



基于机器学习的 搜索语义理解技术

陈翔

百度资深研发工程师

WOTD

World Of Tech
2017年12月1-2日

全球软件开发技术峰会

[深圳站]

报名咨询：010-68478816

议题提交：wot@51cto.com

市场合作：yangxh@51cto.com

商务合作：songjc@51cto.com

媒体合作：yankk@51cto.com

在线咨询（微信）：18401576051

团·购·享·受·更·多·优·惠

5折 优惠（截止8月31日）
现在报名，立省1400元/张

CONTENT

目录

- | 人工智能：搜索的二次进化
- | 通用需求理解模型
- | 领域语义理解模型
- | 深度学习的应用技巧

1

人工智能：
搜索的二次进化

搜索场景智能化



>> 语音搜索

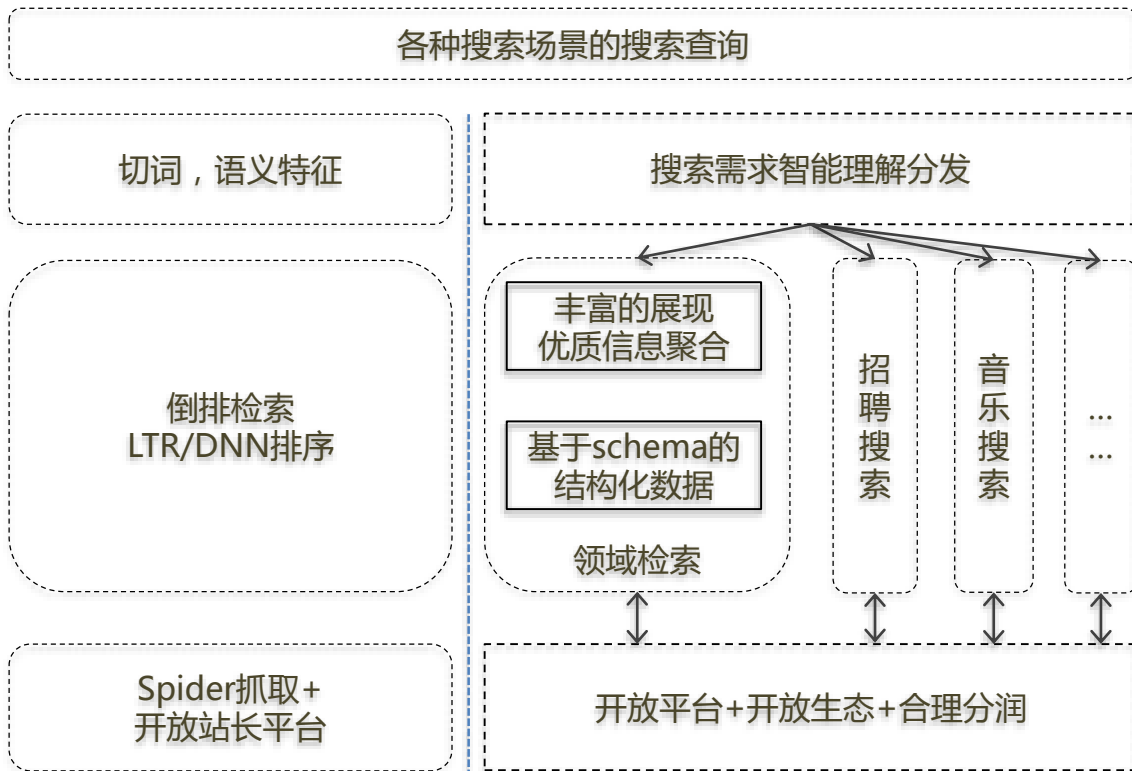


>> 图像搜索



>> 物联网等其他载体

搜索生态进化



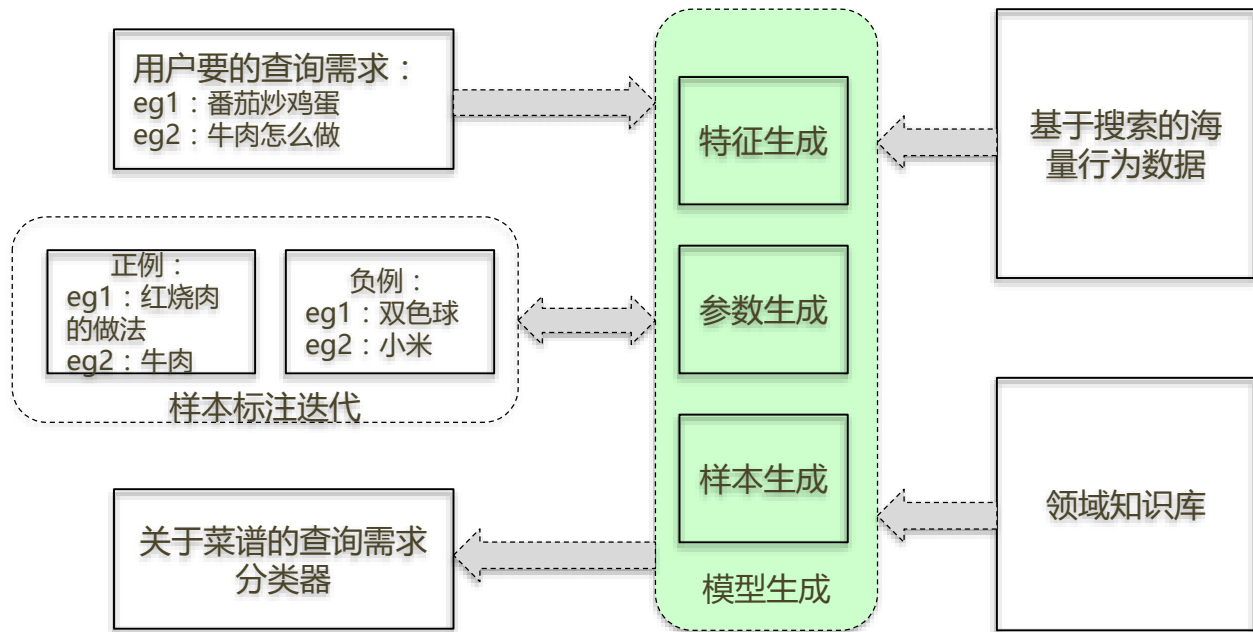
- 通用+领域检索智能聚合
- 基于大数据+机器学习的算法内核
- 基于开放生态的更丰富数据

