《MIX: Multi-channel Information Crossing for Text Matching》分享

深度文本匹配在搜索场景中的应用

移动浏览产品部/自然语言处理组
陈浩蓝　2018.07

目录 / CONTENTS

主要履历

浙江大学	工学学士	计算机学院
阿尔伯塔大学	科学硕士	电子与计算机工程学院
Google	实习工程师	Summer of Code项目/后台开发
腾讯	高级研究员	搜索词解析/**语义搜索**/多媒体搜索/搜索行为预测/广告词挖掘

公司内学术成果

Haolan Chen, Di Niu, Kunfeng Lai, Yu Xu, Masoud Ardakani. Separating-plane factorization models: Scalable recommendation from one-class implicit feedback. ACM CIKM 2016

Haolan Chen, Fred Han, Di Niu, Dong Liu, Kunfeng Lai, Chenglin Wu, Yu Xu. MIX: Multi-Channel Information Crossing for Text Matching. ACM KDD 2018

公司内获奖

2018	腾讯卓越运营奖银奖（金奖空缺）
2017	优秀员工奖
2017	浏览器技术大拿奖
2017	部门经理即时激励奖
2017	五星团队奖
2016	腾讯技术分享奖第6名，6/276 每日头条文章

移动搜索市场需要更符合移动场景的搜索体验 —— 搜索直达

用户价值

操作：高效、便捷
结果：精准、智能
内容：优质、丰富

高效

精准

优质

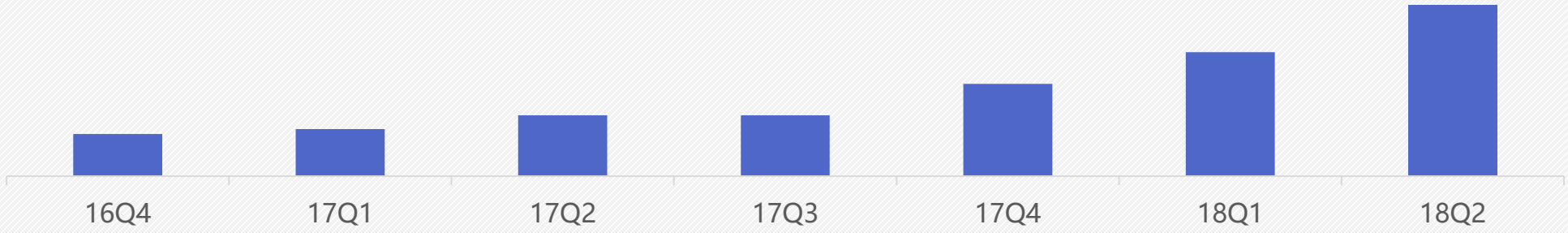
平台布局

亿级流量的自有价值转化
平台流量的再分发
更大的商业空间

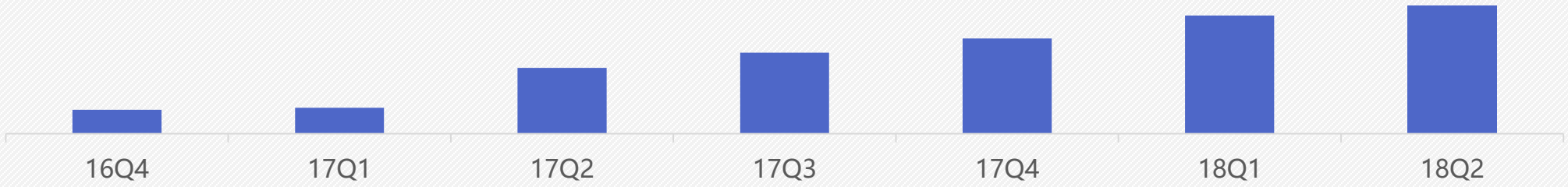
搜索直达

规模和收入均大幅提升

搜索直达规模，相比去年底提升97%



搜索直达收入，相比去年底提升35%



深度文本匹配在搜索直达的应用

一 二 三 四 五

Query请求 → 资源召回 → 资源排序 → 直达曝光 → 用户点击 → 商业计价

Query 补全	精准 召回	CTR 预估模型	UI 模板系统		广告主接入平台
Query 解析	核心 召回	CVR 预估模型	分层实验系统		关键词售卖 推荐系统
需求理解	**语义 召回**	**相关性模型**	模拟召回平台		

底层支撑服务

运营配置平台	质量评估平台	Goodcase 监控平台	资源服务 监控平台	数据统计平台	用户画像平台

研究成果

- 文本匹配演进
- 研究工作介绍
- 模型效果评测

文本匹配的价值与挑战

无监督学习

Bag of words, TFIDF, mean of word vectors, Levenshtein distance, tons of hand crafted rules.

浅层有监督学习

SVM based on handcrafted features.

传统方法的局限

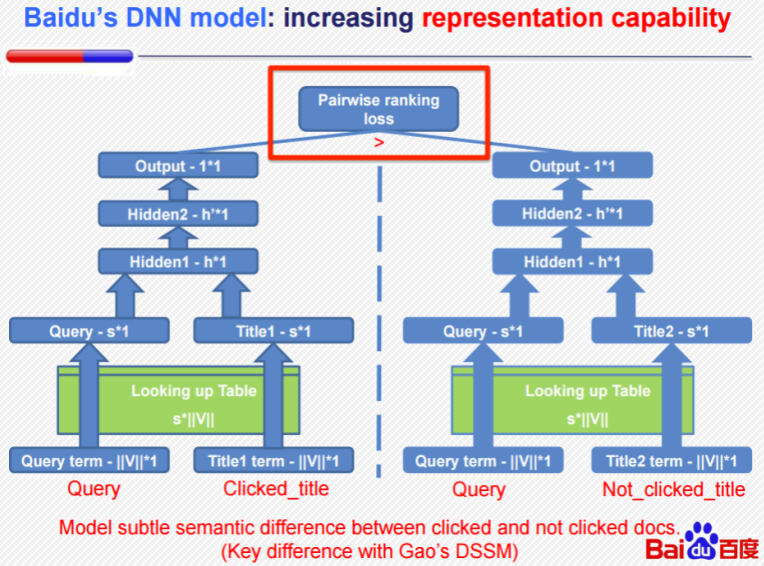
词义局限：菠萝-凤梨，苹果-苹果
结构局限：机器学习-学习机器
知识局限：秦始皇打Dota

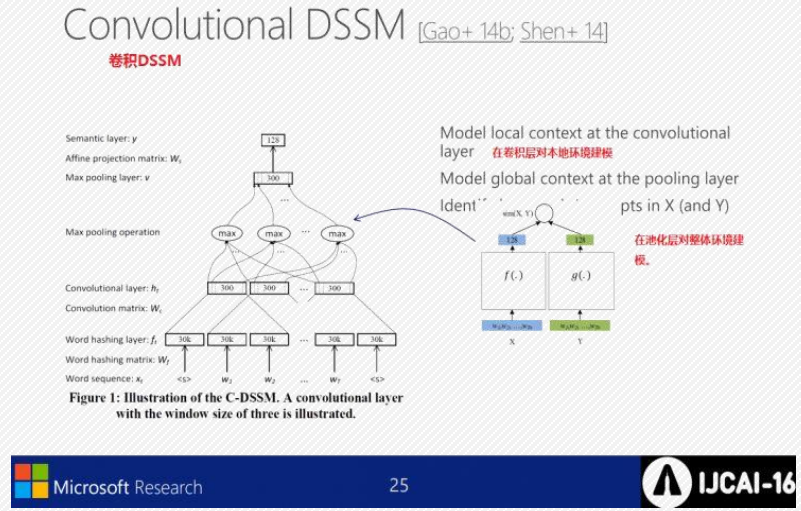
行业趋势：深度文本匹配

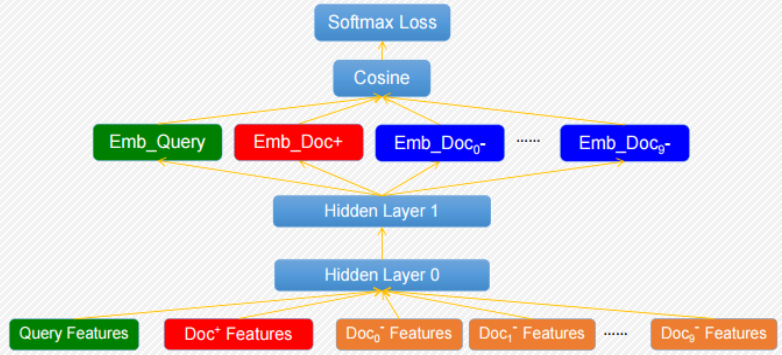
- **Microsoft**
  Deep Structured Semantic Model

