知乎NLP技术应用与挑战

黄 波

huangbo@zhihu.com

WeChat:

2018.05.23

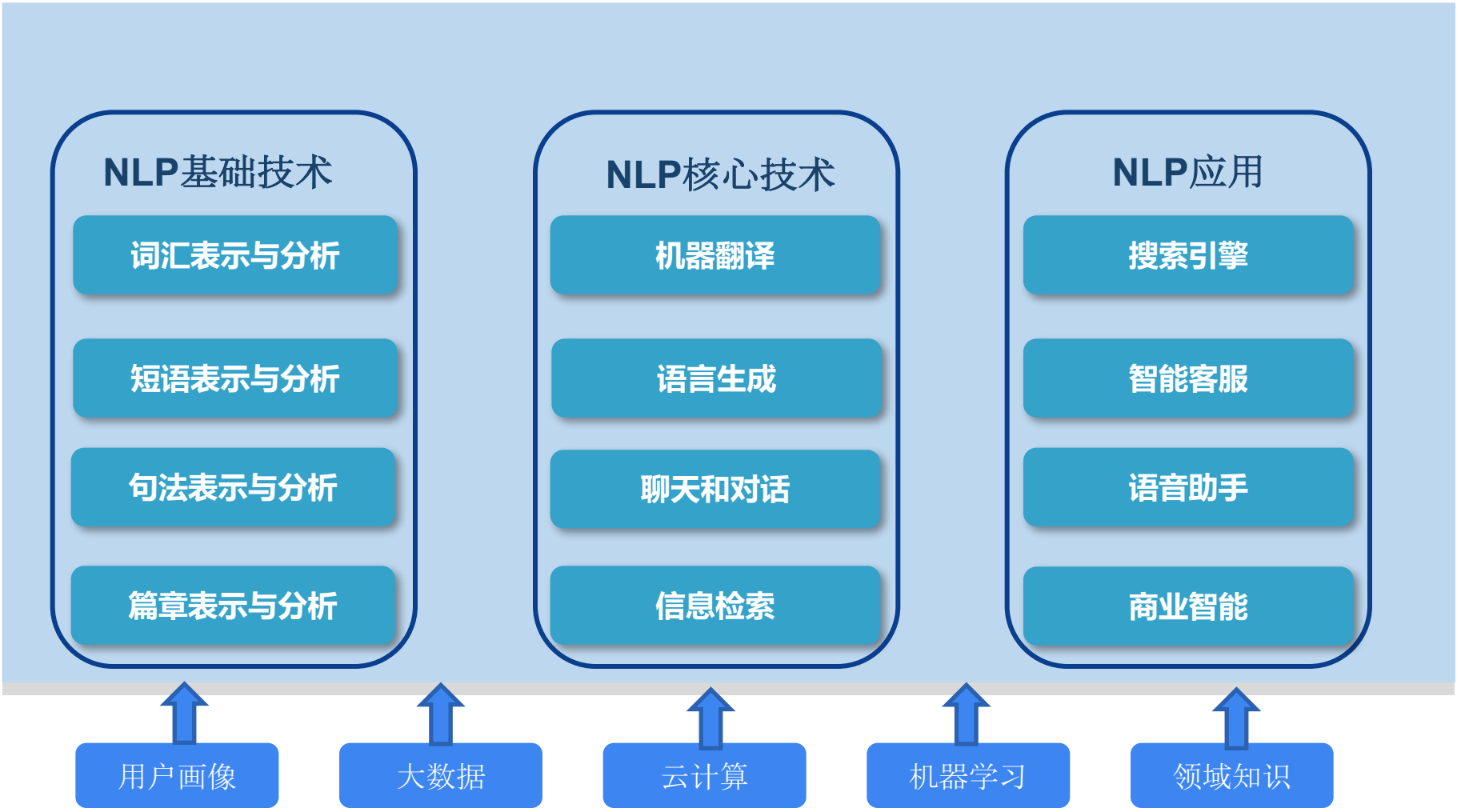
报告提纲

1. NLP回顾
2. 语义表示
3. 语义标签
4. 内容质量
5. 自然语言生成

NLP回顾

❖ NLP 概述

- 人工智能分支
- 机器分析语言
- 机器理解语言
- 机器生成语言



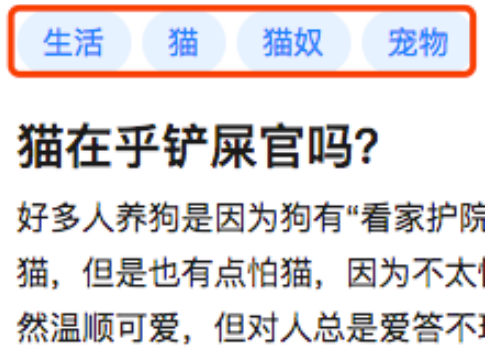
报告提纲

1. NLP回顾
2. 语义表示
3. 语义标签
4. 内容质量
5. 自然语言生成

语义表示 – 背景

❖ 知乎数据有什么特点？

- 每个问题绑定至少一个话题

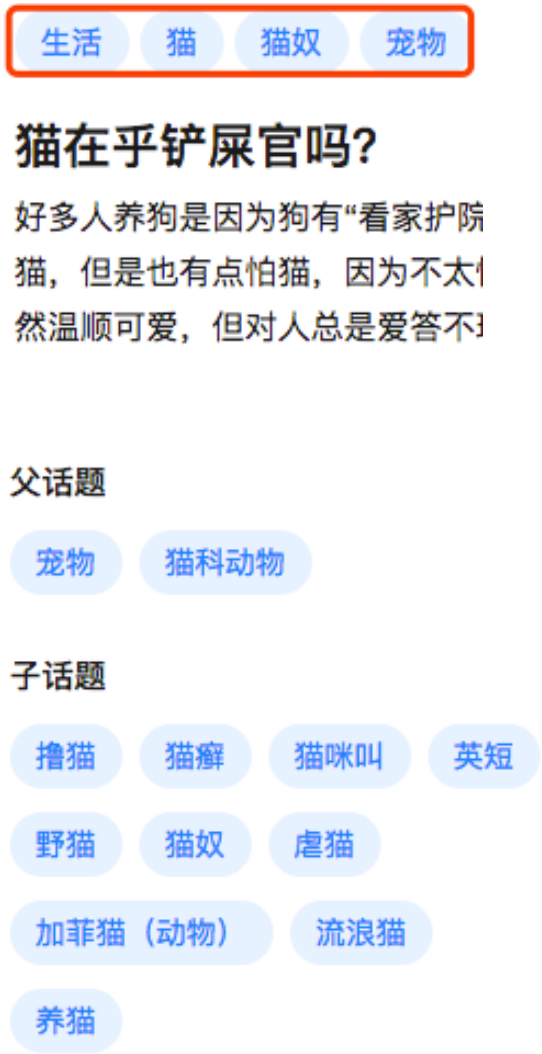


- 层次的话题结构

语义表示 – 背景

- 哪些需要表示？
  - 词：在乎
  - 短语：温顺可爱
  - 话题：猫
  - 话题层次关系：
    - (猫 -> 宠物)， (宠物 -> 猫)
  - 句子：猫在乎铲屎官吗？
  - 篇章：

好多人养狗是因为狗有“看家护院
猫，但是也有点怕猫，因为不太
然温顺可爱，但对人总是爱答不



语义表示 – 词表示

❖ 词向量 离散表示

- one-hot 表示，向量长度为词典大小

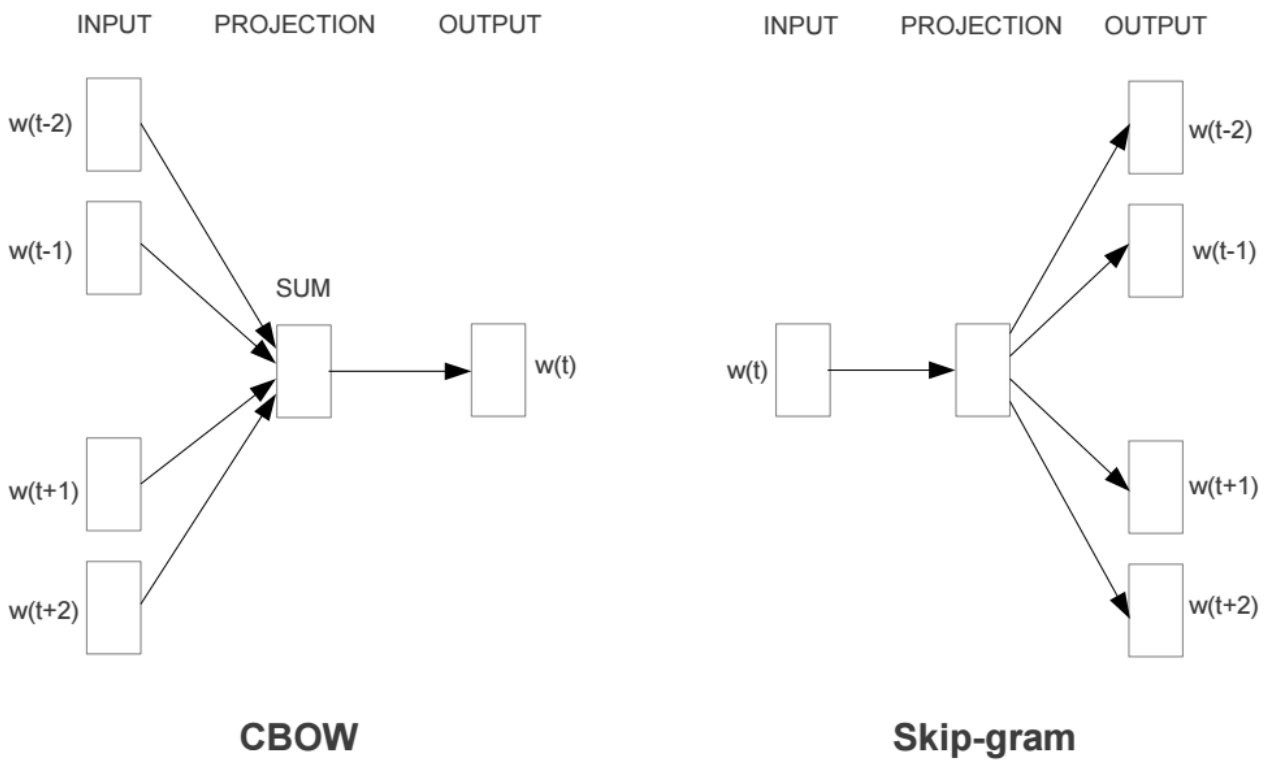
$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \cdots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