2

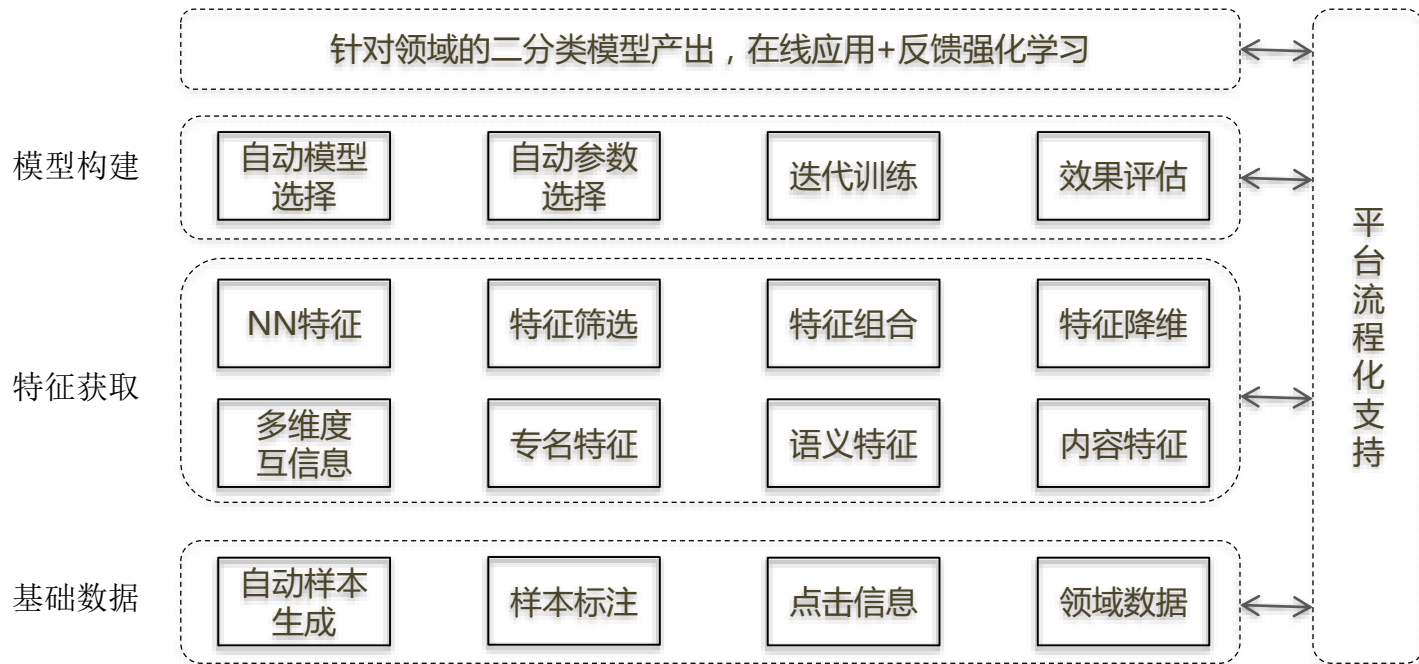
需求理解：

通用语义需求识别模型

需求理解



策略概览



通过随机游走，产生更丰富的特征信息

- Input : 手机百度下载
- Features :
 - Template : [Appname]下载/
手机[Appname]下载
 - Term : 手机百度/下载/手机/百度
 - Related Url-pattern :
 - mobile.baidu.com/
 - dl.pconline.com.cn/download/
 - www.wandoujia.com/apps/
 - ...
 - Sim-query features :
 - Features of “手百下载”
 - Features of “手机百度app下载”
 - ...
- 基于策略控制的随机游走+标注
- 百万维基础特征

Queries Q

q_1 : jobs in chicago
 q_2 : jobs in boston
