QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE
BEIJING 2017

深度学习在电商搜索和聊天机器人中的应用探索

SPEAKER / 程进兴

2017年4月

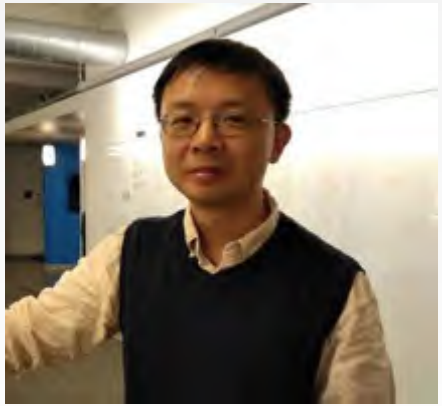
主办方 Geekbang. InfoQ

苏宁国际美国硅谷研究院

苏宁美国硅谷研究院创建于2013年11月，其宗旨是建立高科技人才和专利的蓄水池，推动苏宁持续地创新和转型，为用户提供简约完美的用户体验。

硅谷研究院由来自云计算、大数据、人工智能及深度学习等不同专业背景的工程师、数据科学家及分析师组成。目前包含人工智能、大数据和创新三个实验室。

个人简介



- 程进兴，苏宁美国研究院技术总监，斯坦福大学博士，清华大学本科。 曾在甲骨文，雅虎，微软，沃尔玛实验室等多家公司从事搜索，广告，大数据分析，机器学习，人工智能应用等方面的研发工作。在此期间，发表了10多篇相关领域的研究论文，并有10多项相关领域的专利。

- 业余爱好： 骑行

电子邮箱： jim.cheng@ususing.com

议程

- **深度学习与商品搜索**
  - 矢量化搜索技术简介
  - 基于词语聚类的矢量化
  - 基于用户会话的矢量化
  - 原型评测结果及效果示例

- **深度学习与聊天机器人**
  - 聊天机器人简介
  - 聊天机器人主要模块及架构
  - 深度学习探索
  - 聊天机器人评测结果

目前商品搜索中的一些问题

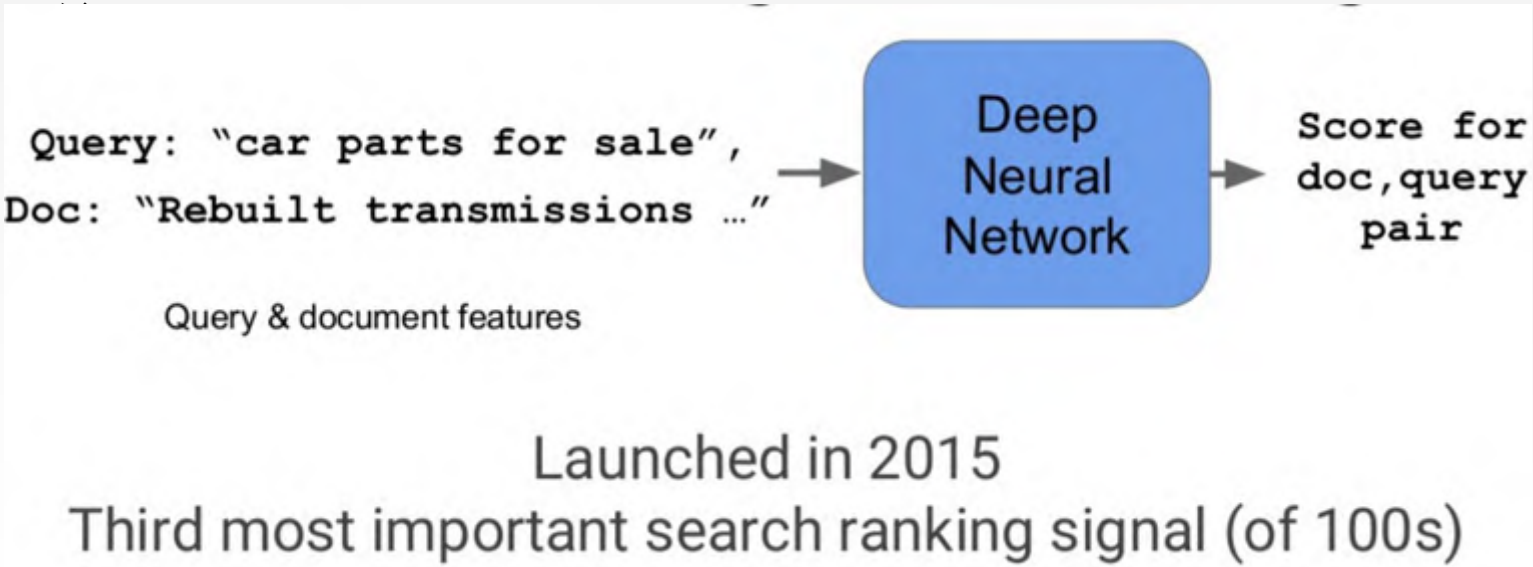
- 语义词汇差异
  - 理发器，理发推子，电推子
  - 血糖计，血糖仪
  - 山地车，死飞，自行车，碟刹，折叠车，公路车，单车