- **百度**
  Baidu Search: Challenges we face, Experiences we learned.

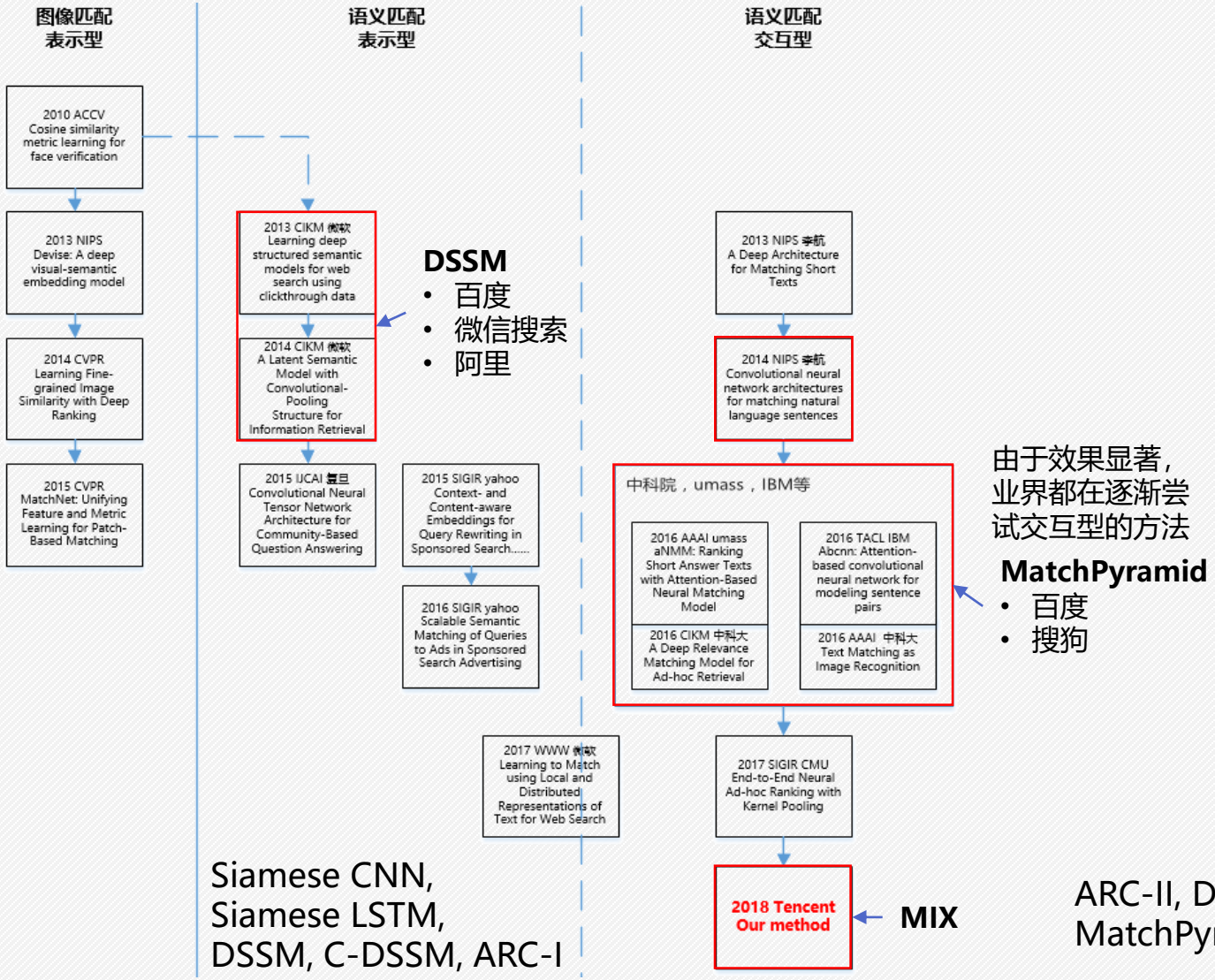
- **Yahoo**
  Ranking Relevance in Yahoo Search



