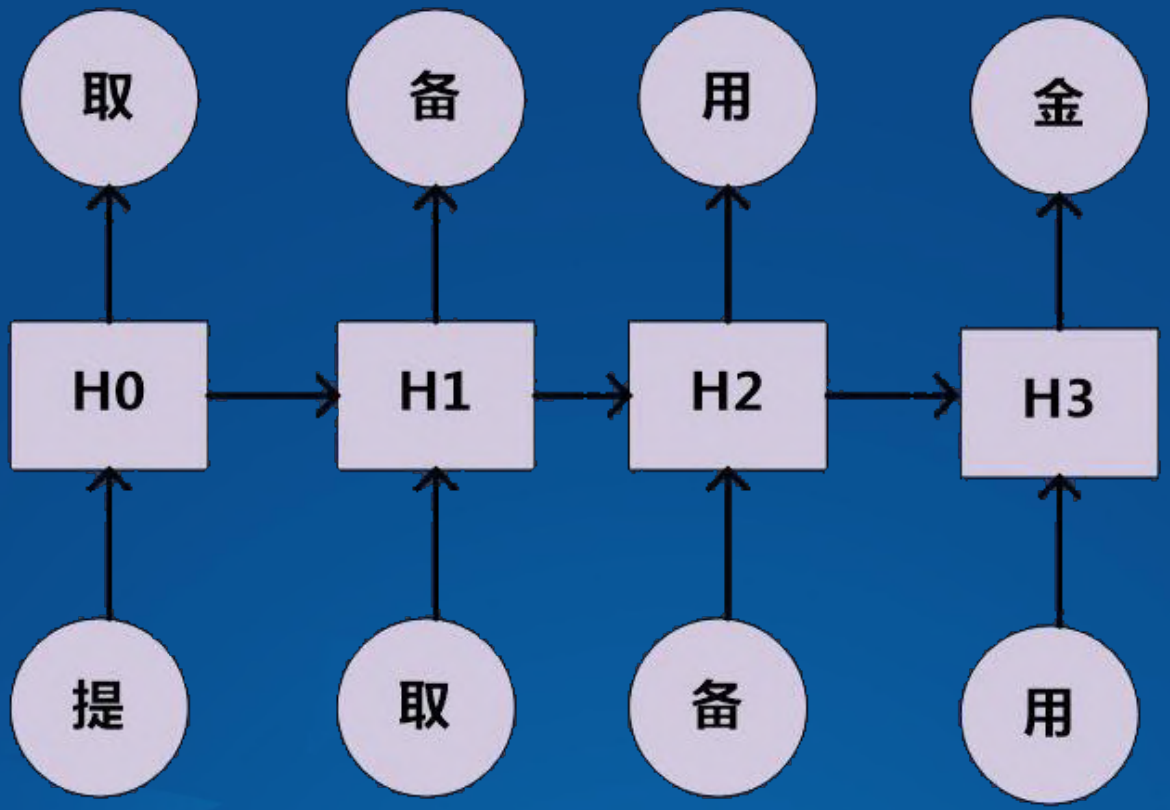
$$h_i = \mathrm{Sigmoid}(t_i)$$

$$s_i = W_{yh}h_i + b_y$$

$$y_i = \mathrm{SoftMax}(s_i)$$

RNN (Recurrent Neural Network)

基于深度学习的财务领域语言模型



RNN语言模型

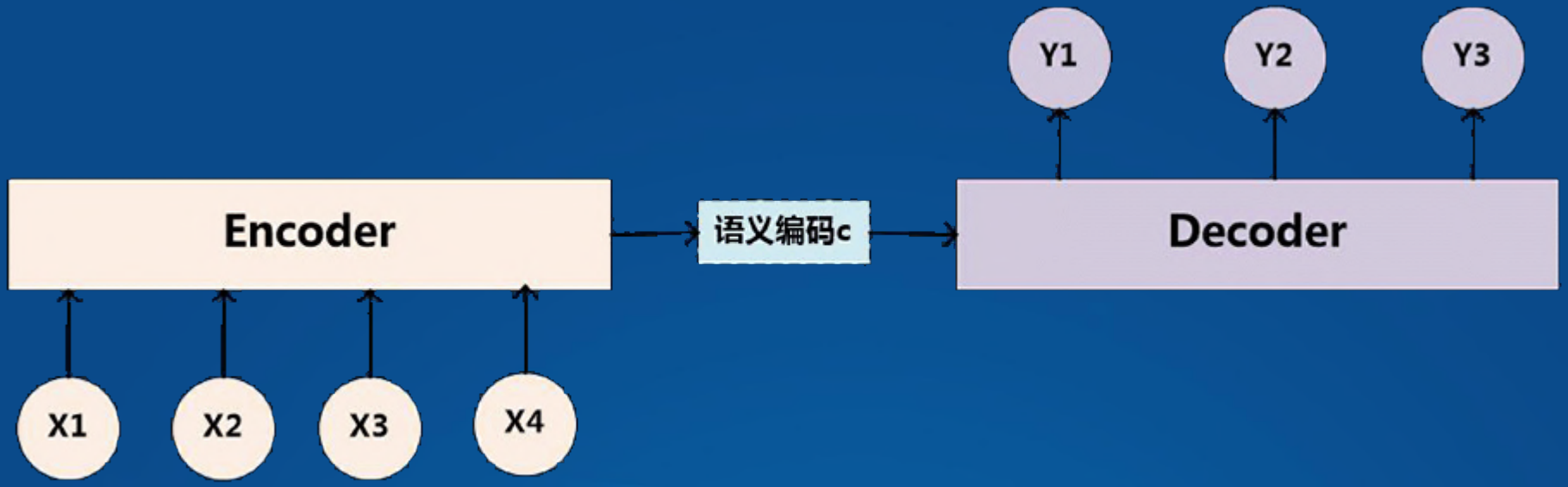
基于深度学习的财务领域语言模型

困惑度计算公式：$$PP(W) = P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \cdots w_N)}}$$

模型	汉语字 Uni-gram	汉语字 Bi-gram	汉语字 Tri-gram	汉语词 Uni-gram	汉语词 Bi-gram	RNN 语言模型
困惑度	733.2	141.3	50.9	129.1	43.4	36