- 解决方案
  - 同义词？
  - 归一化？

  預報 =》预报， 五岁 =》 5岁

人工智能／深度学习在搜索中的应用：网页／电商搜索

- 基于深度学习的（Query, Document）分数是Google搜索引擎中第3重要的排序信



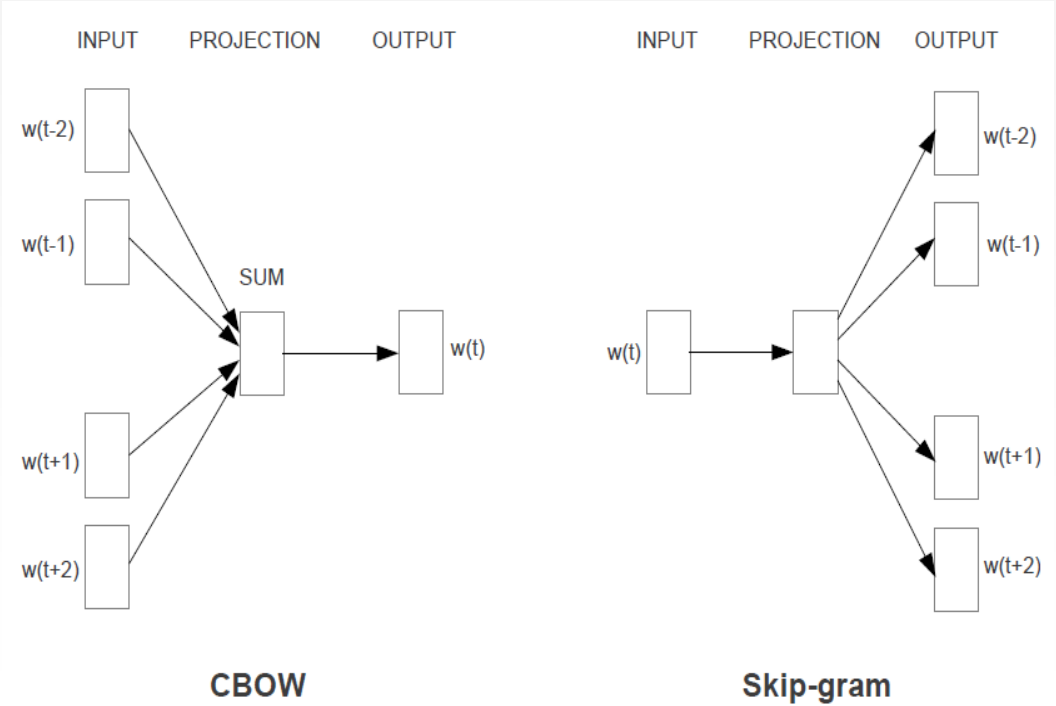
- 亚马逊（Amazon／A9)电子商务搜索引擎中， 深度学习还在实验阶段，尚未进入生产线。

矢量化搜索模型

- **搜索数值矢量化**
  - 传统搜索基于文字匹配，商品包含搜索词或者不包含搜索词
  - 利用深度学习技术，将搜索词和商品全部数值矢量化，将文字匹配转化为数值矢量计算
  - 词语矢量化是进一步进行各种深度学习的基础。