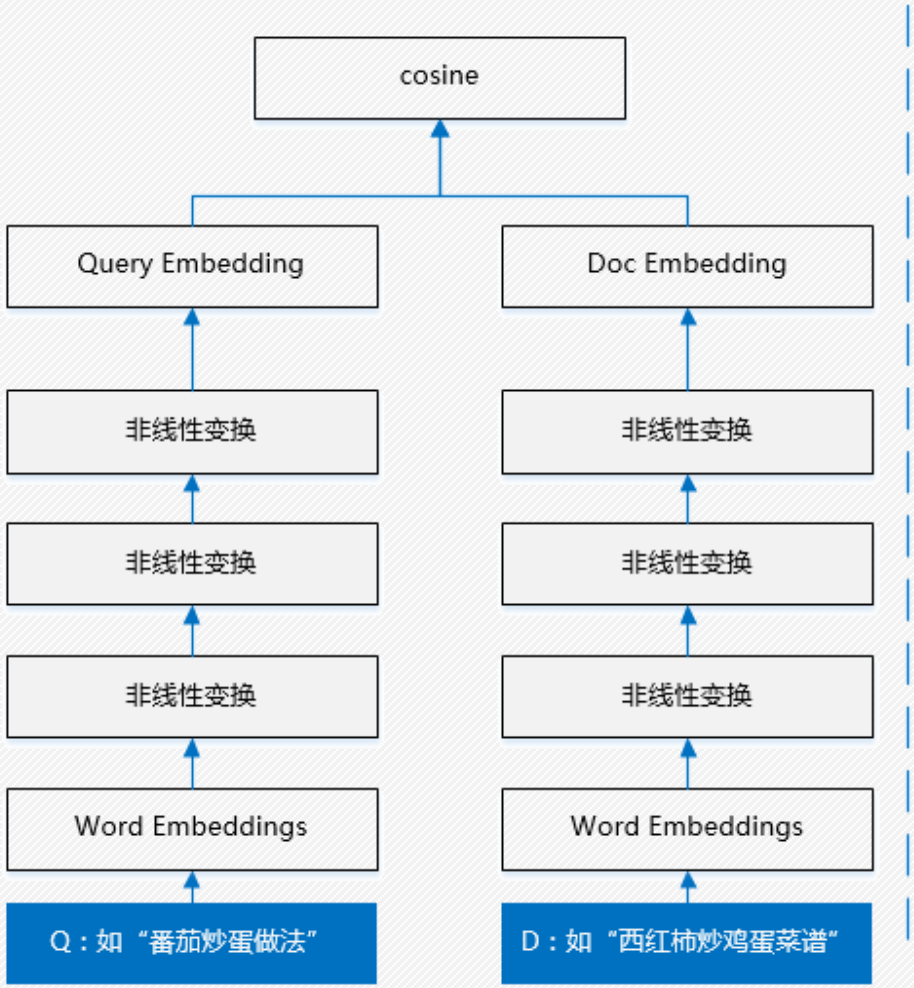




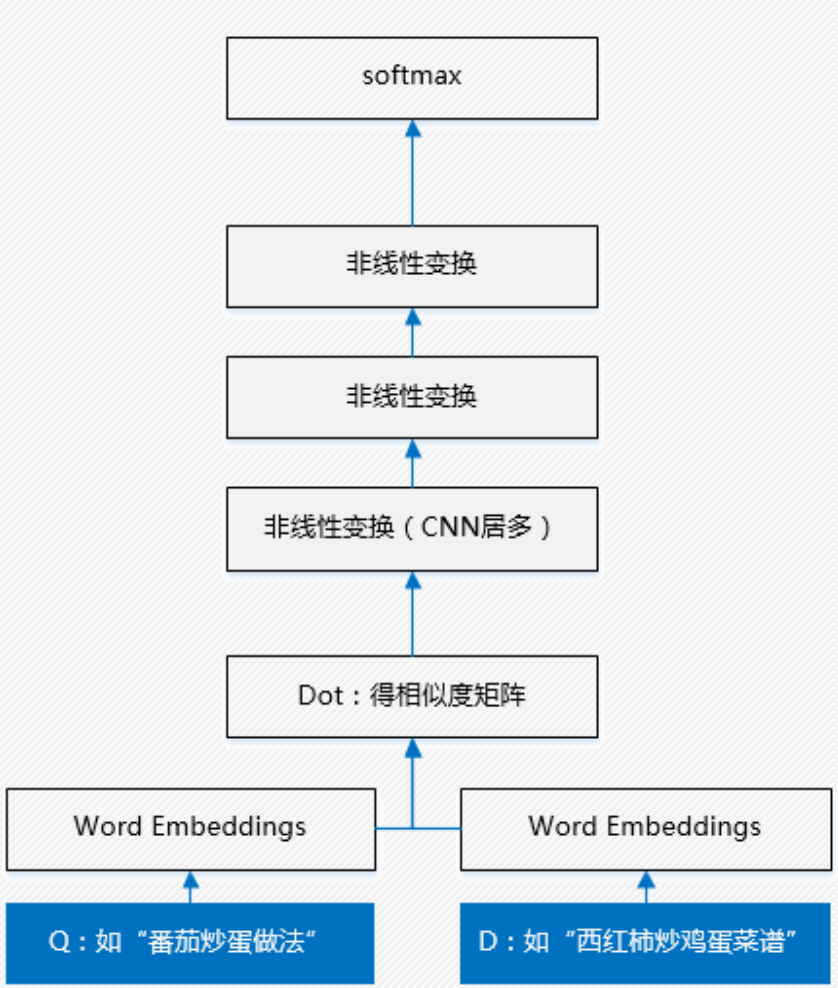
深度文本匹配 发展路线



深度文本匹配 表示型

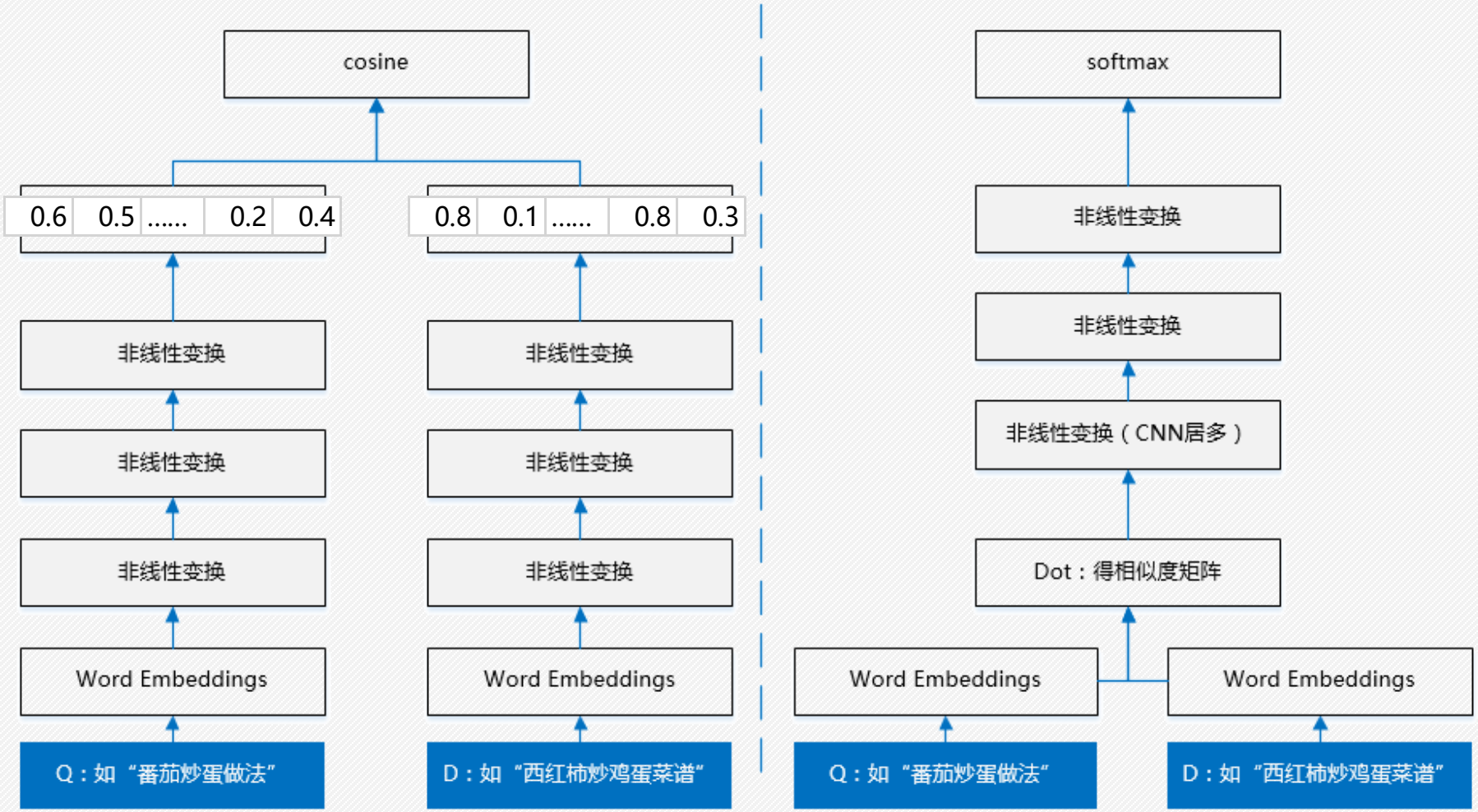


表示型（如DSSM）
Representation-based



交互型（如ARC-II）
Interaction-based

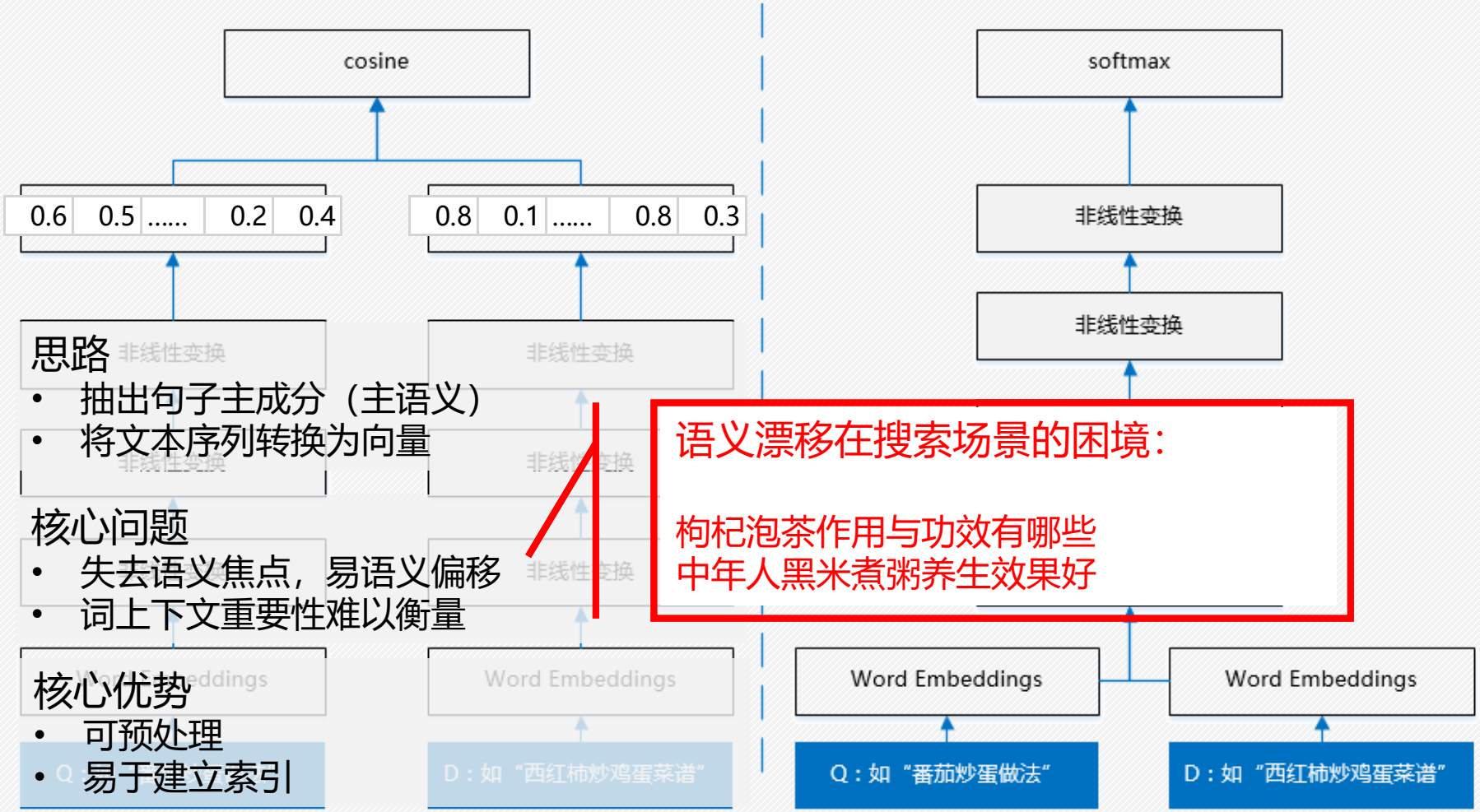
深度文本匹配 表示型