- 缺点：无法计算两个词之间的距离

$$(w^{hotel})^T w^{motel} = (w^{hotel})^T w^{cat} = 0$$

语义表示 – 词表示

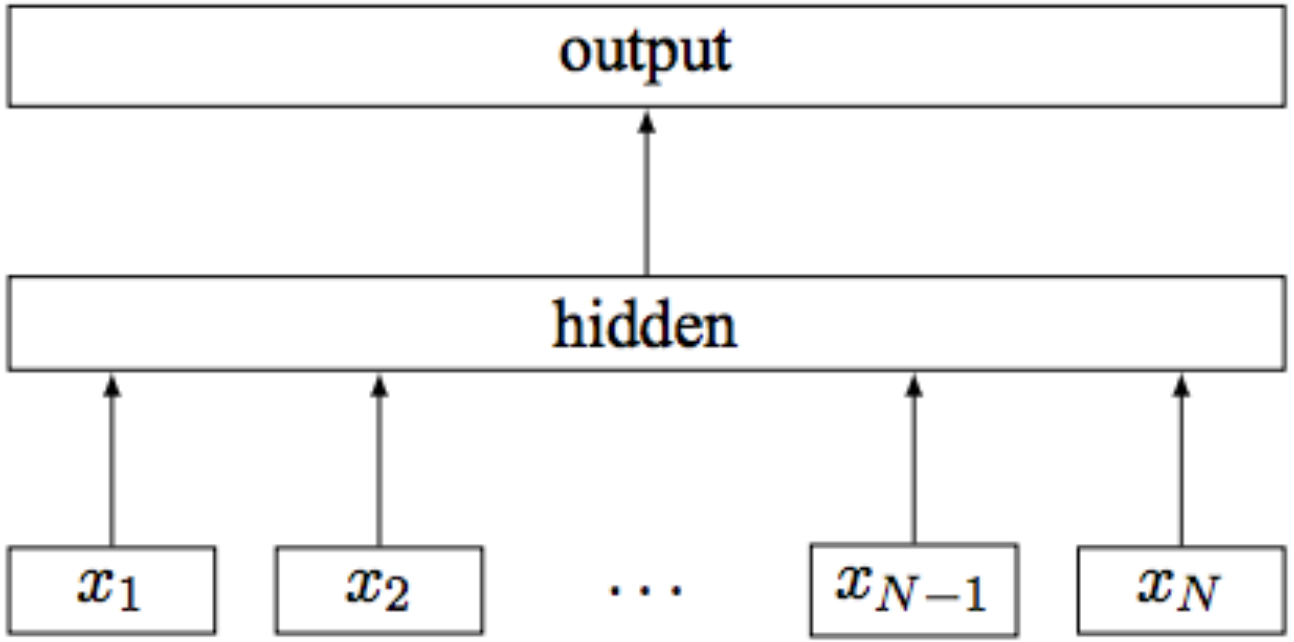
- 词向量 连续表示 cbow & skip-gram



语义表示 – 词表示

❖ 词向量 连续表示 – FastText

- 利用 label 信息
- 加入 n-gram 信息



语义表示 – 句表示

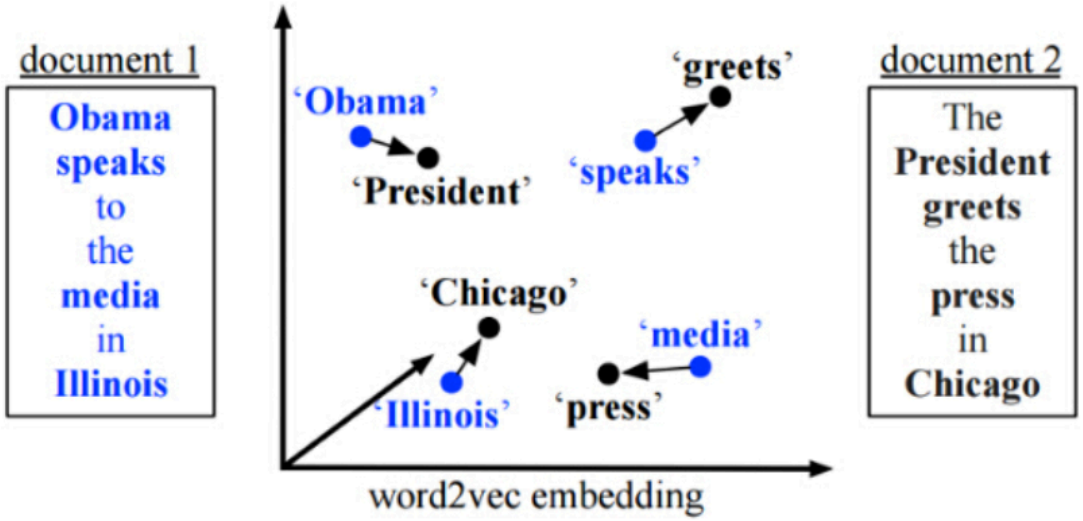
- **词袋模型 (bag of words)**
  - 如何评价罗永浩? = [0, 0, 1, 0, …, 1, 0, 0, 1,0]
- **词向量平均 (bag of embeddings)**
  - 直接平均
    - 如何评价罗永浩? = [0.2, 0.1, …, -0.3, 0.4] + [0.1, -0.2, …, 0.2, 0.5] + [0.5, 0.4, …, -0.3, -0.2]
  - idf 加权平均

语义表示 – 句表示

❖ 词向量平均的问题？

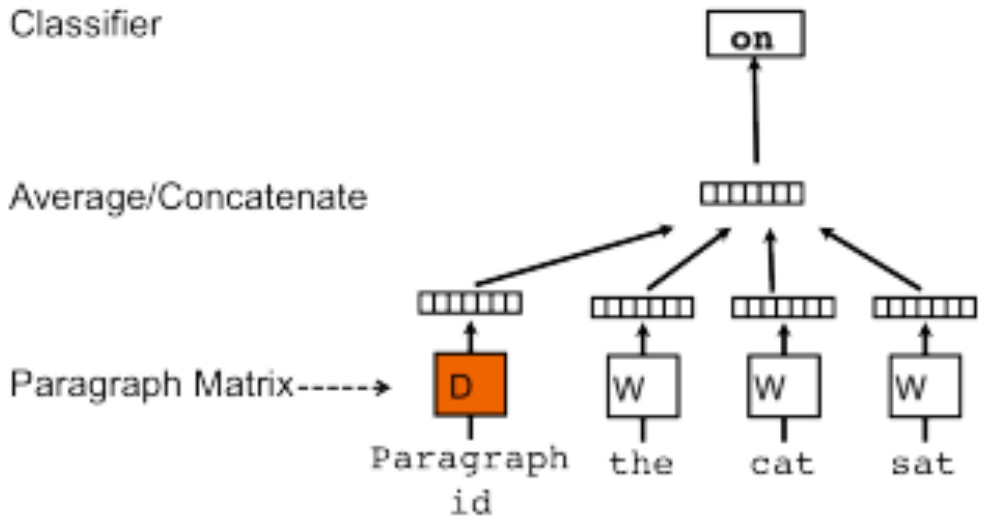
- You are going there to teach not play
- You are going there to play not teach

❖ Word Movers' Distance

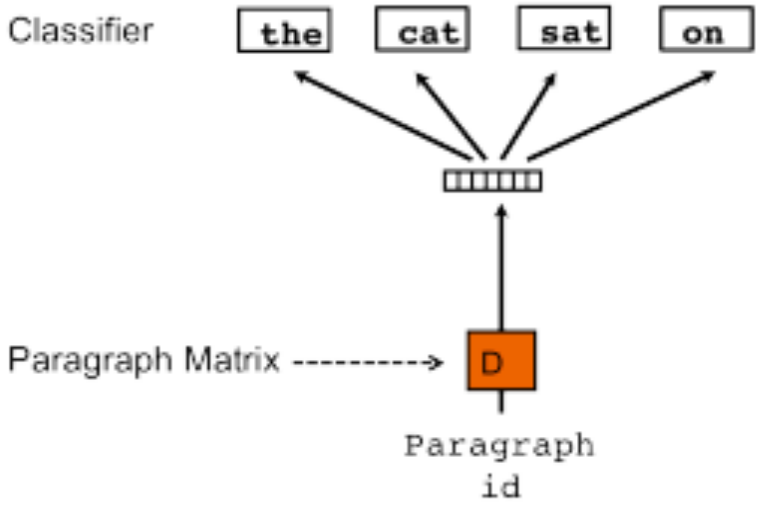


语义表示 – 句表示

❖ Doc2Vec - paragraph vector



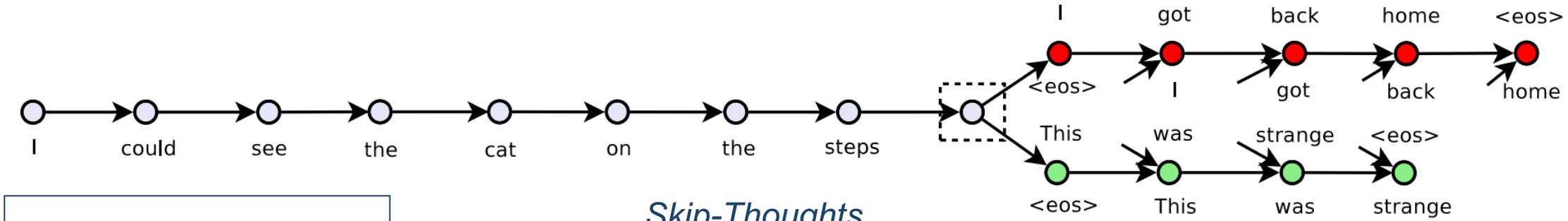
Doc2Vec-DM(Distributed Memory)