- **矢量化模型介绍**
  - Mikolov(Google员工)等人2013发表了两篇关于Word2Vec的文章，成为词语矢量化表示的基础
  - Word2vec的优点：
    - ✓ 词语矢量考虑了上下文及词语之间的语义关系
    - ✓ 复杂词语可以通过矢量计算来实现（如 Vec（北京）＝ vec(东京) － vec(日本) + vec(中国) ）
  - 矢量化模型的现况
    - ✓ 词语的矢量化模型已经有开源实现方案
    - ✓ 句子和文档的矢量化还在摸索阶段，尚不成熟
    - ✓ 已经有一些在词语相似度，舆情分析等方面的应用

词语矢量化模型



CBOW: 通过上下文词语来预测词语本身出现的概率

Skip-gram: 通过词语本身来预测上下文词语出现的概率

基于词语聚类的矢量化模型

- Word2vec等工具可以有效地将词语转化为向量
- 将句子/段落/文章有效转化为向量则有很大的挑战。
  - 简单平均/加权平均容易失去句子等的语义/结构信息
  - 直接以句子为单位进行训练，则训练文本严重不足
- 电商搜索中遇到的主要是句子/短文分析，可以将短文中的词语聚类，挑选具有代表性的词语聚类结果，来表示整个短文
- 传统聚类（如Kmeans)在几何距离的基础上进行聚类，效果不好。利用随机过程做词语聚类可以解决这一问题