表示型（如DSSM）
Representation-based

交互型（如ARC-II）
Interaction-based

深度文本匹配 表示型



思路
- 抽出句子主成分（主语义）
- 将文本序列转换为向量

核心问题
- 失去语义焦点，易语义偏移
- 词上下文重要性难以衡量

核心优势
- 可预处理
- 易于建立索引

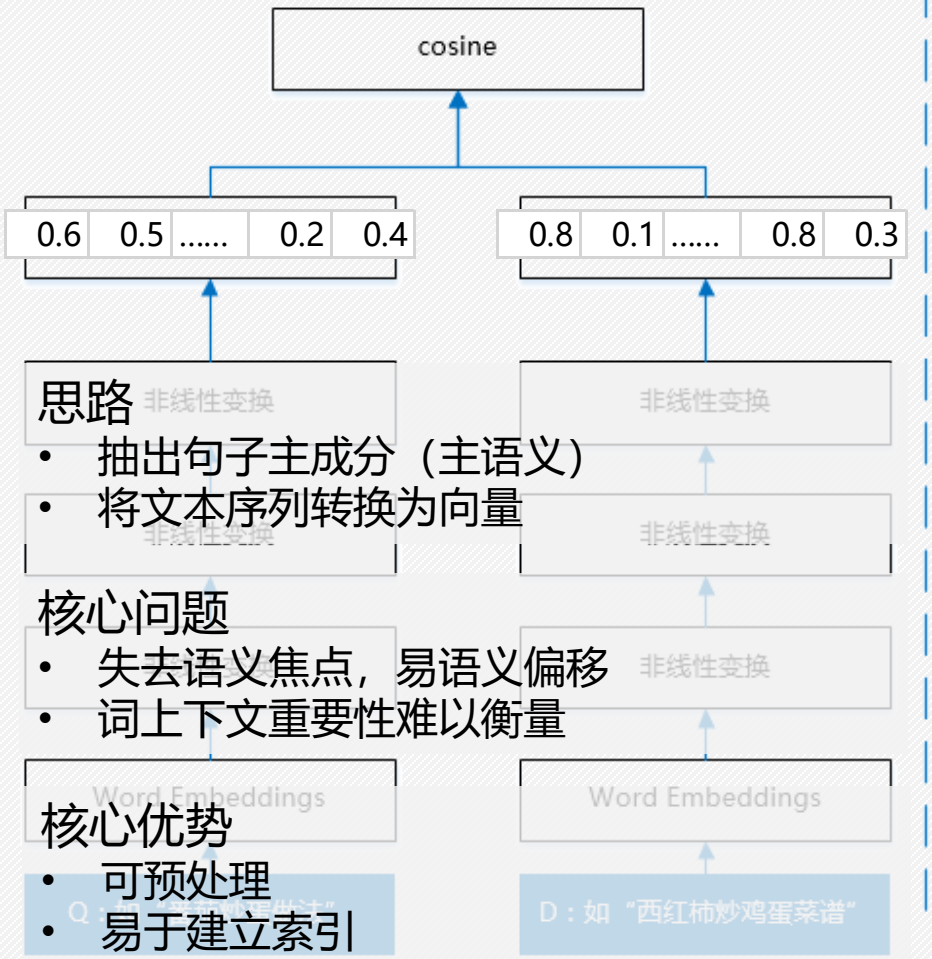
语义漂移在搜索场景的困境：

构杞泡茶作用与功效有哪些
中年人黑米煮粥养生效果好

表示型（如DSSM）