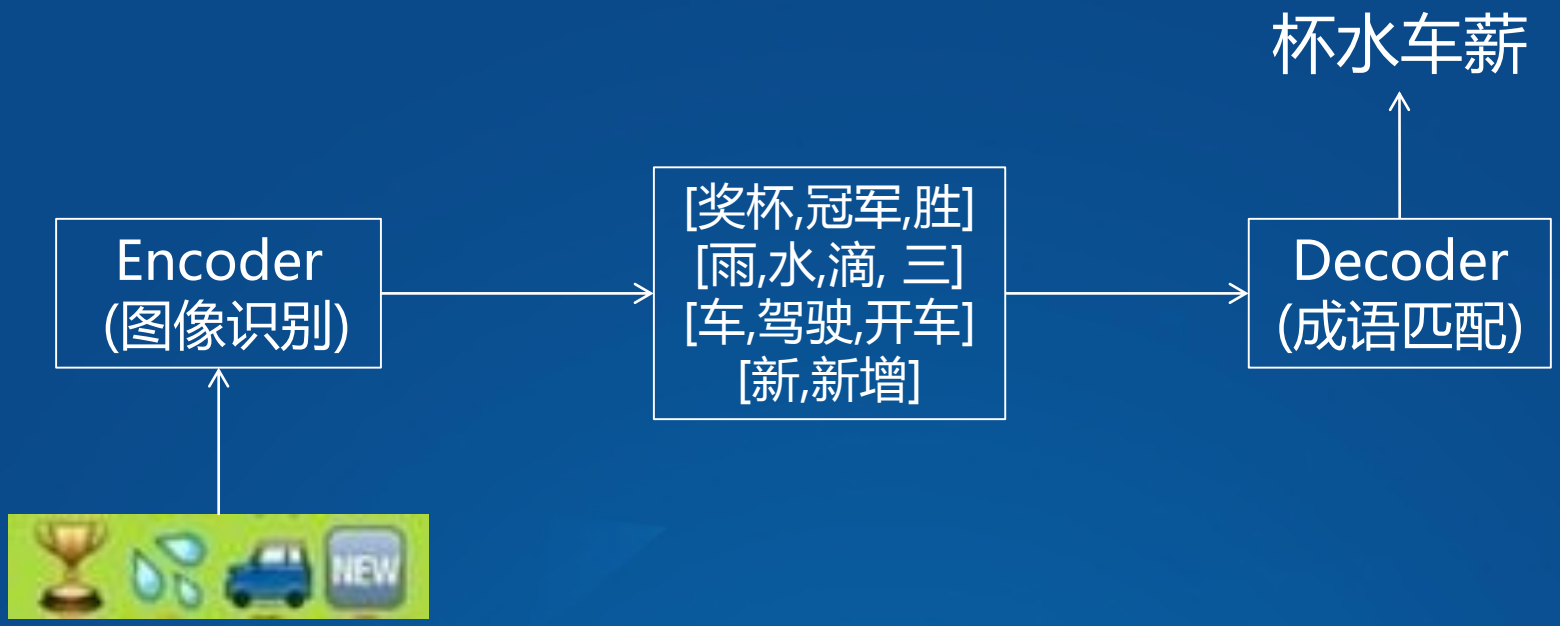
RNN语言模型效果

Encoder-Decoder深度学习框架

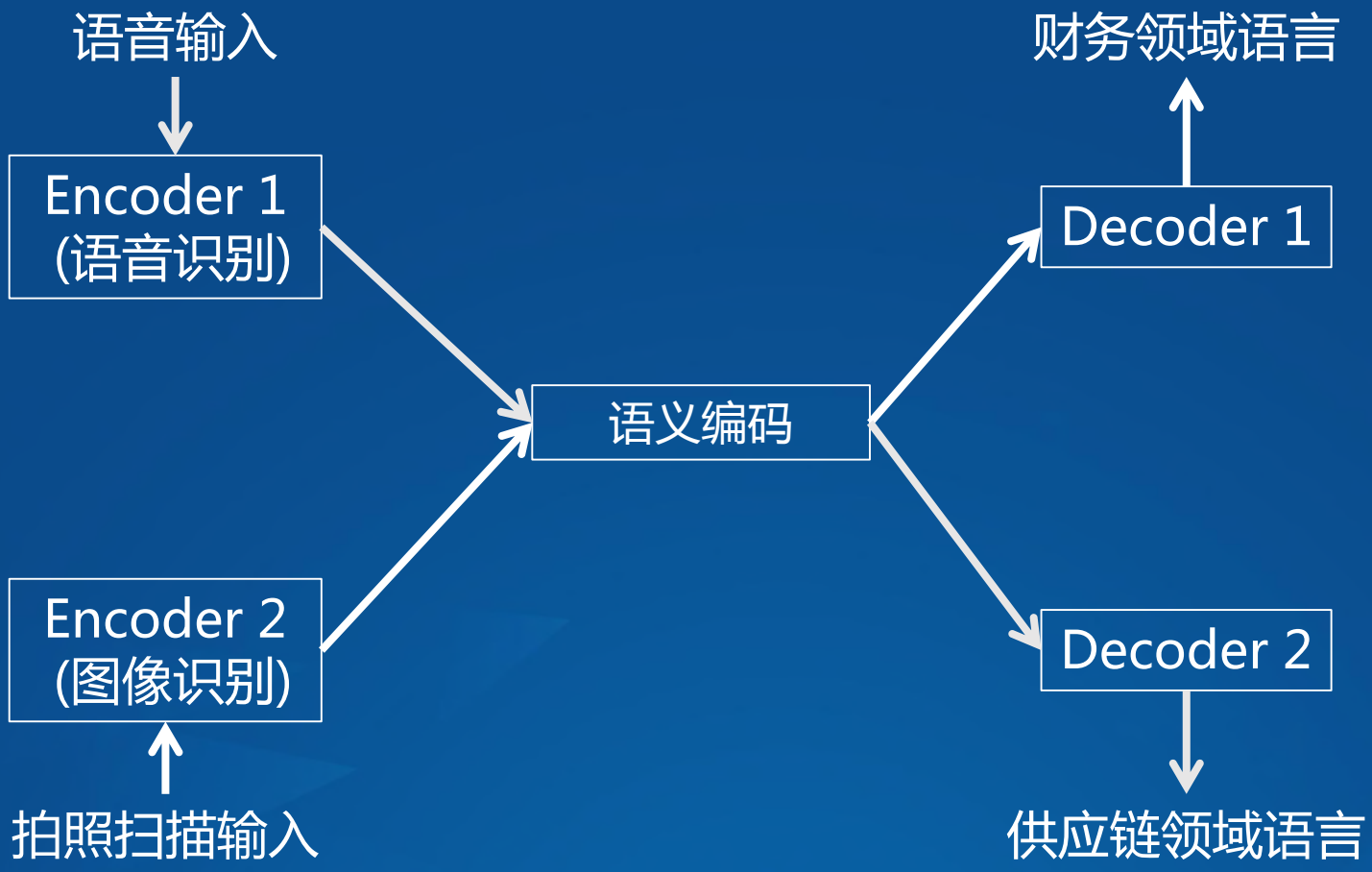


Encoder-Decoder深度学习框架

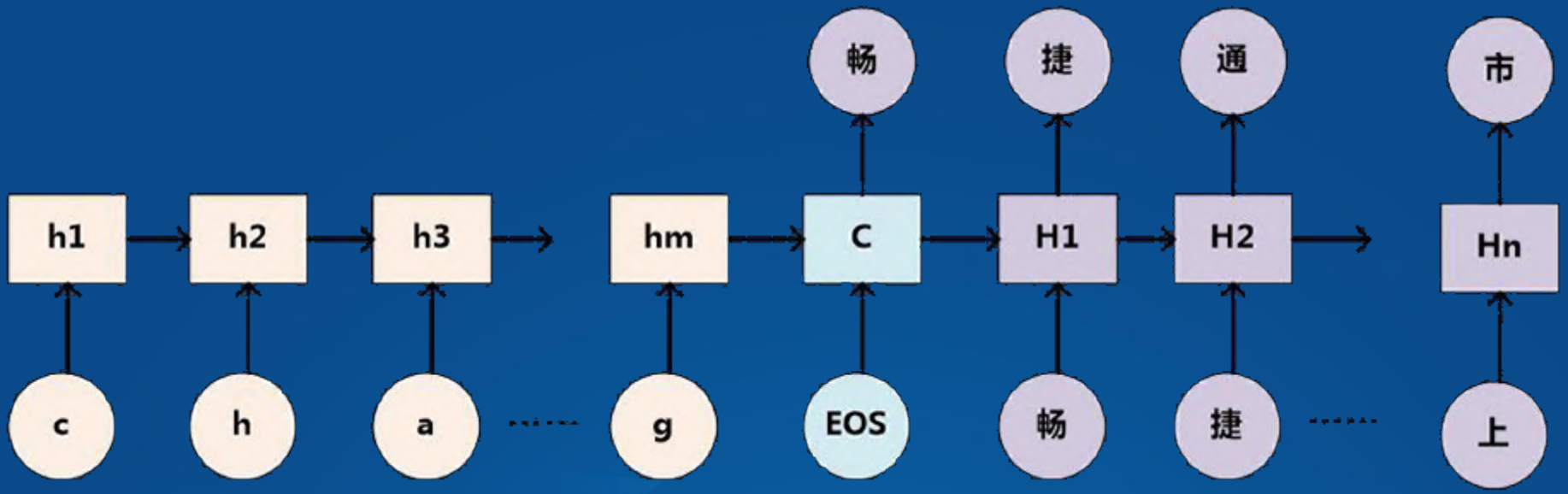