基于词语聚类的矢量化模型

具体的生成cluster的流程如图：

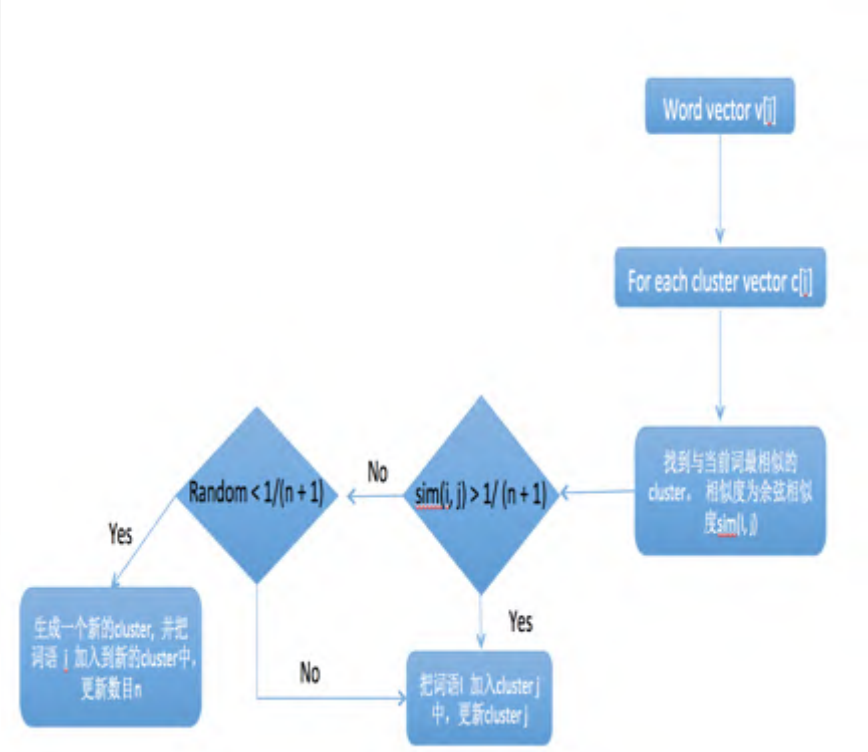
V[i]: 为产品信息里每个词的词语向量(word vector)分数

C[i]: 为聚类(cluster)的vector分数

N: 为cluster的数目

Sim(I, j): 词语i 与cluster j的余弦相似度

Random: 生成一个0 – 1之间的随机数



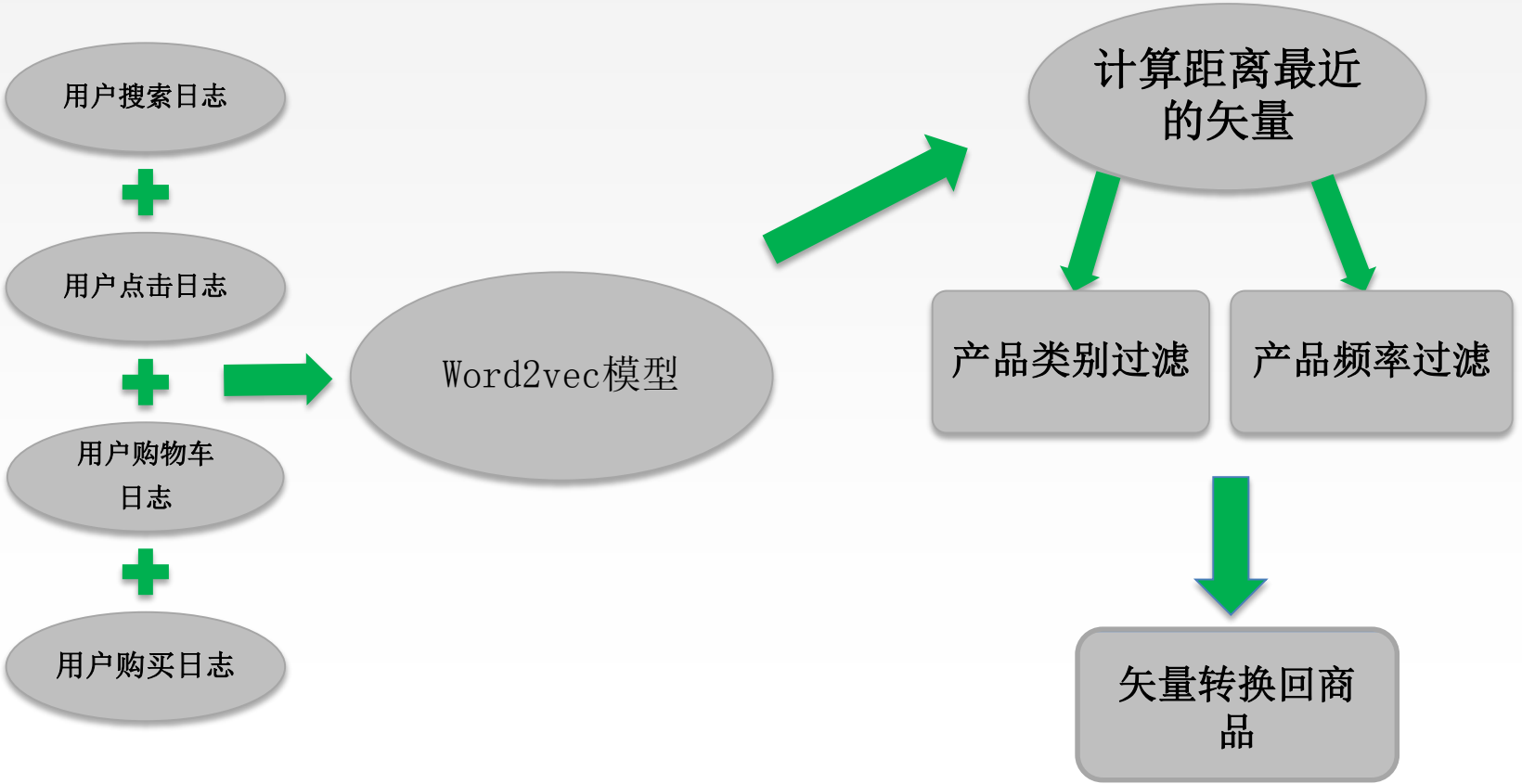
基于用户反馈的矢量化

- 把搜索词和商品文档各自作为整体看待，直接学习训练各自的矢量值
- 通过分析用户每次访问的行为顺序， 构建有“搜索词”和“商品文档”组成的句子
- 训练集是采用苏宁易购的用户搜索日志作为来源。在经过数据清理之后，按照搜索的时间顺序，结合商品的点击，商品放入购物车，商品的购买这些用户行为，而建立的矢量化训练数据