Representation-based

交互型（如ARC-II）
Interaction-based

深度文本匹配 表示型

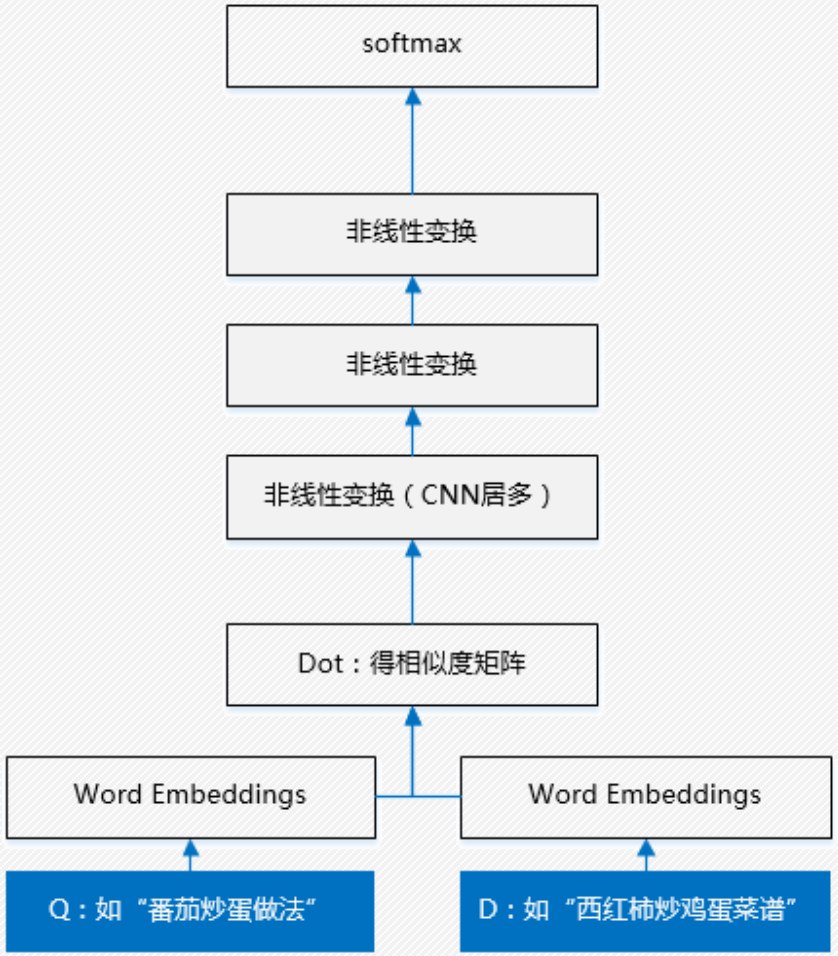


表示型（如DSSM）
Representation-based

思路
- 抽出句子主成分（主语义）
- 将文本序列转换为向量

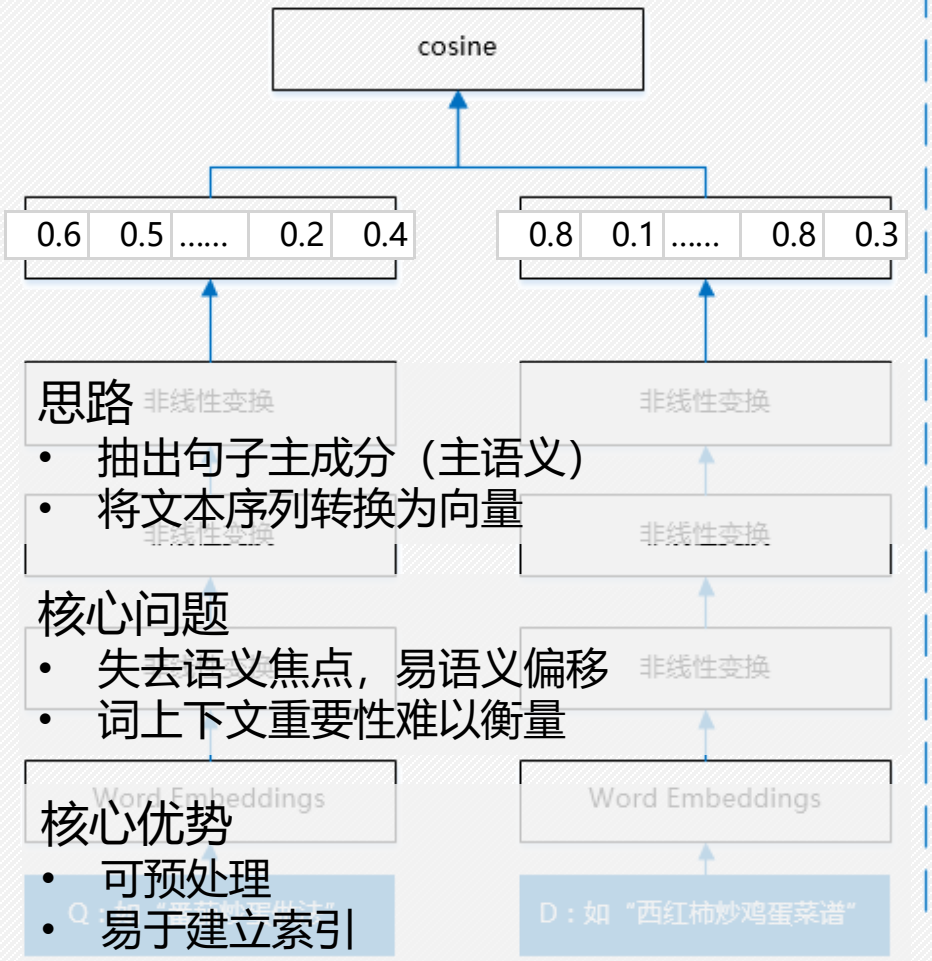
核心问题
- 失去语义焦点，易语义偏移
- 词上下文重要性难以衡量

核心优势
- 可预处理
- 易于建立索引