Encoder-Decoder深度学习框架



Encoder-Decoder深度学习框架

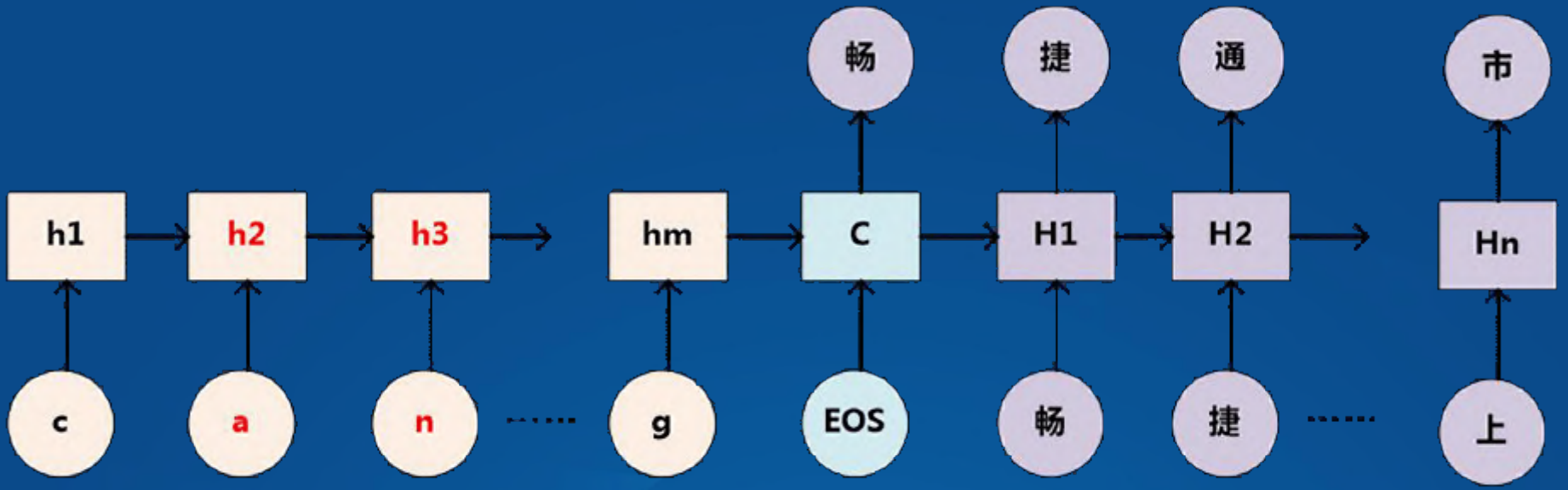


基于Encoder-Decoder框架的深度学习音字转换及纠错模型



Encoder-Decoder音字转换系统

基于Encoder-Decoder框架的深度学习音字转换及纠错模型