小米手机4c，小米手机4s，142074410

美的冰箱 270，美的冰箱645，美的冰箱 330，132268985，美的 2155，美的冰箱，美的冰箱 550

基于用户反馈的矢量化模型



原型评测结果

矢量化搜索引擎与易购传统引擎搜索效果对比 （2016-07-25测试结果）

Queries	Rating on Returned Results		Notes
方太 EM23TS+FD23BE	-2	1	Testing: returned 0 results. PRD: returned 6 relevant results.
三星回音壁	2	1	Testing: returne 15 relevant results and many other brands of sound system.
家用高压洗车机	2	2	
辣鱼	0	1	Both environments returned spicy snacks, but testing's results are less than PRD.
8核4g运行内存	-1	0	Testing: only 10 are relevant in top 64 results. The rest are either laptops, other brands of cell phones, or computers. PRD: 53 returned results contain 36 less relevant results.
手办模型	2	2	
m4800	2	1	Testing: top 64 results are very relevant. PRD: returned 38 relevant results.
美巴喜婴儿床	2	-1	Testing: top 64 results are very relevant. PRD: returned only 8 relevant results.
AOC T3207M	2	1	Testing: returned 3 relevant results and 34 less relevant results. PRD: returned 3 relevant results.
棉包	2	-1	Testing: top 64 results are very relevant. PRD: top 60 results contain only 4 relevant. The rest are irrelevant.
Total score	73	70	
返回结果相关性平均值	1.22	1.17	

效果示例

- 该技术不仅召回与搜索词完全匹配的结果，还可召回与搜索词文本不匹配、但含义近似的结果。

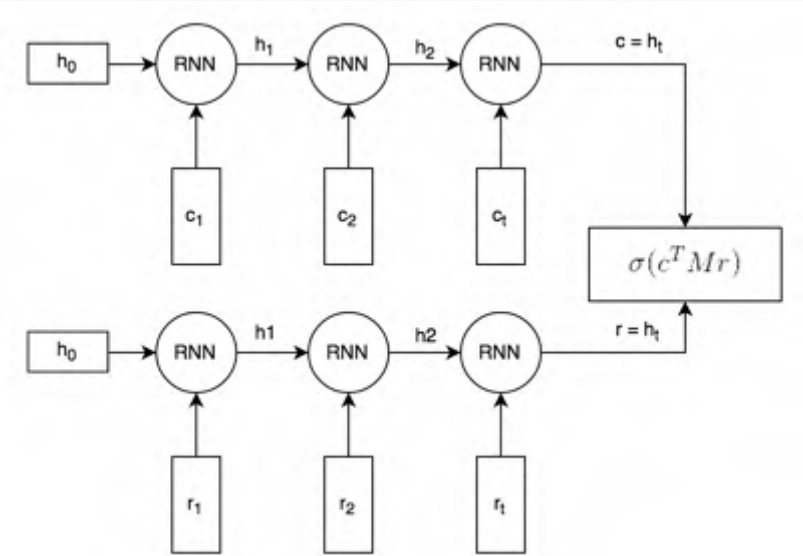
如：经测评，当搜索词为“松下筒灯”， 易购网站返回6个相关结果，
美研方案返回64个相关结果

现有方案

原型系统

正在进行的探索

- 首先进行词语的矢量化
- 词语矢量作为各种深度学习模型的输入值
- 示例深度学习架构： dual RNN（ dual LSTM)
- 利用用户反馈数据来补充训练样本



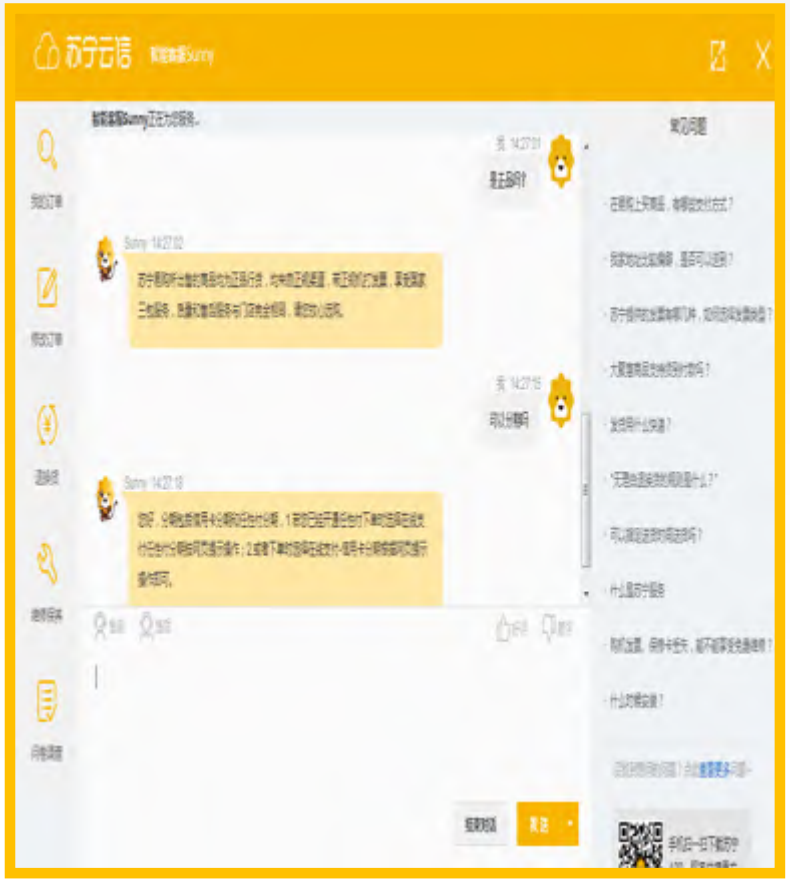
聊天机器人(chatbot)