交互型（如ARC-II）
Interaction-based

深度文本匹配 交互型

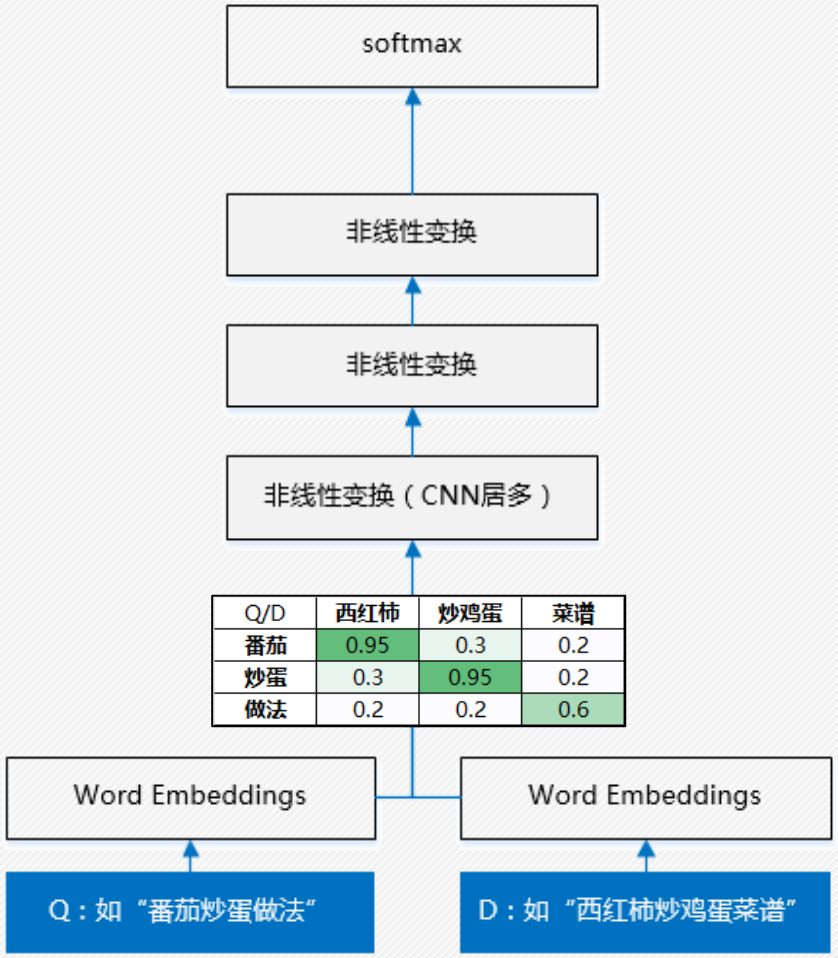


表示型（如DSSM）
Representation-based

思路
- 抽出句子主成分（主语义）
- 将文本序列转换为向量

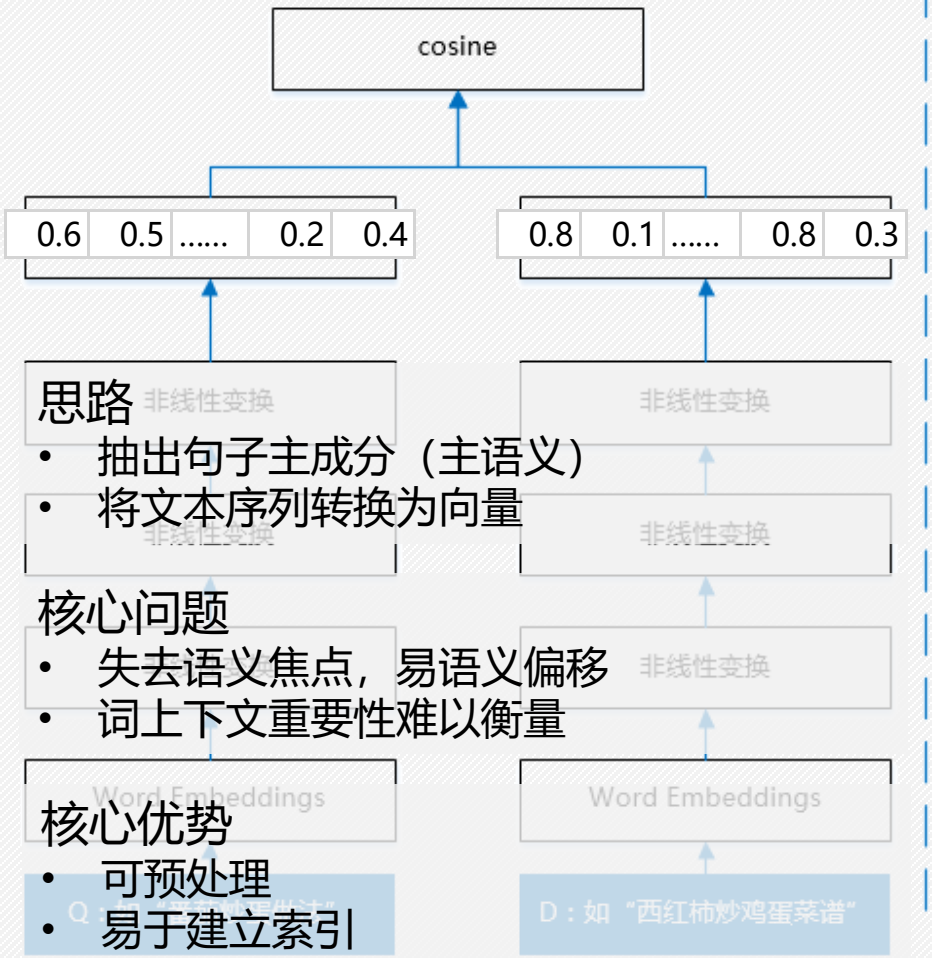
核心问题
- 失去语义焦点，易语义偏移
- 词上下文重要性难以衡量

核心优势
- 可预处理
- 易于建立索引



交互型（如ARC-II）
Interaction-based