 q_3 : jobs in microsoft
 q_4 : jobs in motorola
 q_5 : marketing jobs in motorola
 q_6 : 401k plans
 q_7 : illinois employment statistics

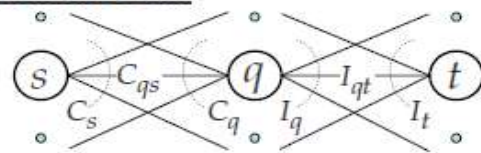
Sites S

s_1 : monster.com
 s_2 : motorola.com
 s_3 : us401k.com

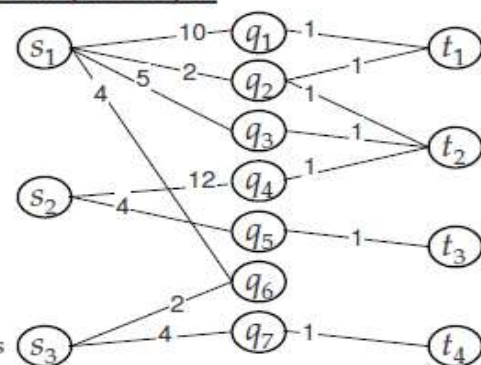
Templates T

t_1 : jobs in #location
 t_2 : jobs in #company
 t_3 : #category jobs in #company
 t_4 : #location employment statistics

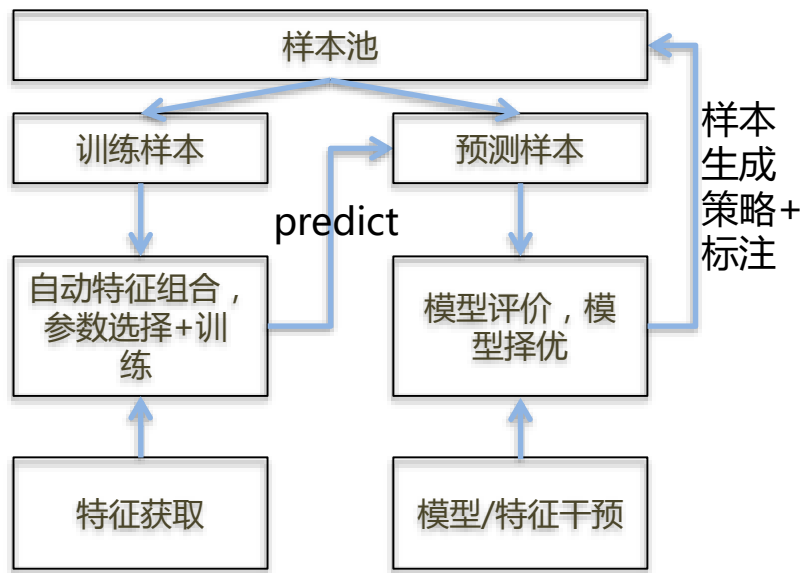
General Form:



Example Graph:



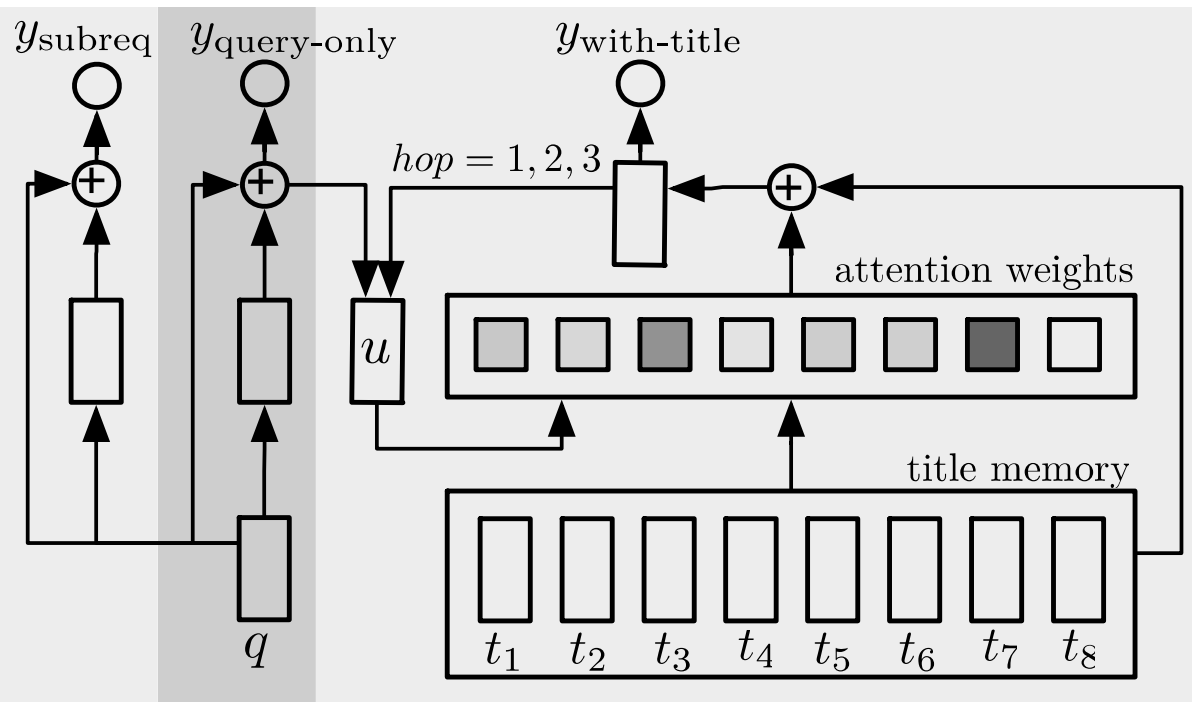
模型生成+强化学习 贴近最终应用的准召优化



Tips

- ◇ 样本挑选—
边界样本+随机样本
+关键特征相关样本
+半监督自动样本挖掘
- ◇ 多种特征组合方案。
- ◇ 多种模型/相关参数
- ◇ 低成本干预手段

用深度学习构建记忆网络 丰富语义信息，提升泛化能力