- 聊天机器人是一种聊天代理，它通过电脑程序设计与人类通过音频或文本进行智力对话。 --维基百科

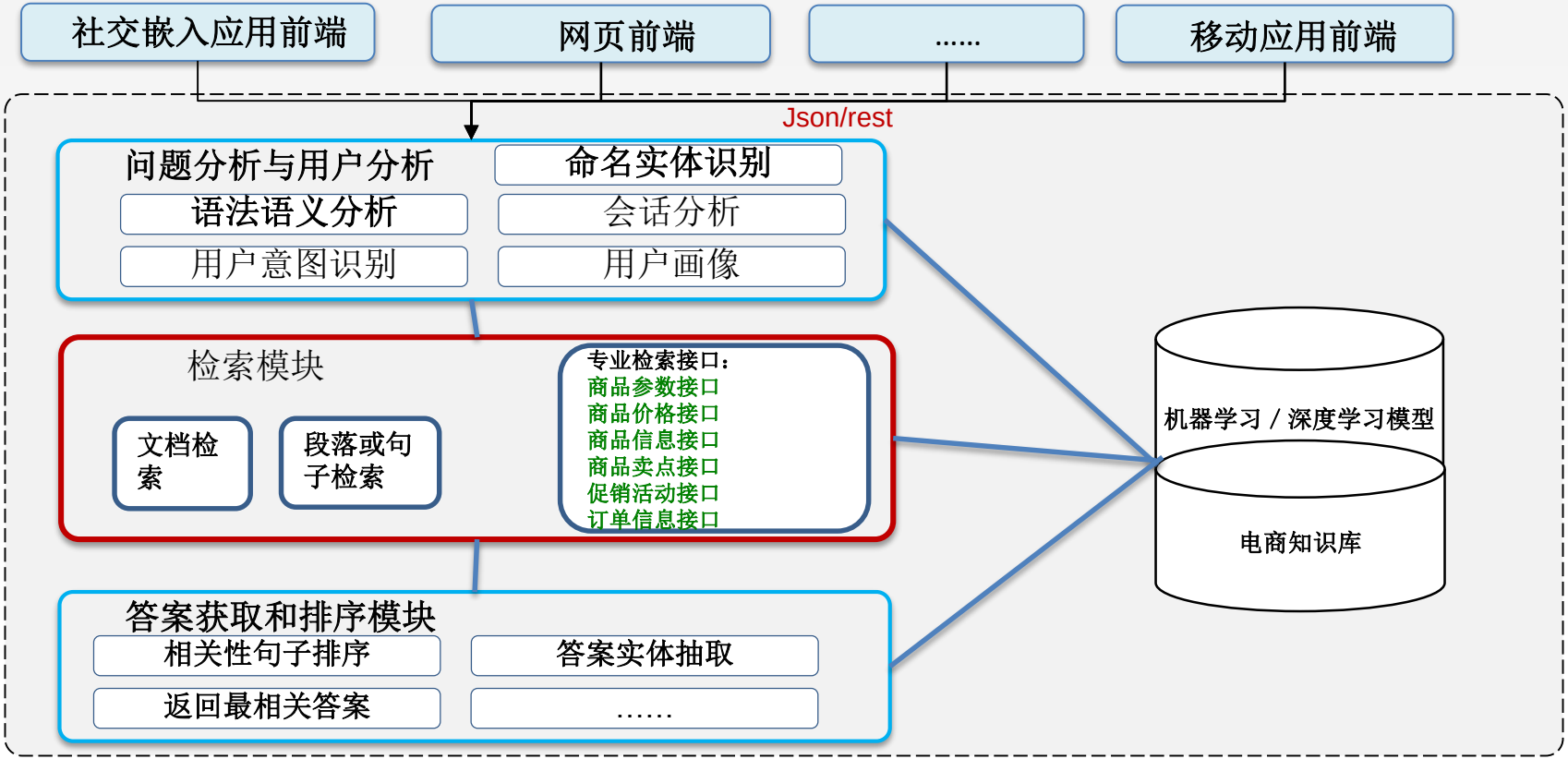
- 未来，聊天应用将被看作是新的浏览器，而机器人程序将成为新的网站。这就是互联网的新开始。--Ted Livingston, CEO of KiK