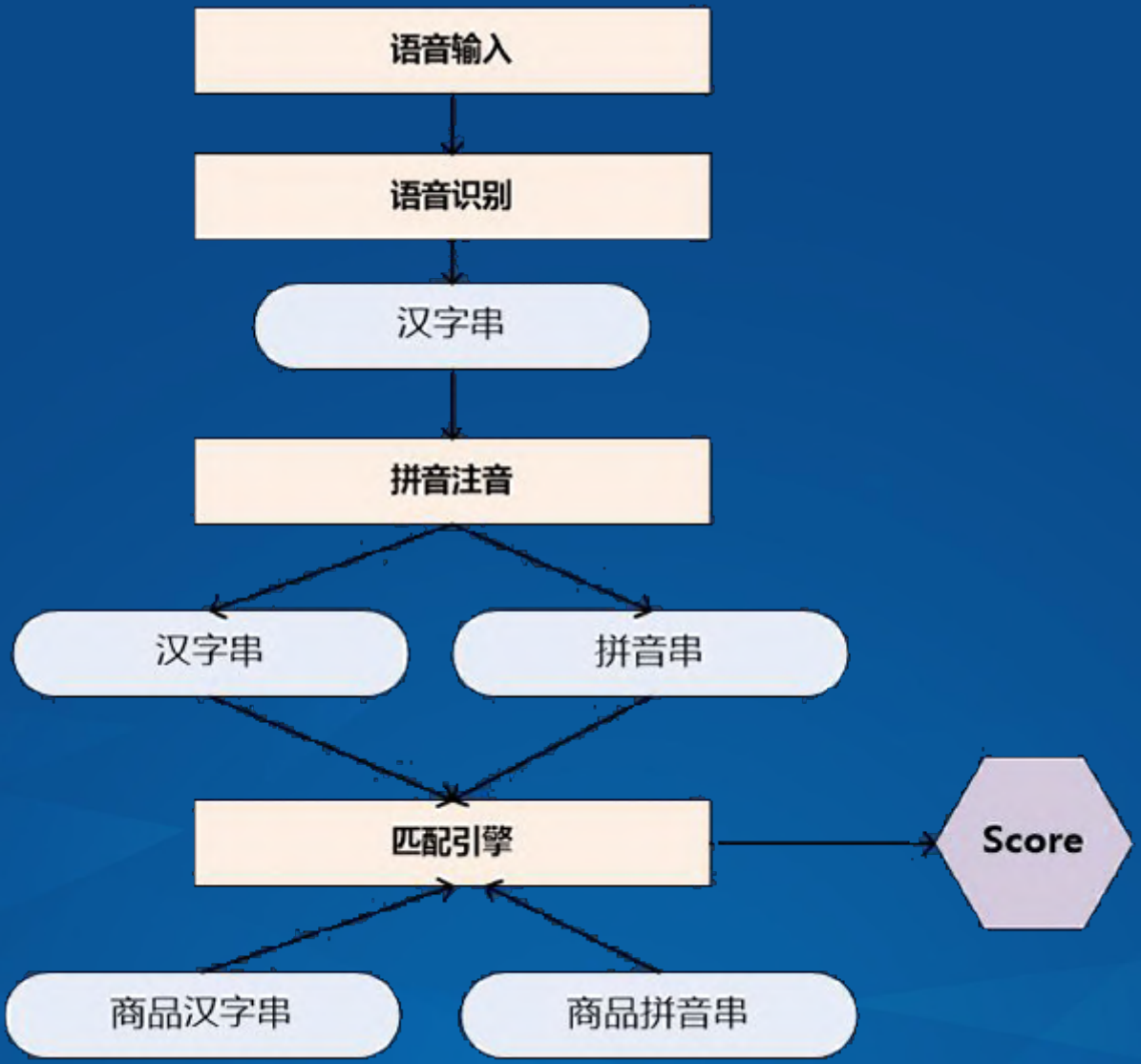
Encoder-Decoder音字纠错模型

拼音智能匹配引擎

拼音	详情
翘舌音	zh/z、ch/c、sh/s
前后鼻音	ang/an、eng/en、ing/in
模糊音	n/l、h/f、l/r
复韵母	an/ai、en/ei、un/ui