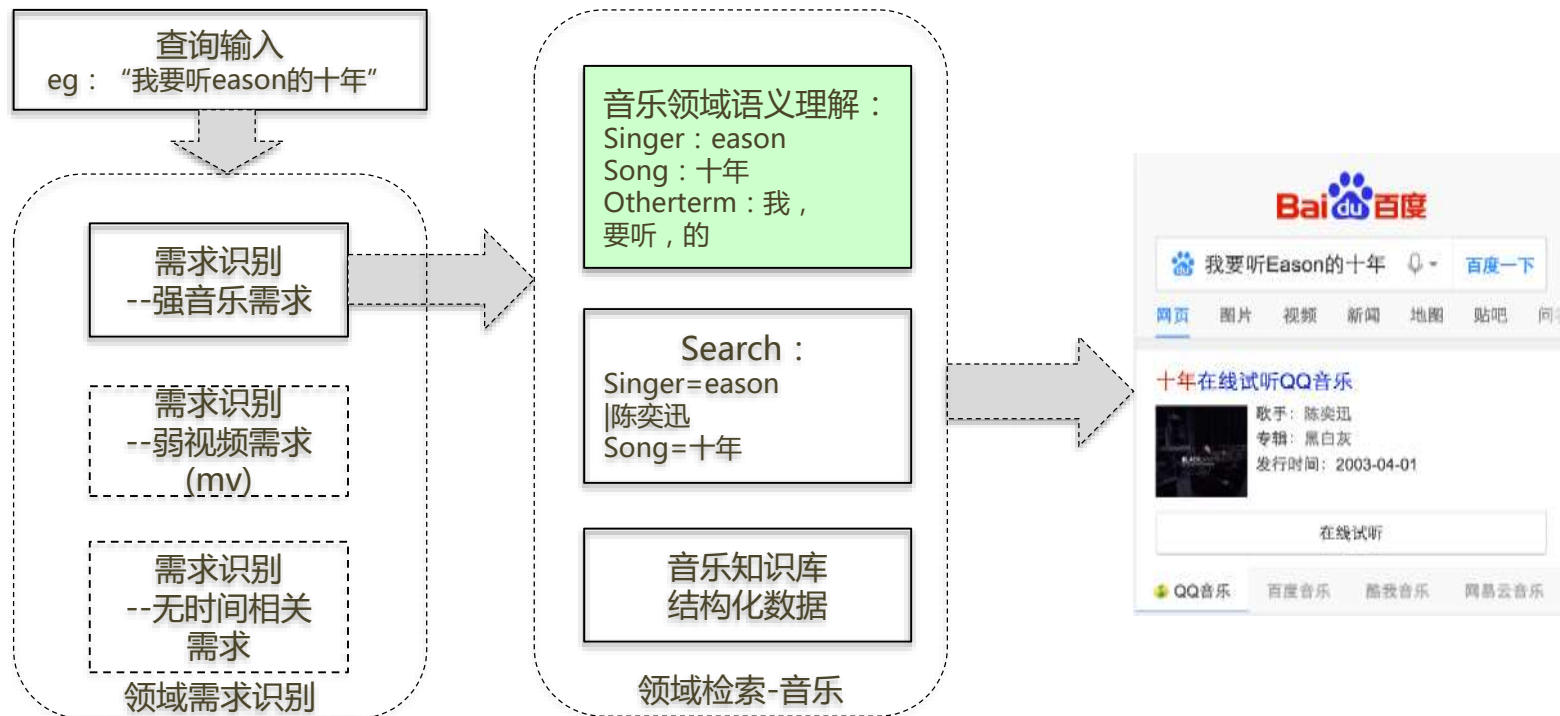
Tips

- ◇ End-To-End, 训练记忆网络, 进一步提升泛化能力
Eg: 局气 -> 餐厅
Eg: 桥西外国语小学 vs 桥东
- ◇ 复用已有的算法和特征, 产生大量高准召样本
- ◇ 深度学习的产出, 作为最终模型的特征输入, 重新参与最终模型生成
- ◇ 效果:
核心评测集 F1: 0.84 -> 0.90