- 聊天机器人将从根本上变革每个用户对人机交互的体验。 --Satya Nadella, Microsoft CEO

应用示例：苏宁易购机器人Sunny，百度度秘，Amazon Echo



系统架构图


社交嵌入应用前端
网页前端
.....
移动应用前端
Json/rest
问题分析与用户分析
命名实体识别
语法语义分析
会话分析
用户意图识别
用户画像
检索模块
文档检索
段落或句子检索
专业检索接口：
商品参数接口
商品价格接口
商品信息接口
商品卖点接口
促销活动接口
订单信息接口
答案获取和排序模块
相关性句子排序
答案实体抽取
返回最相关答案
.....
机器学习 / 深度学习模型
电商知识库

自然语言处理
（Natural Language Processing）

- 最简单地， 用户文字输入的理解可以采用“bag of words”模型。
- 输入的文本可以根据词性、时态等被进一步标签分割。
- 语境信息可以进一步通过word2vec建模。
- 概率语言模型可以用于词汇赋权重。
- 深度神经元网络可以进一步提升自然语言处理的效果
- 电商领域内的各种专业字典（如品牌，产品，型号等）可以协助识别各种实体

命名实体识别

命名识别模块

数码相机不超过 200 元

Word	Type
数码相机	产品
不超过	比较_lte
200	数字
元	单元_currency

iphone 5s 手机保护壳

Word	Type
iphone	品牌
5s	型号
手机	产品
保护壳	附件

西门子洗衣机滚筒

Word	Type
西门子	品牌
洗衣机	产品
滚筒	工作原理

- 人名
- 地名
- 品牌名
- 商品名
- ……
- 机构组织名

识别用户输入中的各种实体是进一步识别用户意图的基础

用户意图识别

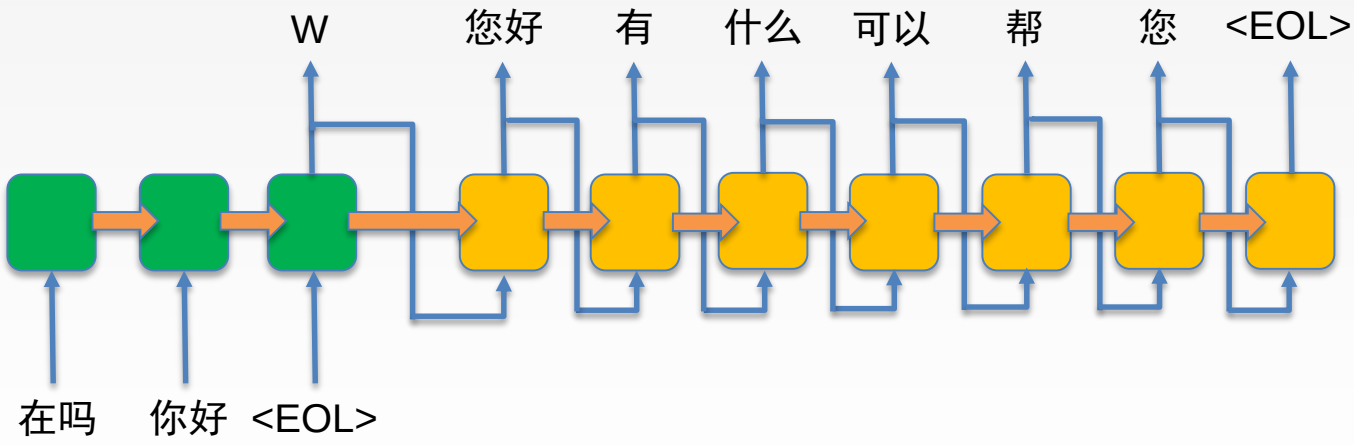
用户意图识别模块

你好，我买了两台空调，想问下安装咋收费的呀？ =》售后服务
问问你，苹果6与6S的运行内存都是1G吗？ =》商品研究
订单能不能改成货到付款？ =》订单查询

- 商品研究
- 下单购买
- 订单查询
- 售后服务
- ……
- 其它闲聊

- 用户意图识别是非常重要的一环。针对不同的意图，可以采用不同的策略回应
- 用户意图识别可以采用深度学习建模分类

深度学习模型： 从会话历史数据中学习回答问题



模型参数：
Dropput rate: 0.5
Learning rate: 0.0001
Embedding dimensions: 1024
Mini-batch size: 32
Number of epochs: 100,000
Number of LSTM layers: 3

数据清理

- 337,190 问答对
- 填充(Padding)
  - 通过填充将输入文本序列转化为固定长度， 并采用一些特殊符号 （EOS, PAD, GO, UNK等）
  - 通过对答案进行逆序处理优化训练结果:
    - ✓ Q : [PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, "在吗", "你好"]