常见方言发音错误

拼音智能匹配引擎



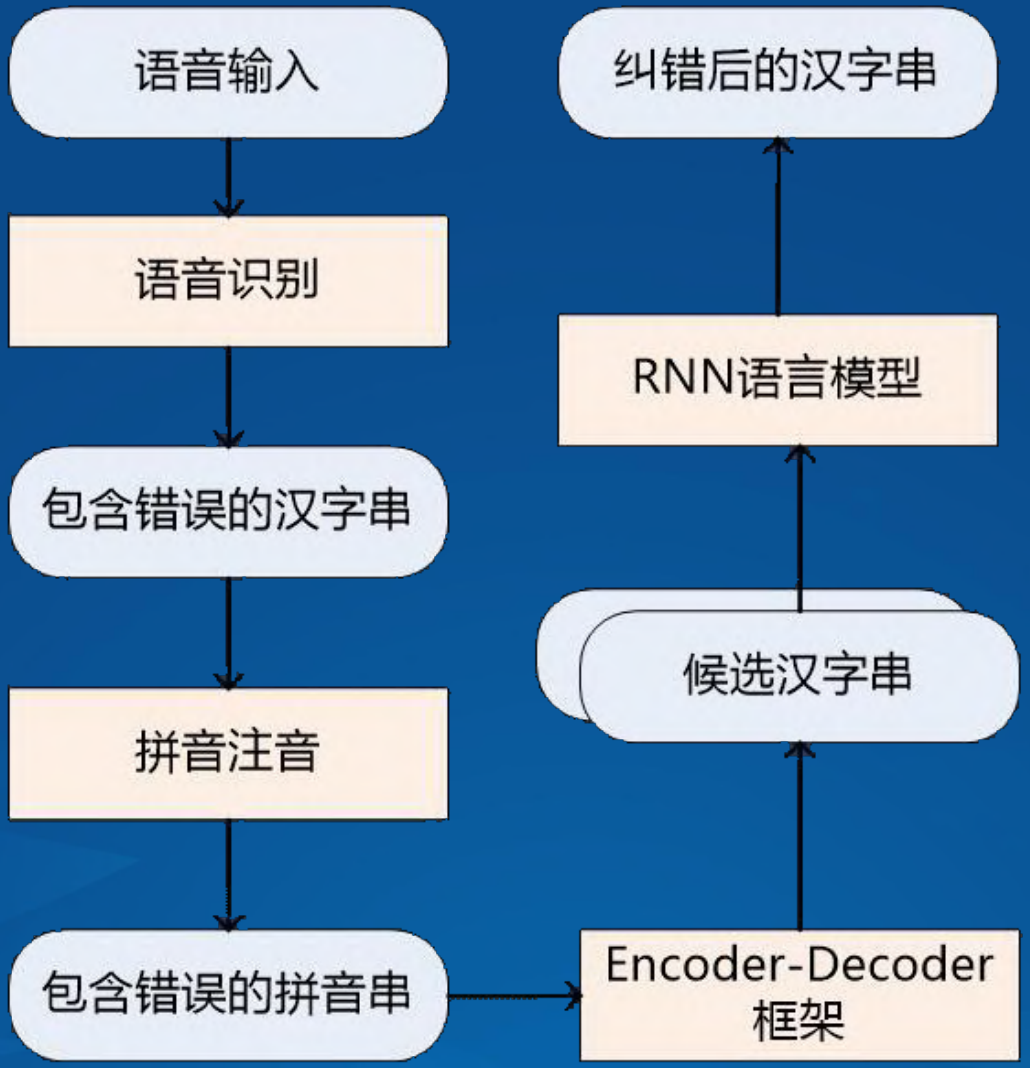
拼音智能匹配引擎

因子编号	因子（以“畅捷通”作为例子）
1	一元汉字匹配因子：{畅，捷，通}{畅，捷，通}
2	二元汉字匹配因子：{畅捷，捷通}{畅捷，捷通}
3	一元单字拼音匹配因子：{chang, jie, tong}{chang, jie, tong}
4	二元单字拼音匹配因子：{changjie, jietong}
5	二元全拼匹配因子： {ch, ha, an, ng, ji, ie, to, on, ng}:{ch, ha, an, ng, ji, ie, to, on, ng}
6	三元全拼匹配因子： {cha, han, ang, jie, ton, ong}:{cha, han, ang, jie, ton, ong}
7	二元简拼匹配因子:{cj, jt}:{cj, jt}
8	三元简拼匹配因子:{cjt}:{cjt}
9	二元韵母匹配因子:{an, ng, ie, on, ng}:{an, ng, ie, on, ng}
10	三元韵母匹配因子:{ang, ie, ong}:{ang, ie, ong}

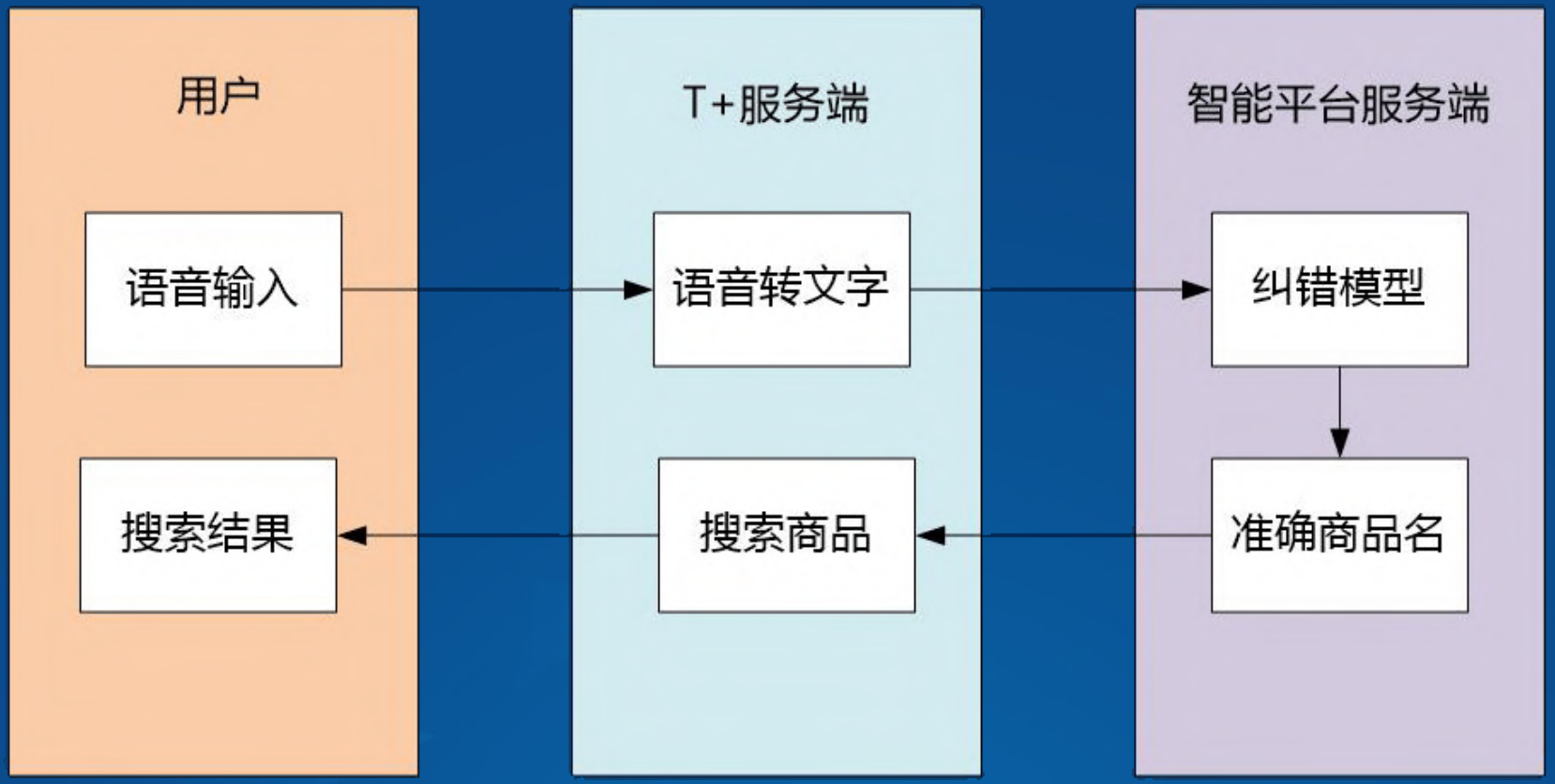
$$\text{Similarity}(\text{UserInput}, \text{EntItem}) = \sum_{i=1}^{k} a_i * Jaccard_i \ (\text{UserInput}, \text{EntItem})$$