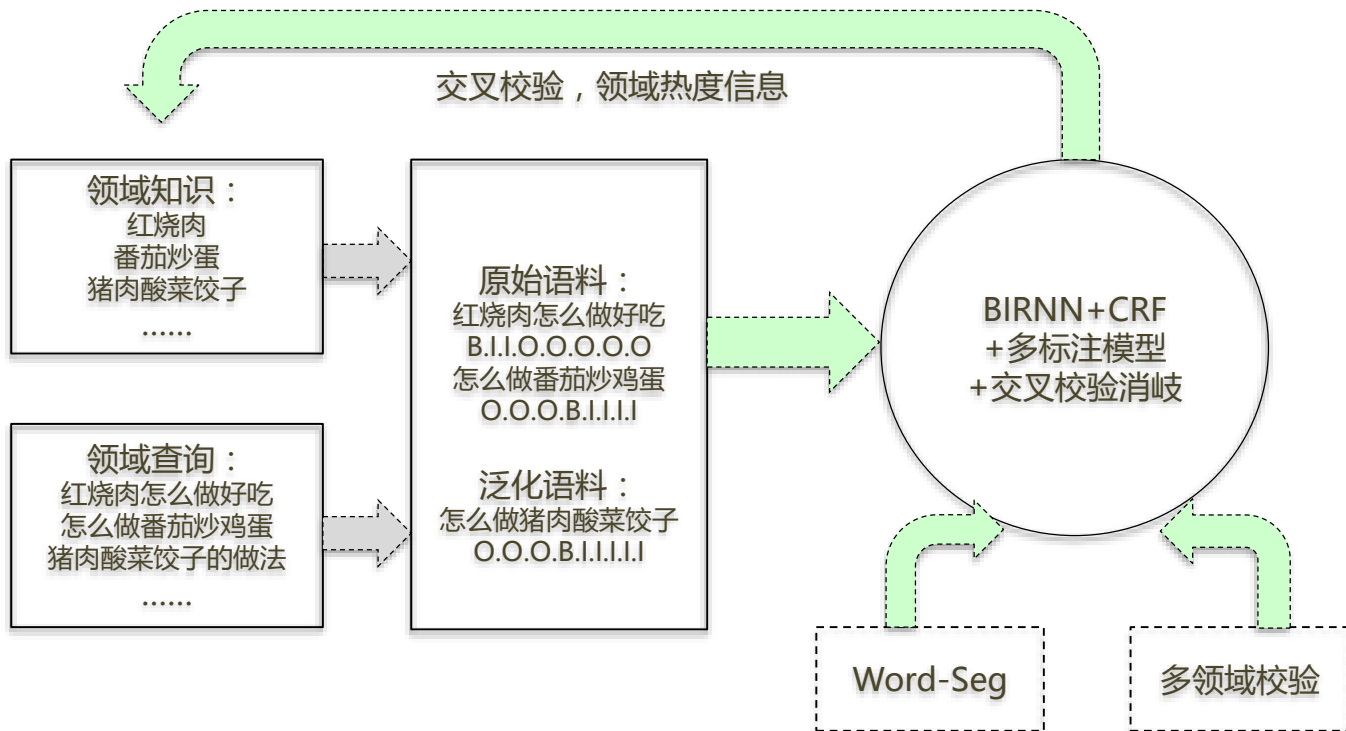
3

领域语义理解模型

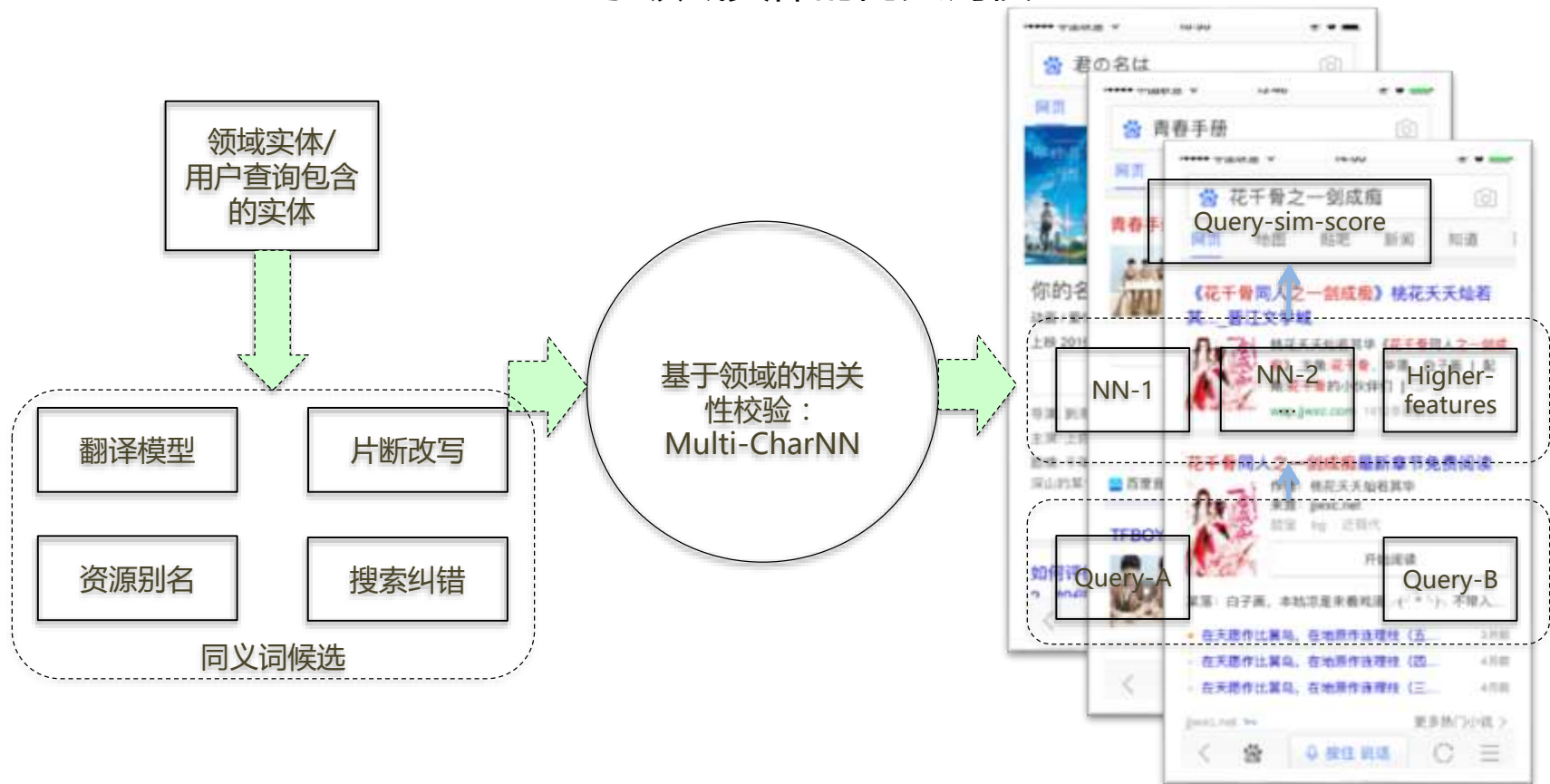