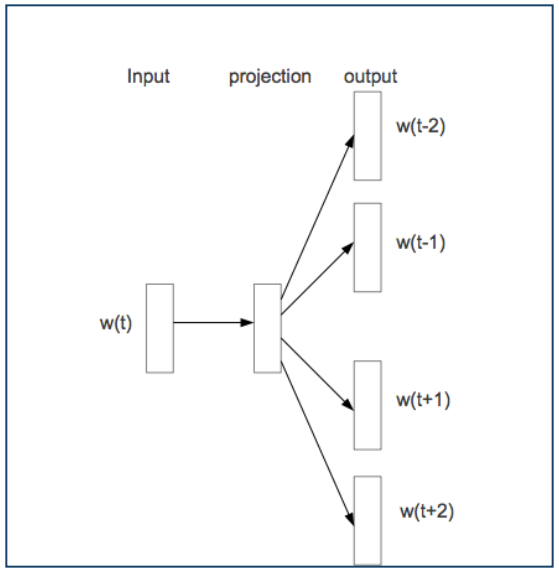
Doc2vec-DB(Distributed BOW)

语义表示 – 句表示

❖ Skip-thought



Skip-Thoughts



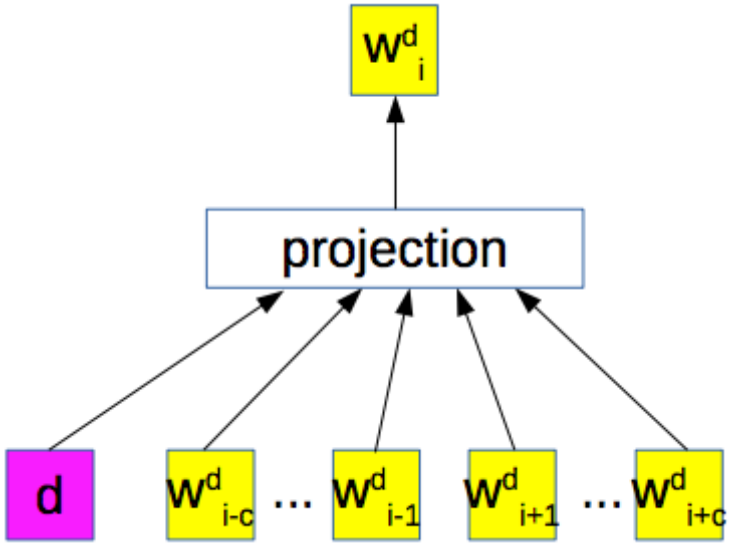
skip-gram

- **生成方式对上下文建模**
- **无监督训练**
- **工业界有较多的标签数据，使用场景少**

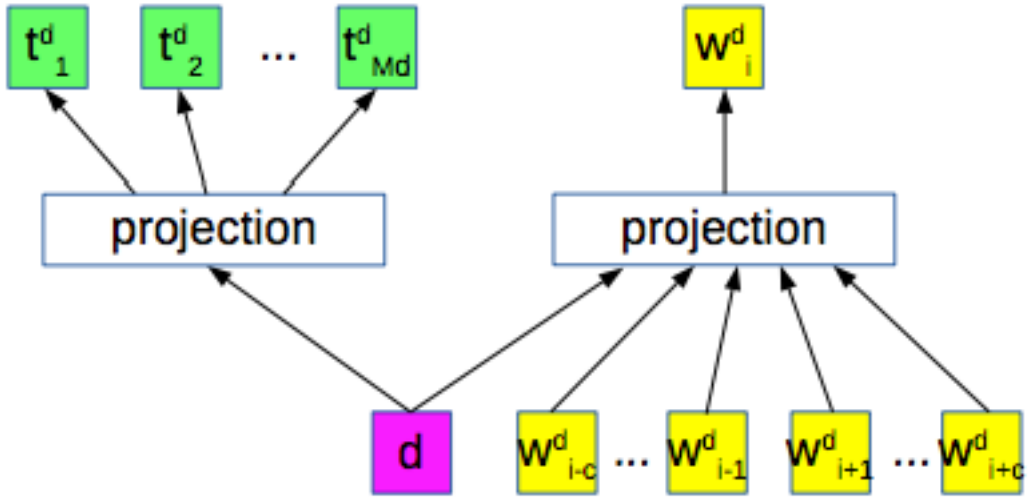
语义表示 – 话题表示

❖ 话题表示 - DocTag2Vec

- 话题、词、句子、文档表示在同一语义空间
- 新话题的增量训练



Doc2Vec(Distributed Memory)



DocTag2Vec

语义表示 – 层次化话题表示

❖ 层次化话题结构有哪些特点？

- 非对称关系
  - 猫 ∈ 宠物
  - 宠物 ∉ 猫
- 多层结构
  - 猫 ∈ 宠物 ∈ 动物 ∈ 物体

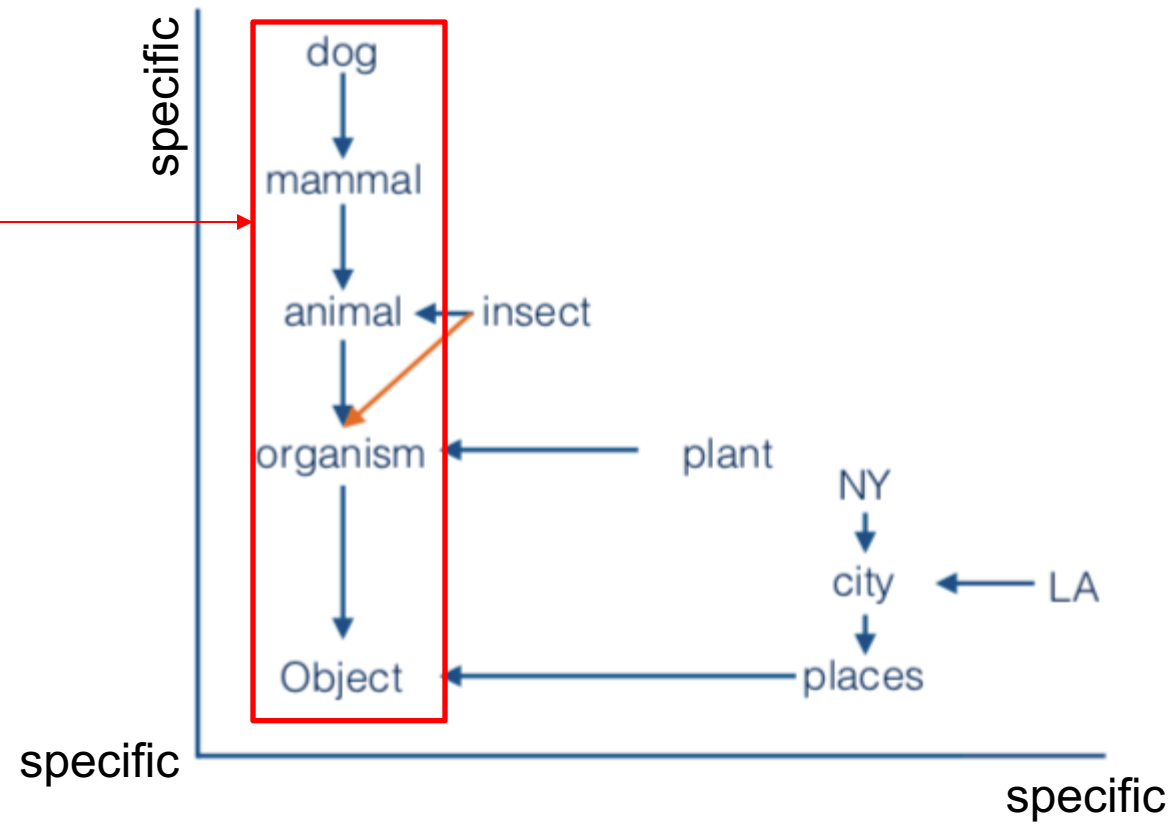
语义表示 – 层次化话题表示

❖ 层次化话题表示：Order Embedding

- 表示为分布在第一项限的向量
- 层次关系？
  - $x \rightarrow y$

$$\forall i, x_i \geq y_i$$

- 话题越靠近原点概念越泛



语义表示应用：相关文章/广告推荐

语义表示应用 – 话题表示应用

❖ 用户兴趣画像扩展

话题兴趣(知乎自带)

id	name
t99902	施一公
t12473	王垠（人物）
t7416	论文
t51843	转专业
t18488	电动车
t8080	惊悚电影
t522	跳槽
t1115	PPT
t161530	瑞克和莫蒂（Rick and Morty）

施一公 → 清华大学

论文 → 科研、学术

电动车 → 特斯拉、新能源汽车

语义表示应用 – 话题图谱构建

❖ 话题图谱清洗与构建：

报告提纲

1. NLP回顾
2. 语义表示
3. 语义标签
4. 内容质量
5. 自然语言生成

语义标签

❖ 语义标签示例

- 一级领域： 科技
- 二级领域： 互联网
- 话题： 罗永浩、锤子科技
- 实体： 坚果手机、苏宁集团、贾跃亭、阿里巴巴、苹果公司
- 时效性： 高时效

罗永浩真要凉了

随后推出的坚果系列，更是直接掉到了千元档，唉，老罗信誓旦旦的逼格，碎了一地。苹果从不做中低端，向苹果看齐的老罗，旗舰手机没做出来，反而坠入千元机市场了。**失去逼格，还是老罗吗?**