匹配因子和相似度计算

综合纠错系统


语音输入
语音识别
包含错误的汉字串
拼音注音
包含错误的拼音串
Encoder-Decoder
框架
候选汉字串
RNN语言模型
纠错后的汉字串

应用场景：T+语音下单商品识别



自然语言处理在企业应用领域的实践

语音下单效果展示


上午9:42
商品列表
(全部) 输入名称|助记码|编码|规格...
最近60天订购
全部商品
新品
100002/红盈袋装泡小黄姜芽2000g(200)
现存量: 1000kg 零售价:
最低售价: 结存成本:
最新售价: 6.00元/kg
权荣其京春荣
说完了
赠品
数量
价格
100004/春蓉鸡精454g(200)
现存量: 1000kg 零售价:
最低售价: 结存成本:
最新售价: 33.00元/kg
赠品
数量
价格
100001/黎红麻辣油168ml(200)
扫描
确定


上午9:42
商品列表
(全部) 输入名称|助记码|编码|规格...
最近60天订购
全部商品
新品
请选择存货
春蓉鸡精100g
春蓉鸡精454g
春蓉泡菜400g
春蓉泡菜200g
豪吉鸡精227g
取消
最新售价: 33.00元/kg
赠品
数量
价格
100001/黎红麻辣油168ml(200)
扫描
确定

语音下单效果展示


上午9:49
商品列表
(全部) 输入名称|助记码|编码|规格
最近60天订购
全部商品
新品
100070/恒顺金标料酒500ml(200)
现存量:
零售价:
最低售价:
结存成本:
最新售价: 元/kg
赠品
数量
价格
与红蚂赖优莹
说完了
100072/周君记麻辣豆腐调料150g(200)
100073/海天鲜味生抽1.9L(200)
扫描
确定


上午9:49
商品列表
(全部) 输入名称|助记码|编码|规格
最近60天订购
新品
请选择存货
黎红麻辣油168ml
黎红麻辣酱350g
可喜可旺袋装红米辣100g
渝航牌青花椒麻辣鱼210g
哈吃青花椒麻辣鱼160g
取消
赠品
数量
价格
100073/海天鲜味生抽1.9L(200)
扫描
确定

语音下单效果展示

用户语音输入	百度语音识别	纠错结果
花之语杭白菊	话剧行白居	花之语杭白菊
鲜花椒油	新发就咬	鲜花椒油
海天老抽800ml	黑天脑抽八百毫升	海天老抽800ml
三五麻辣鱼150g	三无码了鱼150克	三五麻辣鱼150g
黎红鲜花椒油330ml	李红仙化交融330好声	黎红鲜花椒油330ml
黄心猕猴桃	广西猕猴桃	黄心猕猴桃
白皮花生	白皮花生	花生（白皮）