领域语义理解



基于深度序列标注模型的迭代



基于领域实体的同义词模型

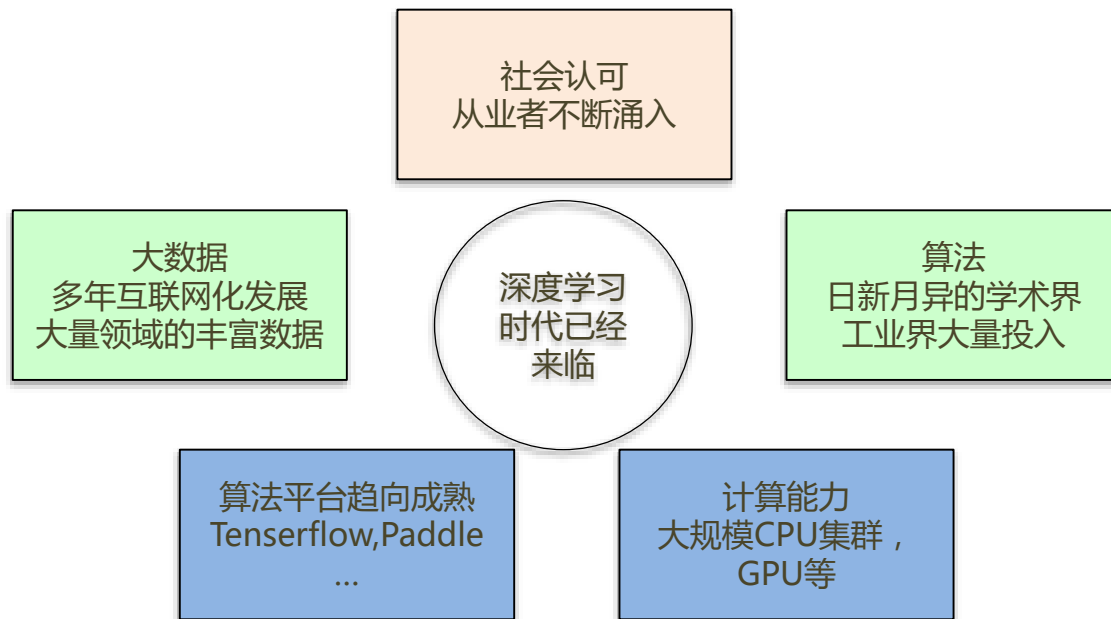


4