深度文本匹配 交互型

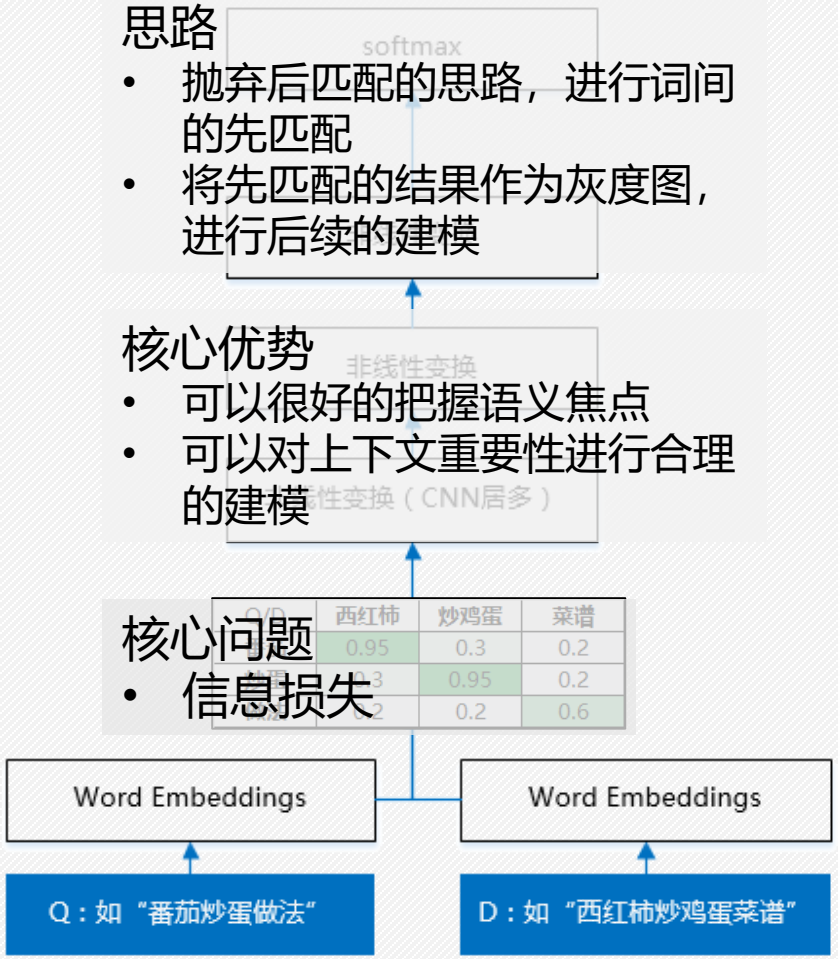


思路
- 抽出句子主成分（主语义）
- 将文本序列转换为向量

核心问题
- 失去语义焦点，易语义偏移
- 词上下文重要性难以衡量

核心优势
- 可预处理
- 易于建立索引

表示型（如DSSM）
Representation-based



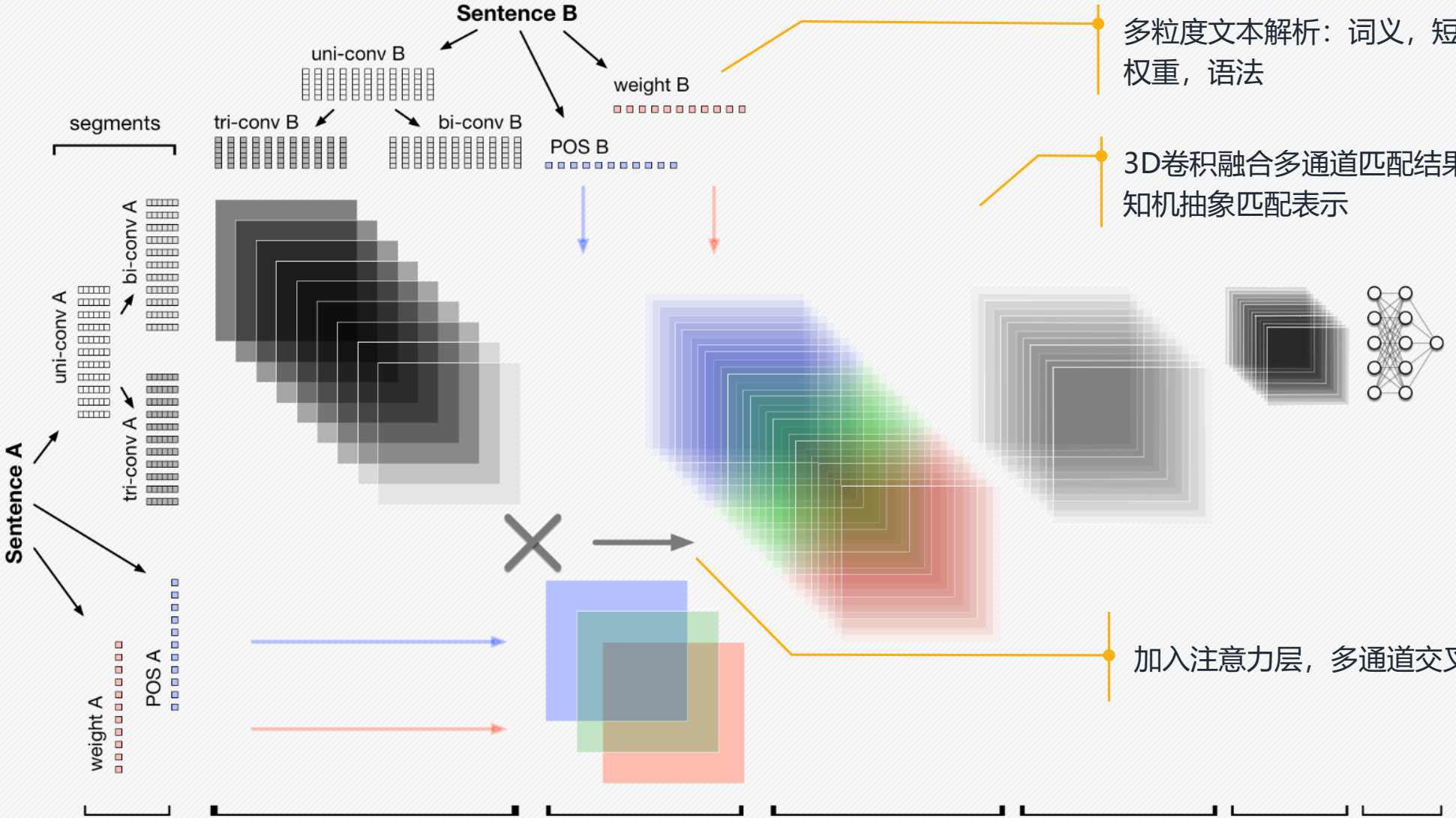
思路
- 抛弃后匹配的思路，进行词间的先匹配
- 将先匹配的结果作为灰度图，进行后续的建模