不仅是产品危机，6年来，锤子还经历了供应链危机，严重缺货，良品率低，让粉丝一等再等。更严重的危机来自钱，资金链几次面临断裂，按老罗的话，工资都发不出来了。

为了筹钱，老罗也是拼了，找苏宁求入股，找阿里做股权质押；据说，还求上了尚在国内的老贾。

语义标签

❖ 多种粒度标签的要求

- 分类，完备、尽量正交的分类体系，保证任一问题/文章能分到某个类别
- 实体，高准确度，保证热门实体被召回
- 话题，高准确度，同一个问题/文章可打上多个话题

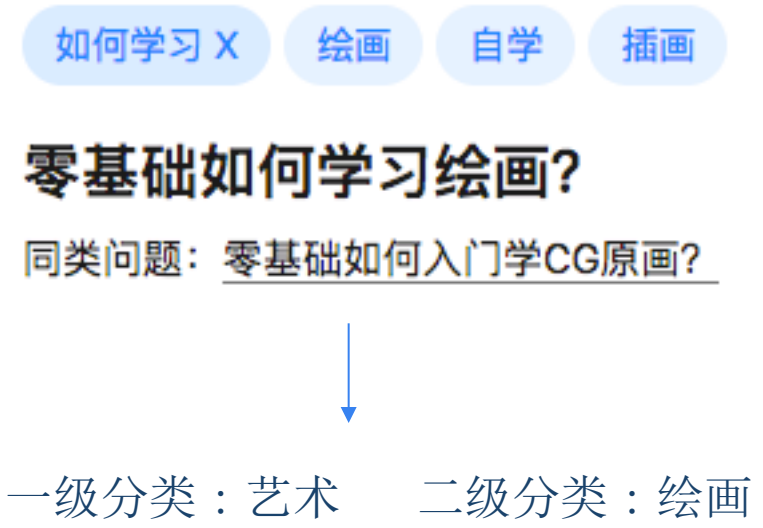
语义标签 – 领域分类

❖ 问题？

- 只有分类体系，站内话题数据，无类别标注数据

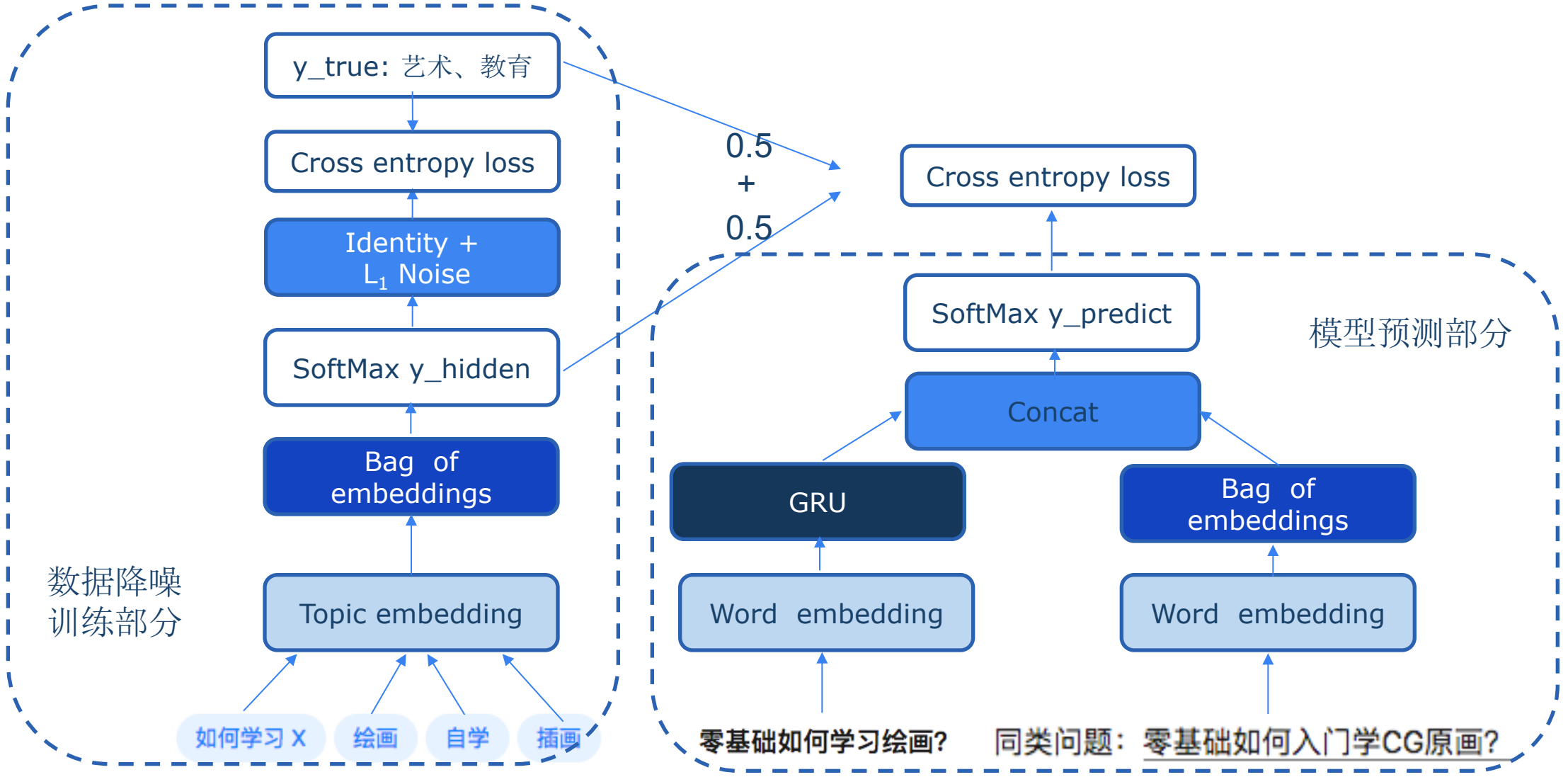
❖ 类别训练数据构建

- 话题到类别映射
  - 绘画 -> 艺术
  - 自学 -> 教育
- 子话题递归
  - CG绘画 -> 绘画 -> 艺术
- 相关话题 PMI



语义标签 – 领域分类

❖ 基于话题映射的带噪声训练数据怎么训练？



语义标签 - 实体识别

罗永浩真要凉了

随后推出的坚果系列，更是直接掉到了千元档，唉，老罗信誓旦旦的逼格，碎了一地。苹果从不做中低端，向苹果看齐的老罗，旗舰手机没做出来，反而坠入千元机市场了。**失去逼格，还是老罗吗?**

不仅是产品危机，6年来，锤子还经历了供应链危机，严重缺货，良品率低，让粉丝一等再等。更严重的危机来自钱，资金链几次面临断裂，按老罗的话，工资都发不出来了。

为了筹钱，老罗也是拼了，找苏宁求入股，找阿里做股权质押；据说，还求上了尚在国内的老贾。

分词&词性标注

罗永浩 nr 真 要 v 凉了 v
……

抽取候选

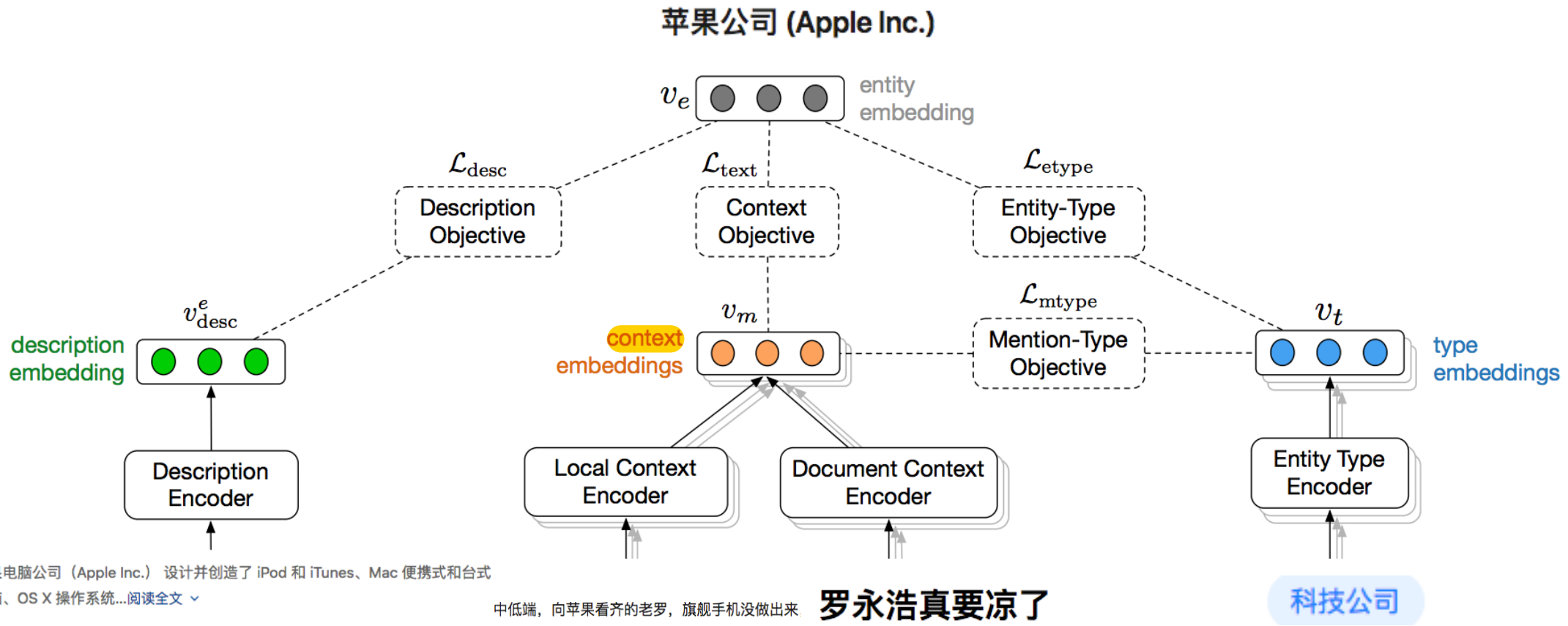
坚果手机、坚果(植物)、苹果公司、苹果(水果)
……

消歧&相关性计算

坚果手机 0.9、苹果公司 0.6
……

语义标签 – 实体识别

❖ 融合实体类别、上下文、描述的实体表示



语义标签 – 实体识别

❖ 基于实体表示的消歧与相关性计算

- P(苹果公司|苹果) = P_{prior}(苹果公司|苹果)
  + P_{text}(苹果公司|苹果)
  - (P_{prior}(苹果公司|苹果) * P_{text}(苹果公司|苹果))