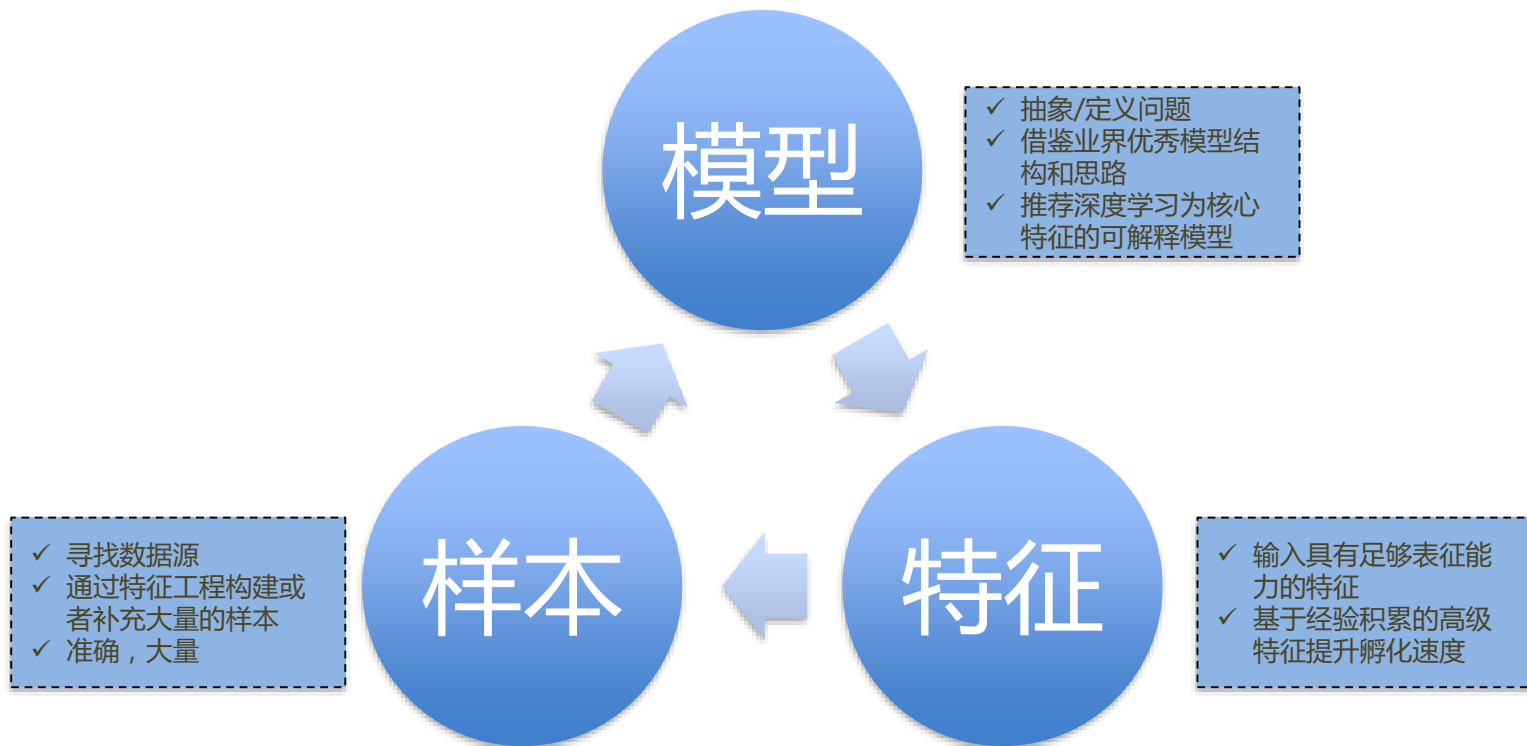
如何利用深度学习

做更多的策略，做更好的效果

深度学习现状



深度学习应用Tips





THANK YOU