纠错结果示例

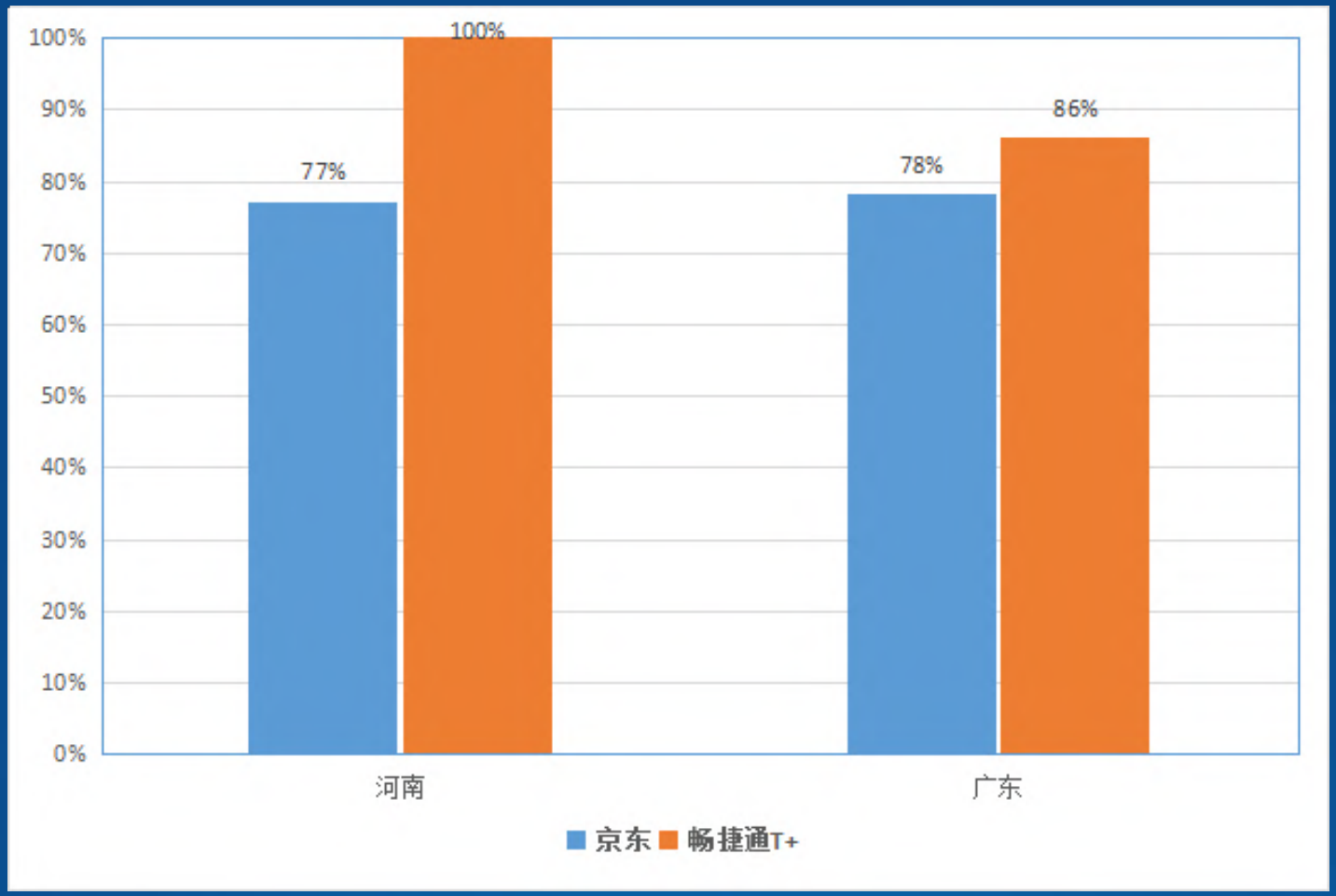
语音下单效果展示

	水果类（941种商品）	五金类（665种商品）	食品类（889种商品）
准确率	87%	95%	90%

	普通话（带口音）	北方省份（山东、东北等）	南方省份（广东、江苏、福建、湖南等）
准确率	> 95%	> 90%	> 86%

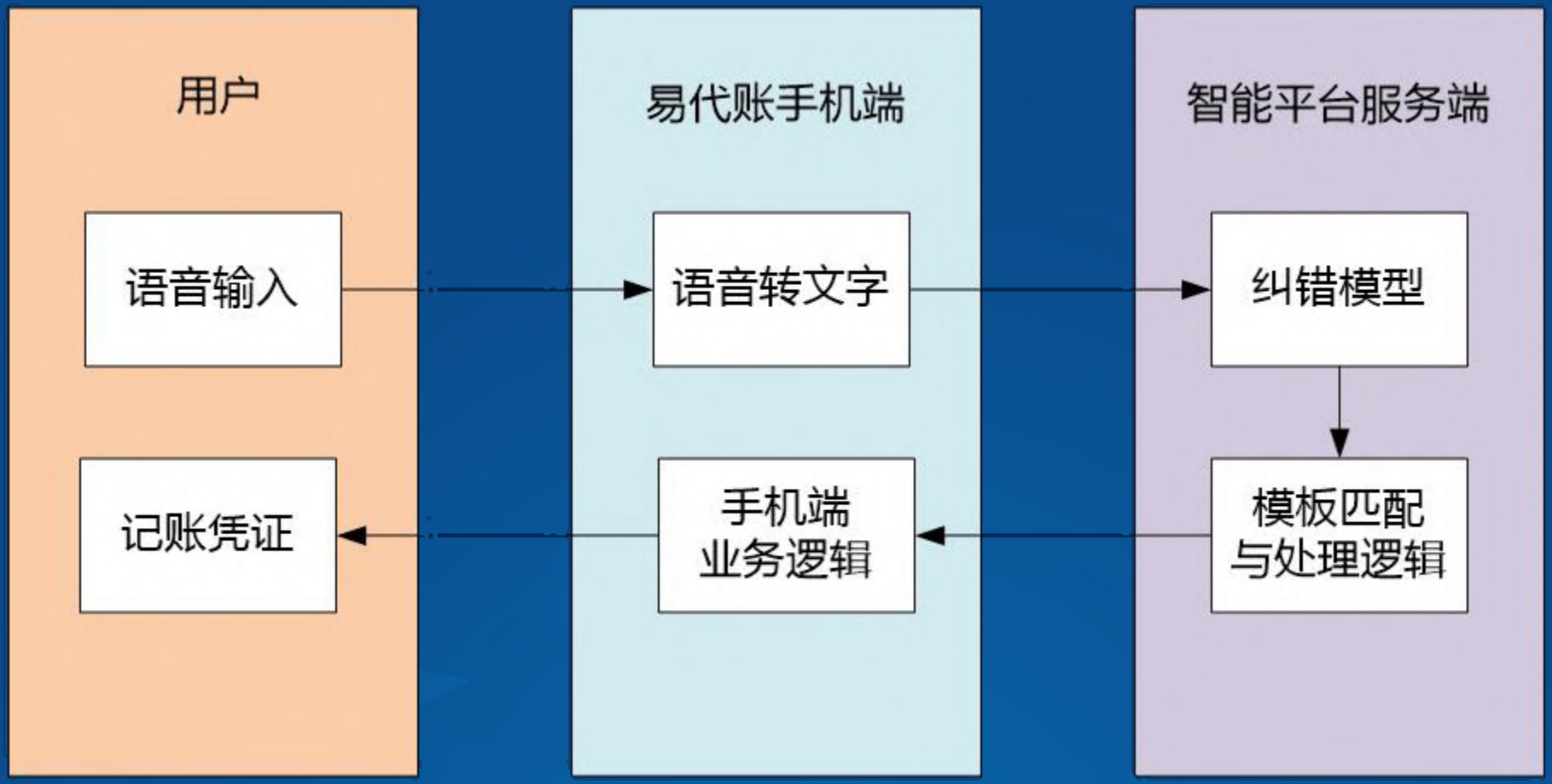
分类商品和方言识别率

语音下单效果展示

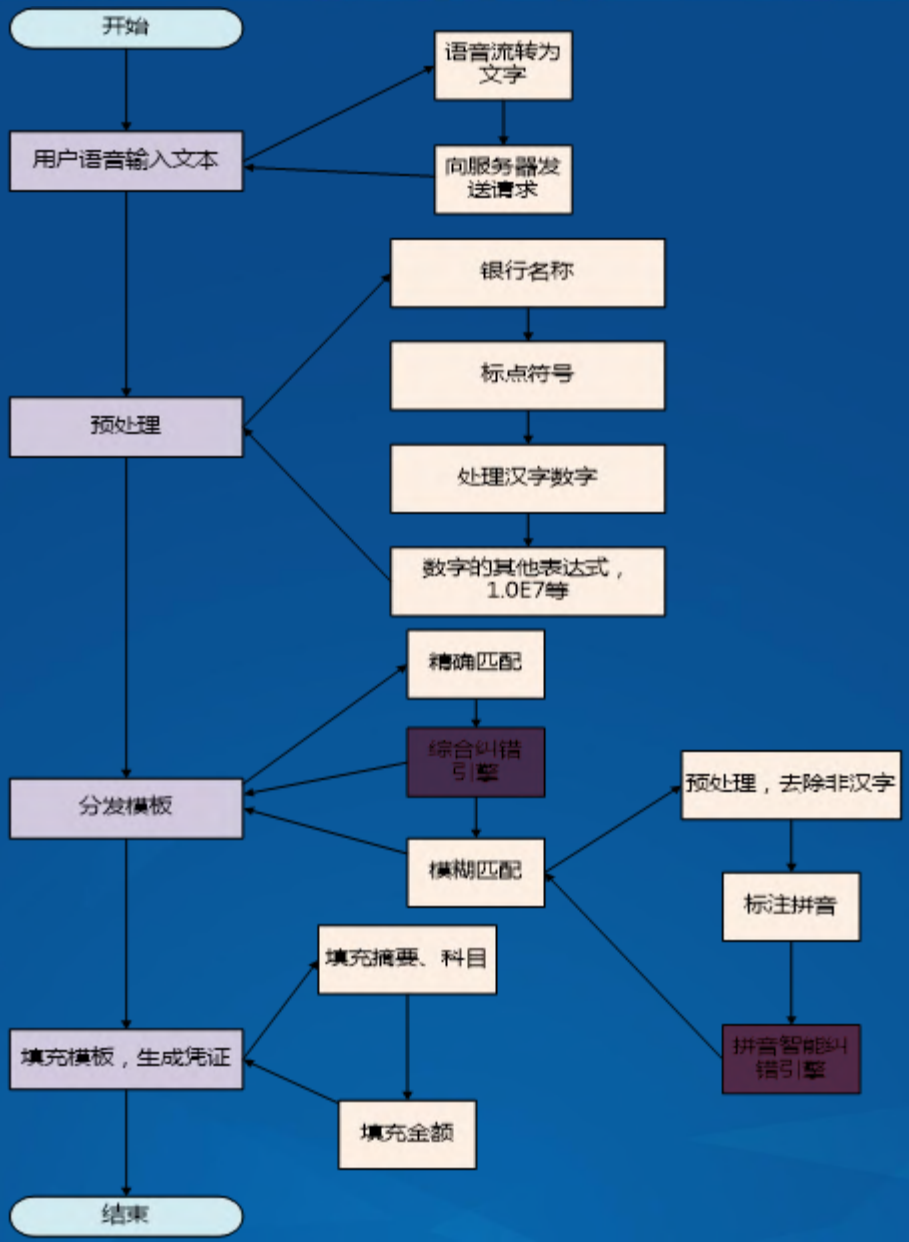


同类产品南北方言识别

应用场景：语音做账



语音做账


开始
语音流转为文字
用户语音输入文本
向服务器发送请求
银行名称
标点符号
预处理
处理汉字数字
数字的其他表达式，1.0E7等
精确匹配
综合纠错引擎
分发模板
预处理，去除非汉字
模糊匹配
标注拼音
填充摘要、科目
拼音智能纠错引擎
填充模板，生成凭证
填充金额
结束

语音做账效果展示

用户语音输入

买进原材料，银行支付材料款60000，预付40000，现金为17000。

语音识别结果

没经验材料以后这个材料款6万，预付4万邪剑仙为17000。

纠错及模板匹配

摘要	科目	明细科目	借方金额	贷方金额
买进原材料	原材料		100000	
	应交税费	应缴增值税（进项税额）	17000	
	银行存款			60000
	库存现金			17000
	预付账款			40000

语音做账效果展示


买点原材料银行支付材料款6000元预付4000元现金支付17000元。
您可以这样说:
识别中...


返回
新增凭证
记字第004号
2015-09-30
摘要 买进原材料
科目 1002 银行存款
借方 0.00
贷方 6000.00
摘要 买进原材料
科目 1403 原材料
借方 10000.00
贷方 0.00
摘要 买进原材料


科目 1123 预付账款
借方 0.00
贷方 4000.00
摘要 买进原材料
科目 22210101 应交税费-应交增值税-进项税额
借方 17000.00
贷方 0.00
摘要 买进原材料


科目 22210101 应交税费-应交增值税-进项税额
借方 17000.00
贷方 0.00
摘要 买进原材料
科目 1001 库存现金
借方 0.00
贷方 17000.00
合计: 借方27000.0
贷方27000.0
制单人:张金宝1426225423
附单据 0 张
保存再记
保存

语音做账效果展示


买点原材料银行支付材料款6000元预付4000元现金支付17000元。
您可以这样说:
识别中...


返回
查看凭证
更多
2015-09
记-004
2015-09-30
摘要 买进原材料
科目 1123 预付账款
贷方金额 40,000.00
摘要 买进原材料
科目 1403 原材料
借方金额 100,000.00
摘要 买进原材料
科目 22210101 应交税费-应交增值税-进项税额
借方金额 17,000.00
摘要 买进原材料
科目 1002 银行存款
贷方金额 60,000.00
摘要 买进原材料
科目 1001 库存现金
贷方金额 17,000.00

The 8th China Cloud Computing Conference

Thank you