- P_{text}(苹果公司|苹果) = $P(v_e|v_m)$

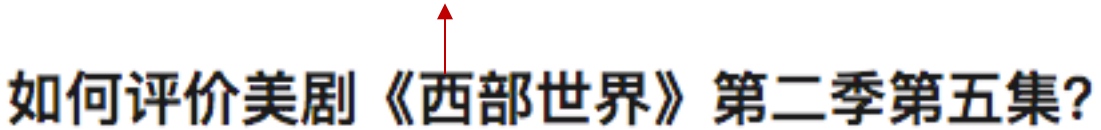
语义标签 – 话题标注

❖ 问题定义

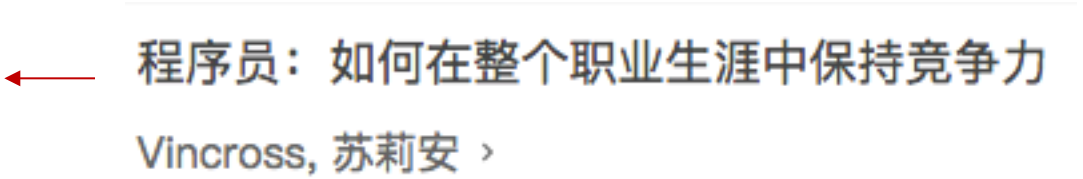
- 给定一段文本，从给定话题集合中匹配出相应话题

❖ 应用场景

- 问题话题标注
  - 西部世界、西部世界第二季、美剧、人工智能

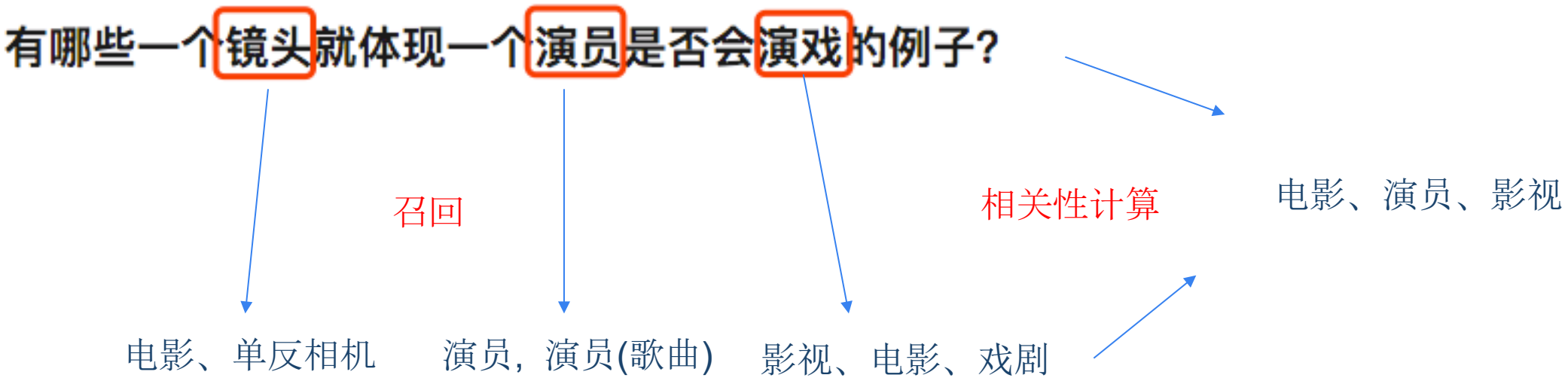


- Live 话题标注
  - 程序员、职业发展



语义标签 – 话题标注

- 话题标注策略 召回 + 相似度计算



语义标签 – 话题标注

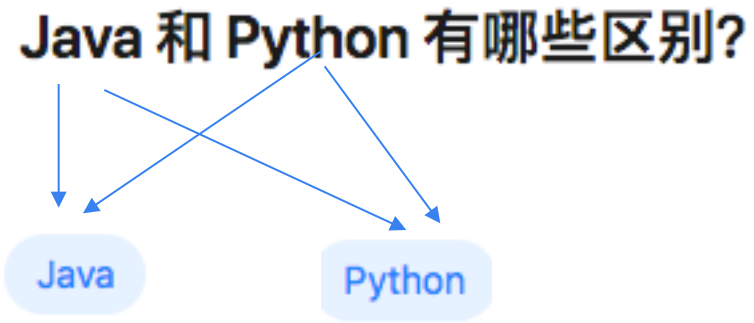
❖ 候选话题召回

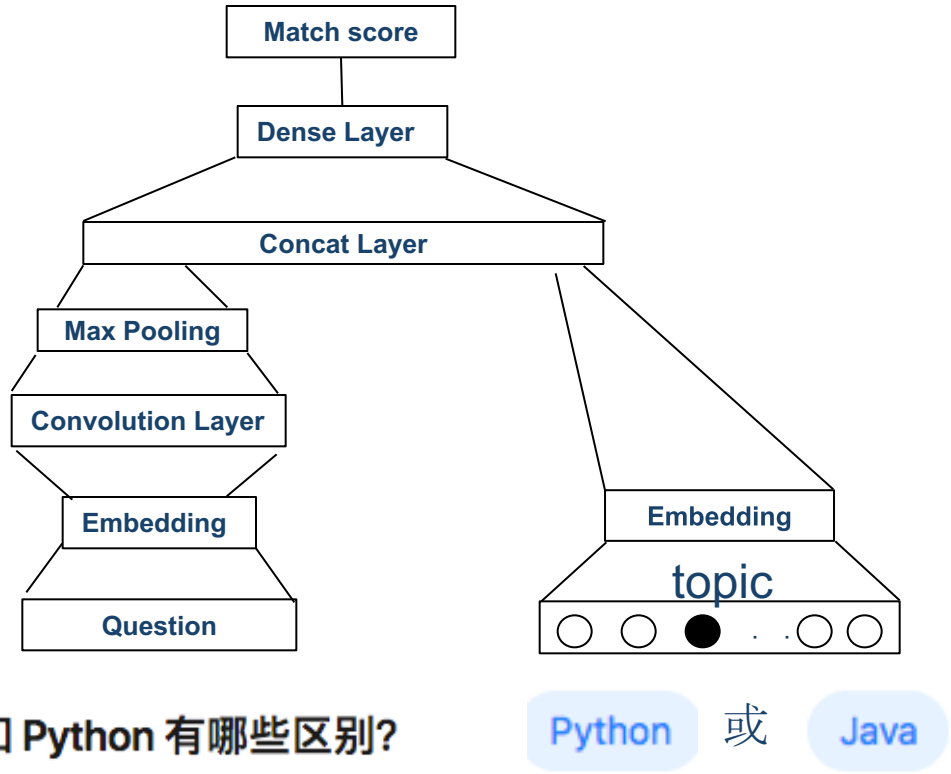
- 热门话题
- 基于词和话题的 PMI 和 NDG
- 话题和话题的 PMI
- 完全匹配话题
  - 话题 alias：(科比 ->科比·布莱恩特)
- 话题子父节点关系