    - ✓ A : [GO, "您好", "有", "什么", "可以", "帮", "您", EOS, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD, PAD]
- training: 70%; validation: 15%; test: 15%
- 词语集大小：6000（最常用词语）
  - 覆盖96.81% 的词语
  - 词语集之外的词用UNK表示.

样例结果

77 out of 256 random samples from test set
q : [你 工 号 多少]; a : [我 这边 帮 您 联系 一下 快递 那边]
q : [不合 适 就 退 不然 真 不 合 理]; a : [亲]
q : [你 现在 到 unk 镇 哪里 提 就 不能 帮 我 查 下 吗]; a : [这个 是 苏宁 发货 的]
q : [我的 增 票 认证 已经 成功 为什么 还 没有 给 我 寄 发票 呢]; a : [您好 请问 有 什么 可 以 帮 您 的 呢]
q : [可以 有 人员 上门 帮忙 安装 调试 吗]; a : [安装 师傅 上门 安装 的]
q : [零钱宝 没 得 钱 呢]; a : [您 可以 在 易付宝 里 申请 一下]
q : [他 说 等下 就 送]; a : [那 您 可以 先 联系 下 厂家 的 电话]
q : [不是 可以 上 网 看 优 酷 unk 吗]; a : [不支持 的 哦]
q : [好 的 到 时候 再 说 吧]; a : [好 的]
q : [多久 能 到货]; a : [您 稍 等]
q : [m 的 小 吧]; a : [嗯 嗯]
q : [我 急 着 用 哎 我的 unk 已经 在 现场 等 着 了 空调 没 装 好 他 不能 unk]; a : [这个 是 需要 您 自 己 承担 的]
q : [谢谢 啦]; a : [不 客气 的]
q : [你 哪 里 查 不到 吗]; a : [稍 等]
q : [冰箱 内 的 冷藏 的 unk 小 铲子 都是 自带 的 对 吧]; a : [是 的 呢]
q : [你好]; a : [您好 请问 有 什么 可 以 帮 您 的 吗]

聊天机器人评测结果

- 从实际客户回话中抽样选取约1700个问题
- 问题覆盖不同的应用场景
- 经过4个版本迭代后， 机器人问题回答的准确率和覆盖率均达到业界领先水平

测评情况						
例总数	已回答用例数	无法回答用例数	回答正确用例数	意图识别正确用例数	用例回答率（已回答用例数/用例总数）	用例回答正确率（回答正确用例数/用例总数）
1723	1721	2	747	1108	99.88%	43.35
1723	1722	1	758	1119	99.94%	43.99
1723	1721	2	1045	1245	99.88%	60.65
1723	1719	4	1226	1396	99.77%	71.15

Questions and Answers