语义标签 – 话题标注

❖ PointWise Matching CNN

- 缺点：
  - 模型无法区分经常共现的兄弟节点

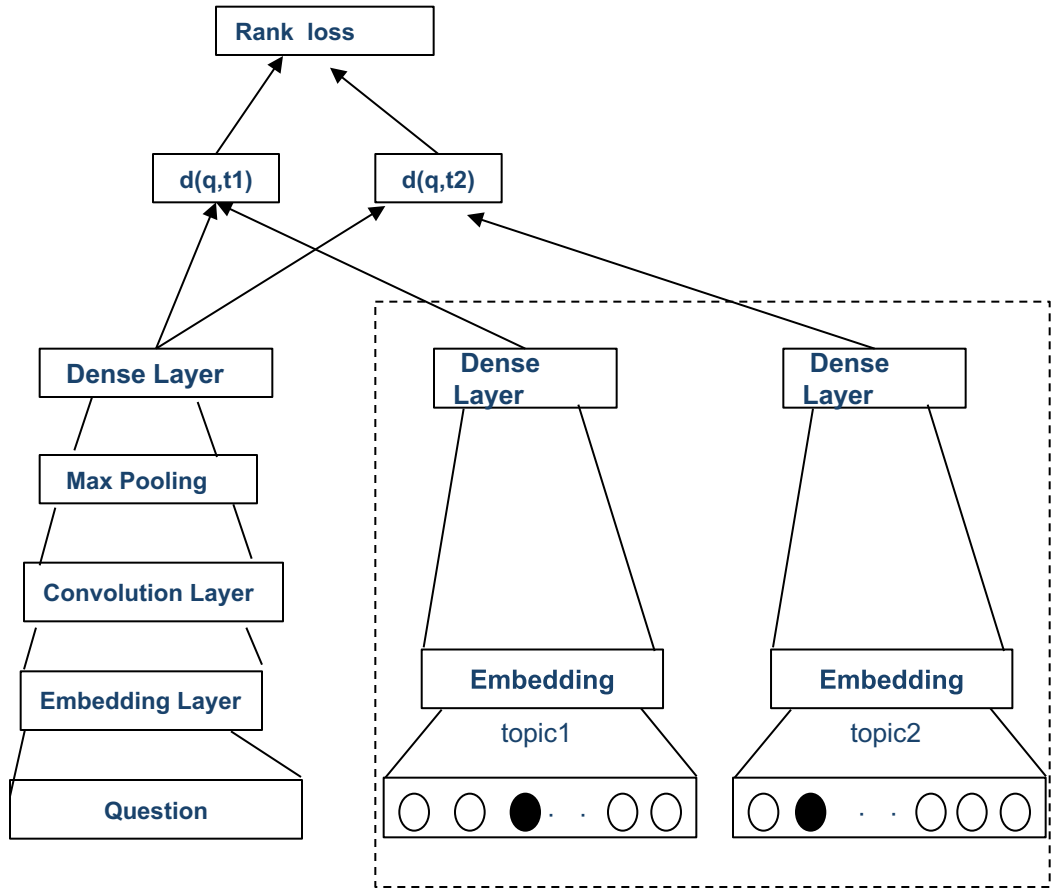




语义标签 – 话题标注

❖ PairWise Matching CNN

- 对兄弟节点负采样
  - Python -> java
  - 北京 -> 上海
- 对儿子节点负采样
  - 艺术 -> 绘画



如何理解Python装饰器?

Python +

Java -

语义标签 – 话题标注

❖ 未来的工作

- 与分类一起做multi-task
- 模型加入话题结构

❖ 相关比赛

- 知乎看山杯： https://biendata.com/competition/zhihu/
- NLPCC 2018 知乎话题标注：
  http://tcci.ccf.org.cn/conference/2018/dldoc/taskgline06.pdf

语义标签 – 时效性标签

- 高时效

机器学习 计算机科学 NIPS

如何看待 NIPS 2018 submission达到近 5000 篇?

- 低时效

中国历史 中国古代历史 历史人物

中国历史上有哪些被低估的皇帝?

- 周期性

大学新生

大一开学自我介绍怎么让他人记住?

作为大一新生 要自我介绍了 班里女生比较少 我想借自我介绍的机会让女生对我印象深刻 不要太奇葩

语义标签 – 时效性标签

❖ 高低时效二分类

- 训练数据
  - 文本的被浏览数据
- 模型
  - TextCNN文本分类模型

❖ 周期性

- 话题的被搜索行为数据分析，知乎指数、百度指数

语义标签应用：个人性化新闻推荐

程序员 – 张三

运营 – 小丽

报告提纲

1. NLP回顾
2. 语义表示
3. 语义标签
4. 内容质量
5. 自然语言生成

内容质量

应用

内容质量分级

核心技术

低质特征

- 低俗&谩骂
- 答非所问
- 广告
- 标题党
- 虚假谣言

价值特征

- 专业性
- 权威性
- 观点倾向
- 可读性
- 内容深度
- 内容源质量

后验特征

- 用户显式反馈
- 用户隐式反馈

数据

内容文本

用户行为

内容质量 – 专业性

❖ 难点：没有训练数据？

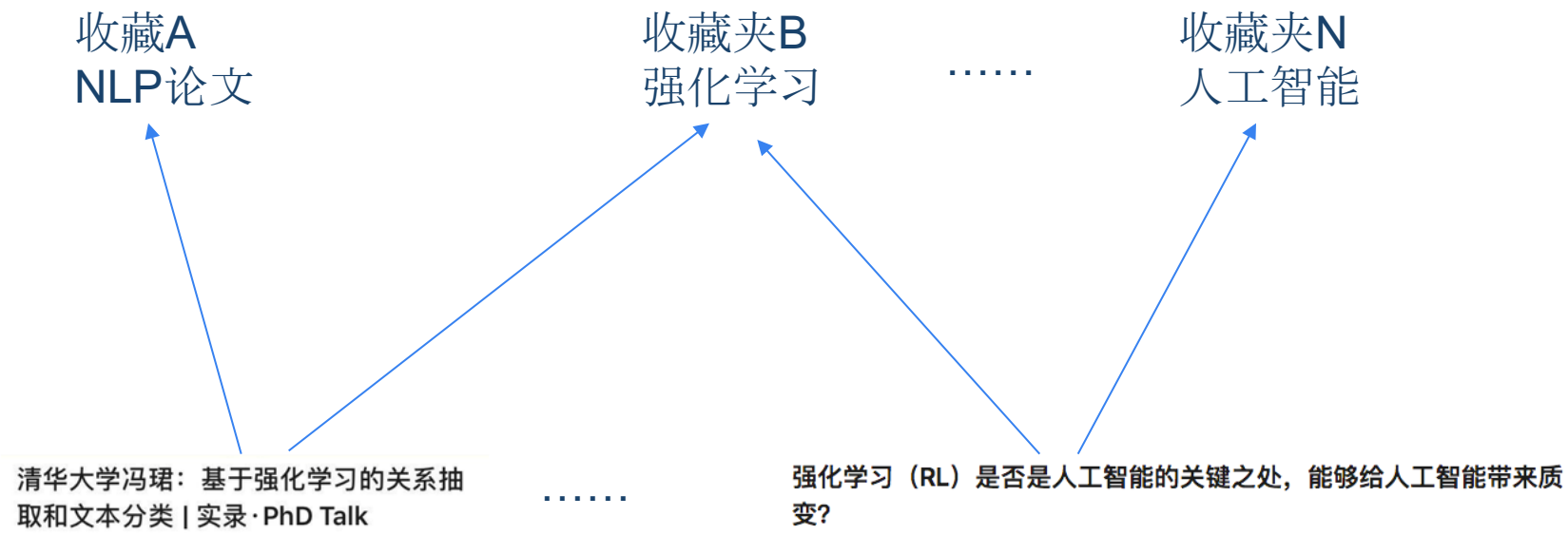
- 专业回答/文章 -> 用户更倾向于收藏
- 抖机灵回答 -> 用户更倾向于点赞

抖机灵回答

某收藏夹内容

内容质量 – 专业性

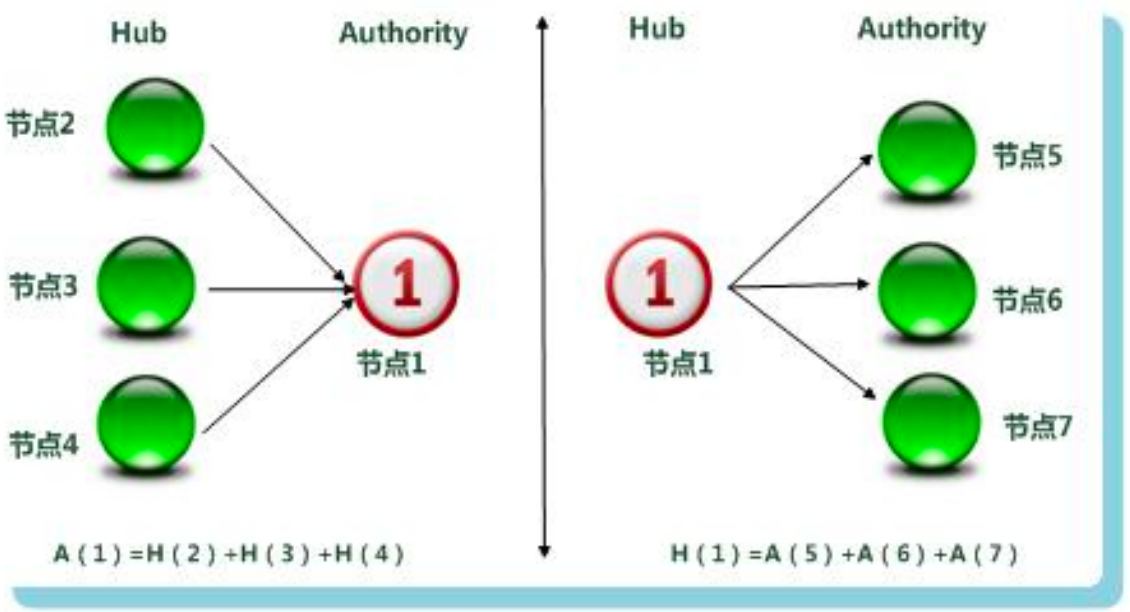
❖ 收藏夹数据



内容质量 – 专业性

❖ 专业性识别 - Hits 算法

- 基本假设1：一个好的“Authority”页面会被很多好的“Hub”页面指向；
- 基本假设2：一个好的“Hub”页面会指向很多好的“Authority”页面



内容质量 – 专业性

❖ Hits 算法的问题？

- 未考虑作者与收藏者特征
  - 在算法中加入作者和收藏者权重
- 部分段子/经历也经常被收藏
  - 基于标题语义的专业性模型

大家在医院实习的时候都遇到过哪些有趣或者奇葩的老师?

非专业

为什么去医院看感冒要做心电图?

专业

❖ 未来的工作

- Graph model

内容质量 – 答非所问

❖ 示例

跳绳的好处有哪些？可以锻炼哪些肌肉？

A：心肺功能比之前有提高。

B：有助于提高身体的乳酸阈值。

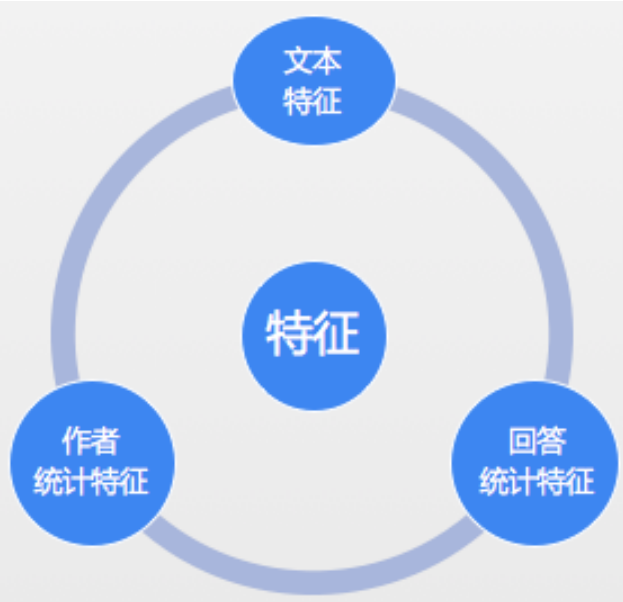
C：有助于提高身体的协调性。

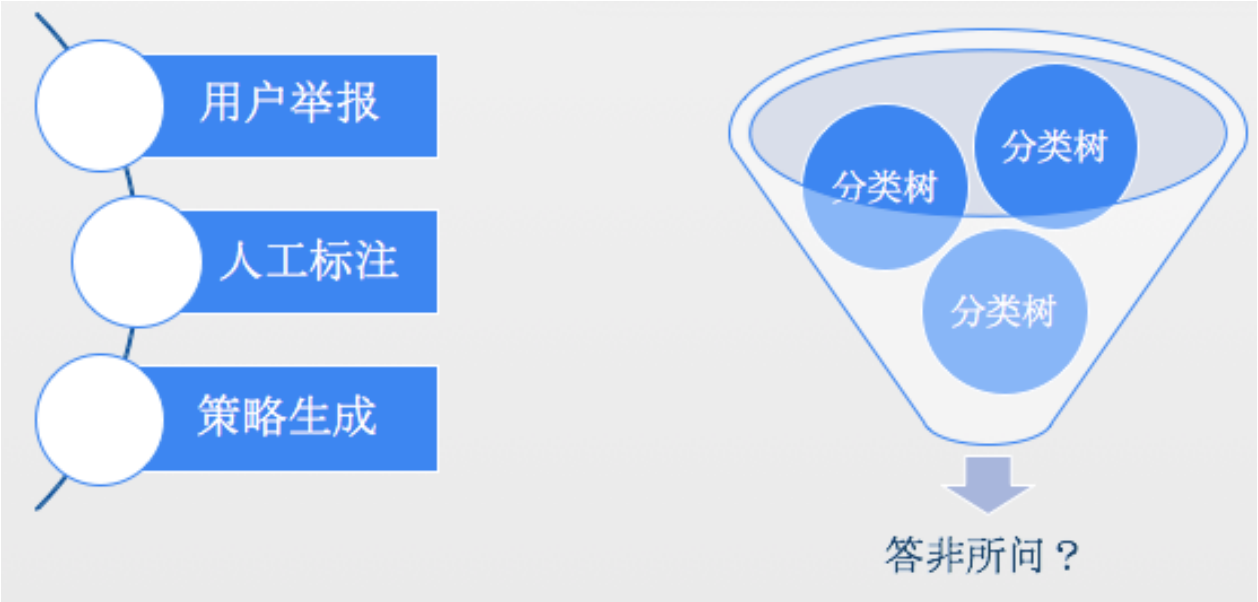
D：谢谢，我去买了跳绳。

请问，以上哪个答案是答非所问？

内容质量 – 答非所问

- 随机森林

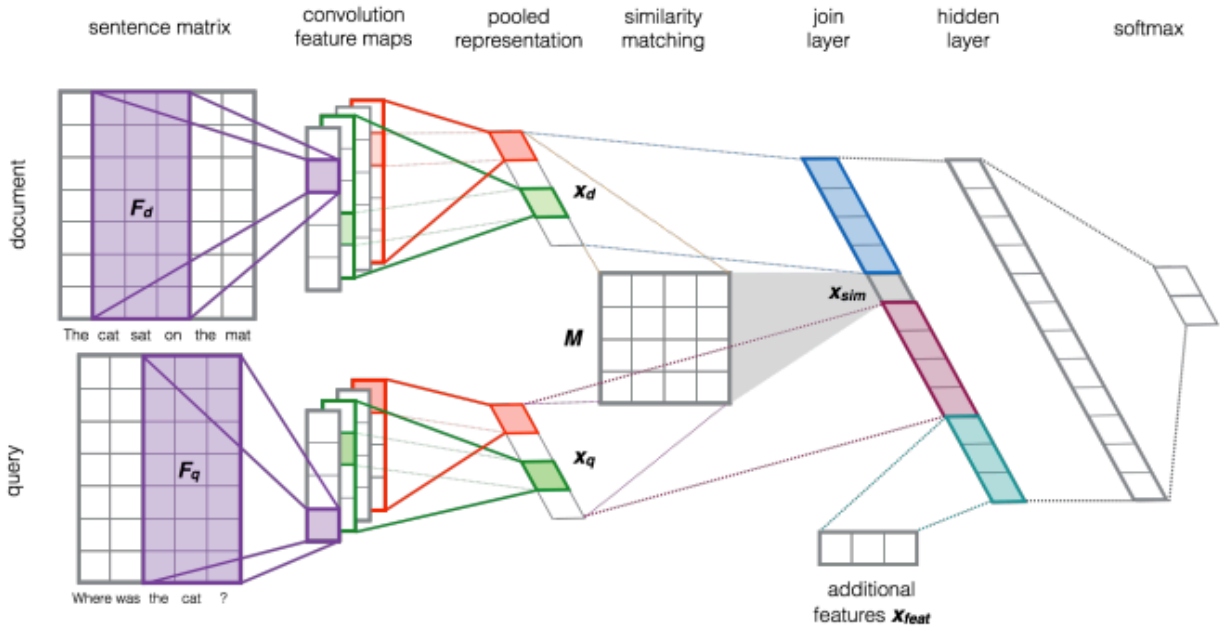




特征工程 → 样本构造 → 模型训练

内容质量 – 答非所问

- **随机森林：**
  - Precision 97%
  - Recall 58%
  - 自动折叠
- **CNN 模型：**
  - Precision 78%
  - Recall 80%
  - 人机配合



答非所问：Siamese 网络结构

内容质量 – 应用

❖ 机器人瓦力

- 每天清理约 **5000** 条新产生的「答非所问」内容
- 现存的 **115** 万条「答非所问」内容

❖ 专业性内容作者扶持

- 专业内容加量分发
- 作者专业度评价

报告提纲

1. NLP回顾
2. 语义表示
3. 语义标签
4. 内容质量
5. 自然语言生成

自然语言生成 – 自动提问

愿景：

- **对新闻自动问题**

上班雪天路滑 司机开车较快 车轮打滑出了车祸 副驾驶当场死亡 ⟶ 雪天开车有什么技巧？

- **对Query 到问题**
  - 跳槽 薪资 ⟶ 跳槽时怎么谈薪资？ 你会因为薪资问题跳槽吗？
- **对电子书自动提问**
- **对电影自动提问**

自然语言生成 – 自动提问

❖ 挑战

- 业界研究任务：生成客观问题， 问法单一，通常有标准答案
- 我们目标：生成主观问题，问法多样，生成一对多的问题
- 业界数据：SQuAD，Knowledge Graph
- 我们数据：知乎站内问题回答数据
  - 噪声多
  - 类型杂

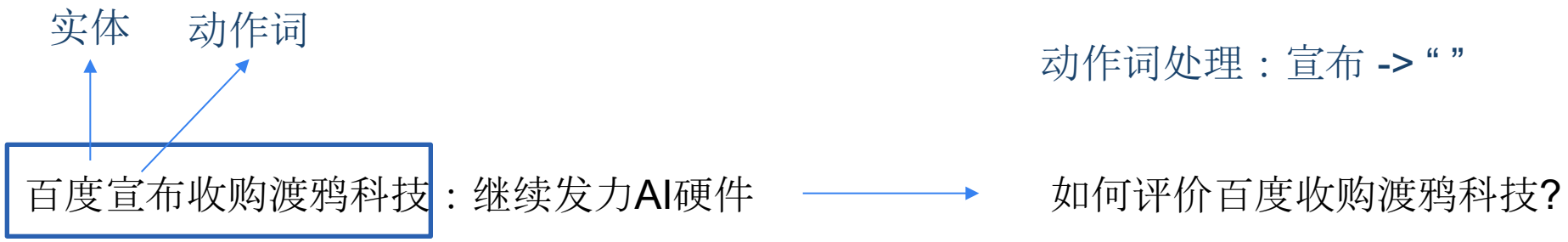
自然语言生成 – 自动提问

一期：基于新闻的问题生成

- 人工配置模板
- 自动挖掘模板
- Seq2seq 模型

自然语言生成 – 自动提问

❖ 人工配置模板： 如何评价 + 部分标题内容



❖ 问题

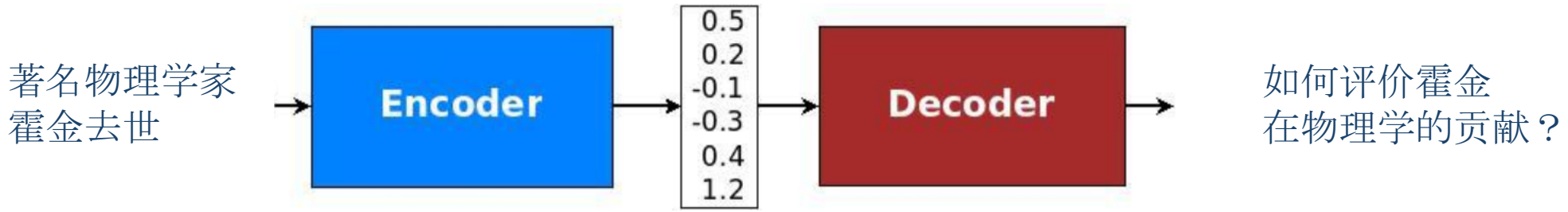
- 模板命中率较低
- 提问角度单一

自然语言生成 – 自动提问

- **自动挖掘模板**
  - 如何评价杨振宁在物理上的贡献 -> 如何评价x在y上的贡献
  - 杨振宁在物理上能排第几 -> x在y能排第几
- **预测模板 + 填空**
  - 新闻 “著名物理学家霍金去世” ->
    - 如何评价霍金在物理学上的贡献
    - ….
    - ….
- **问题重排序**
  - 模板与新闻相关性
  - 填空后问题的流畅度
- **问题**
  - 知乎站内大部分为非时效模板，在新闻上命中率较低

自然语言生成 – 自动提问

❖**seq2seq 生成模型**



❖思考：训练数据怎么来？

自然语言生成 – 自动提问

- **训练数据 – 多源混合**
  - 问题描述 -> 问题
  - 问题回答 -> 问题
    - 回答按点赞数筛选
    - 选回答与问题相似度较高的段落
  - 新闻正文 -> 标题
    - 选取正文与标题比较相似的段落

自然语言生成 – 自动提问

模型效果

- 留学生持霰弹枪击退3名非法入侵男子当场击毙1人
  - 美国有哪些合法的枪支？
  - 留学生如何保护自己？
- 女子掌握公司内幕后离职家门口被陌生人放一颗心脏
  - 被同事陷害了怎么办？
- 习近平这一年忙了啥 看看他的2016业绩表民生改善7次下基层调研扶贫开发
  - 2017<unk><unk>规划将将对北京经济发展带来哪些影响？

自然语言生成 – 自动提问

❖End2end 生成的问题

- 问题时效性较差
- 与新闻的相关度较弱
- 存在oov的情况

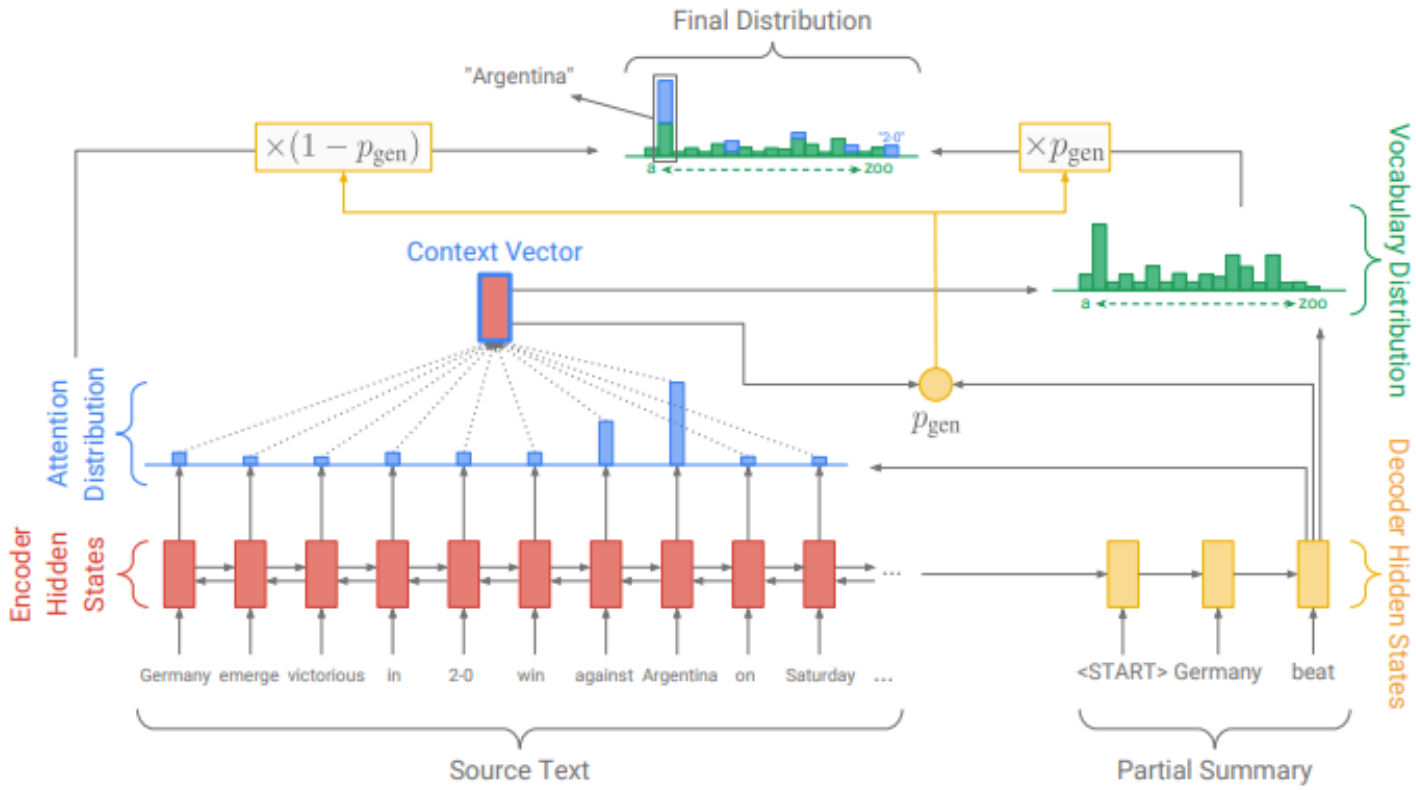
❖怎么改进？ 新闻标题＋模型生成后缀

- 新闻“国土部调控发挥作用一二线城市住宅地价增速回落”＋后缀 “是否会导致通货膨胀？”
- 贵州男子回乡途中贪便宜分到冥币被骗7万元 ＋如何处理？
- 日媒日本自卫队将实施设想台海两岸冲突演习＋是否有什么阴谋？

自然语言生成 – 自动提问

❖其他值得尝试的方案？

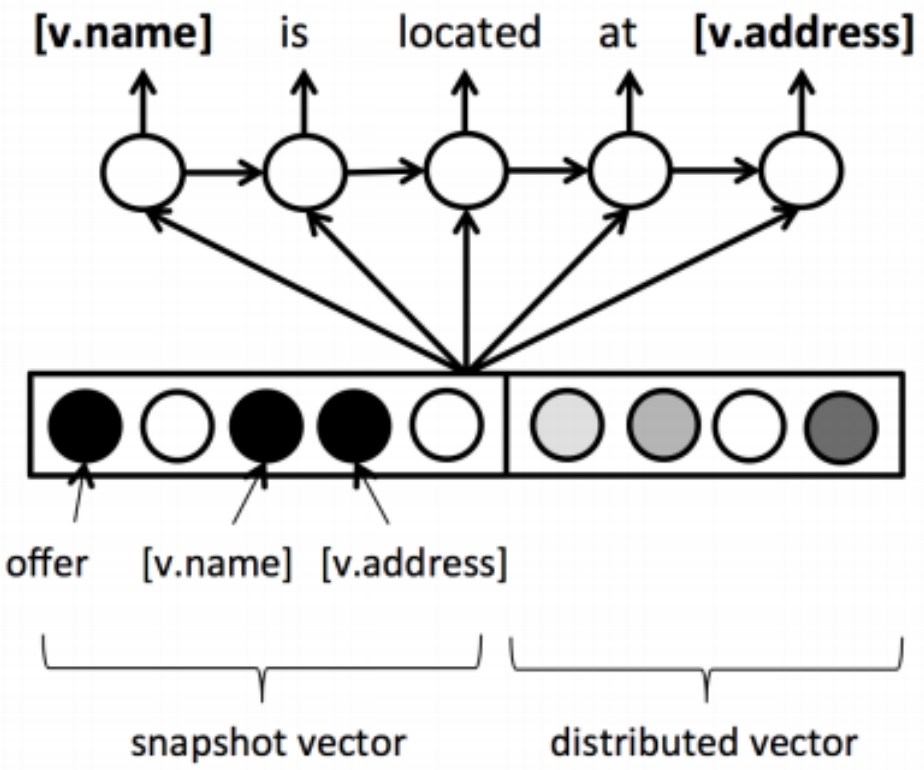
- Pointer generator network – 解决 oov 问题
  - $P_{\text{gen}} \in [0, 1]$
    - 从词典生成的概率
  - $1 - P_{\text{gen}}$
    - 从原文拷贝的概率



自然语言生成 – 自动提问

❖其他值得尝试的方案？

- Conditional Generation Network – 解决相关性问题
  - 扔进几个关键词 -> 拼成一个问句
    - 百度、陆奇、离职 -> 陆奇为什么会从百度离职？
  - 关键词提取
    - 标题实体
    - 标题正文出现次数较高的词



自然语言生成 摘要生成展望

❖ 知乎文本摘要的价值

- 单个长回答的核心内容概述，节省用户浏览时间
- 多种观点聚合，方便用户快速了解大部分的观点等
  - 成都的发展前景怎么样？上海交大和中科大各有什么优势？

❖ 挑战？

- UGC内容，不同领域、作者的回答风格迥异 –> 配置模板难度大、收益小
- 训练数据怎么来？ -> 模型训练数据准备成本高
- 生成式摘要的可读性怎么去衡量 -> 线上应用效果难保证
- 部分开放式问题的主观回答本身就不可聚合 -> 问题筛选成本

Thank You!

Questions ?