核心优势
- 可以很好的把握语义焦点
- 可以对上下文重要性进行合理的建模

核心问题
- 信息损失

交互型（如ARC-II）
Interaction-based

Multi-channel Information Crossing(MIX) Model

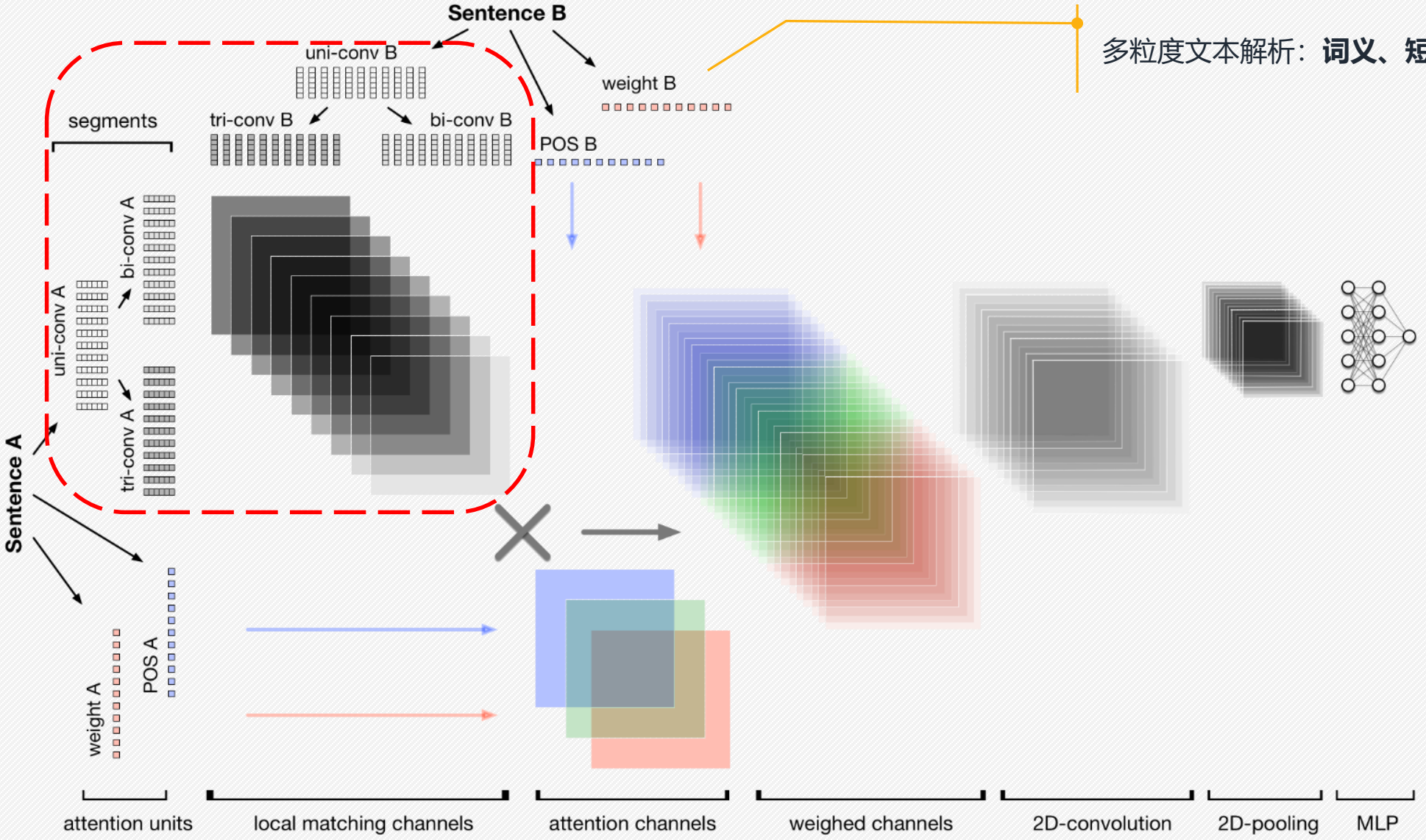


多粒度文本解析：词义，短语，权重，语法

3D卷积融合多通道匹配结果，多层感知机抽象匹配表示

加入注意力层，多通道交叉匹配

1. 多粒度匹配



多粒度文本解析：**词义、短语**

1. 多粒度匹配 Case study

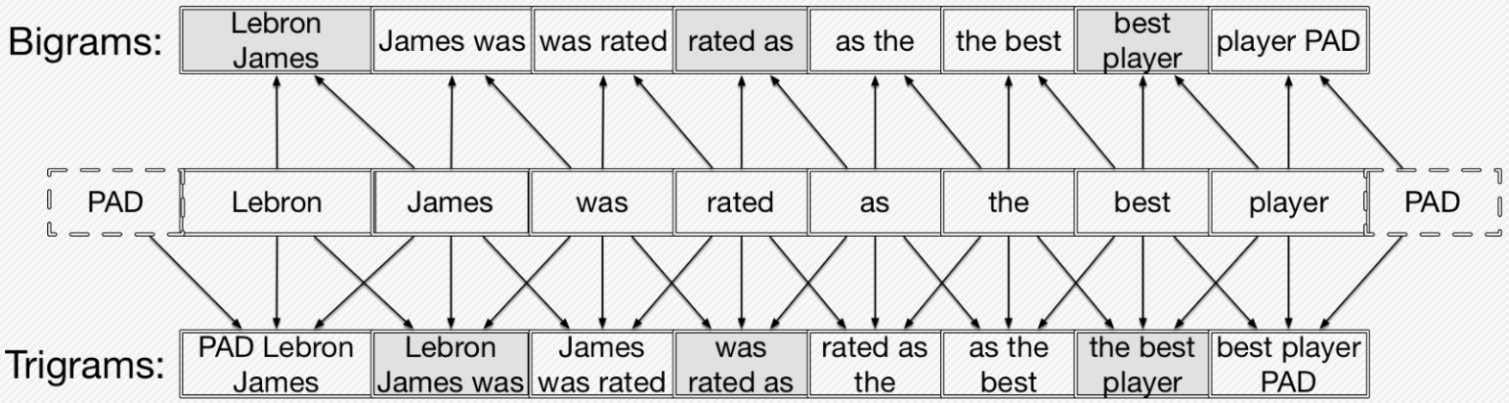
单粒度的问题 & 深度学习的局限

	place	of	interest
senic			
spot	0.2		

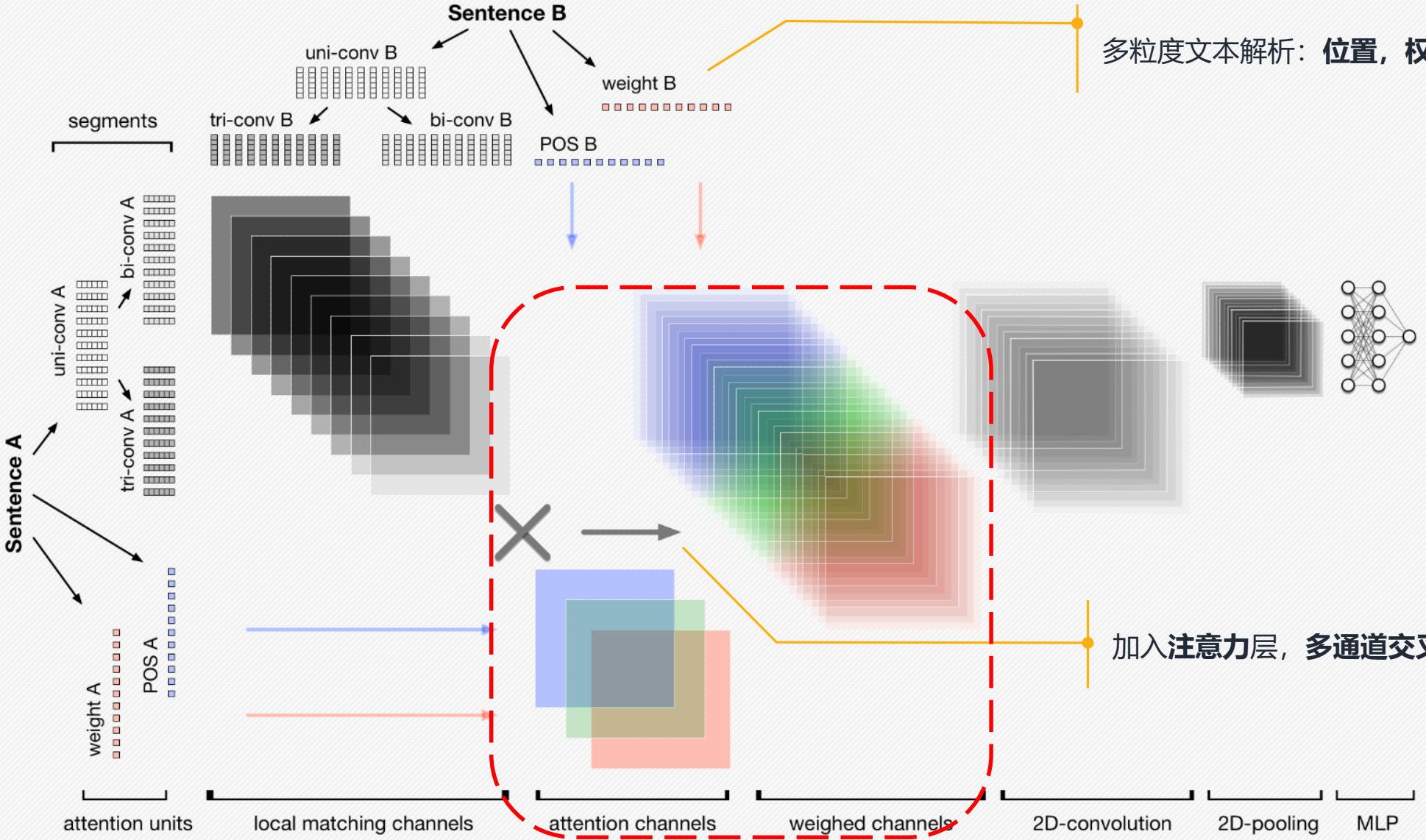
	in	all
all		1.0
in	1.0	

	hard	work
work		1.0
hard	1.0	

多粒度解析对短语的有效捕捉



2. 注意力机制 & 多通道交叉



多粒度文本解析：**位置，权重，语法**

加入**注意力**层，**多通道交叉**匹配

2. 注意力机制 & 多通道交叉 Case study

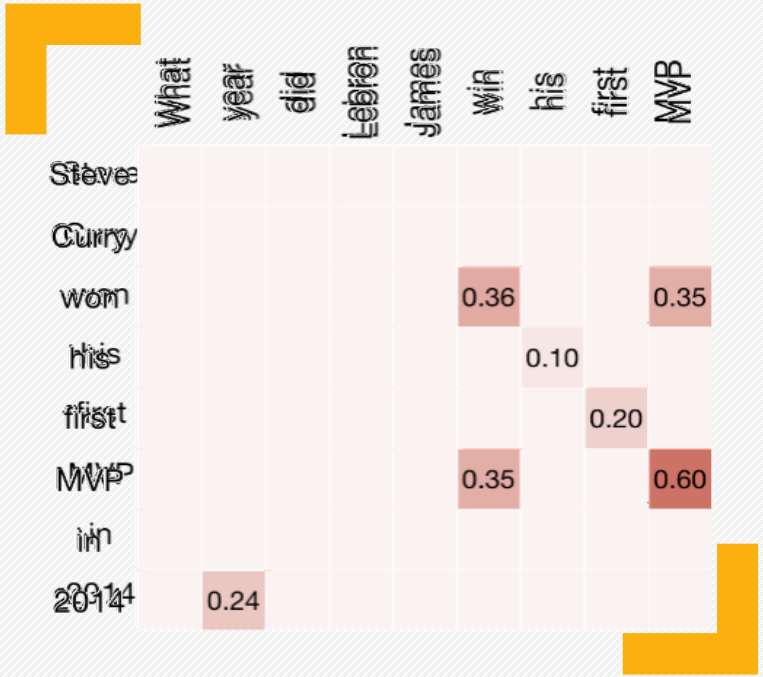
匹配词权重对匹配的影响 词匹配

Query

What year did Lebron James win his first MVP ?

Doc 1

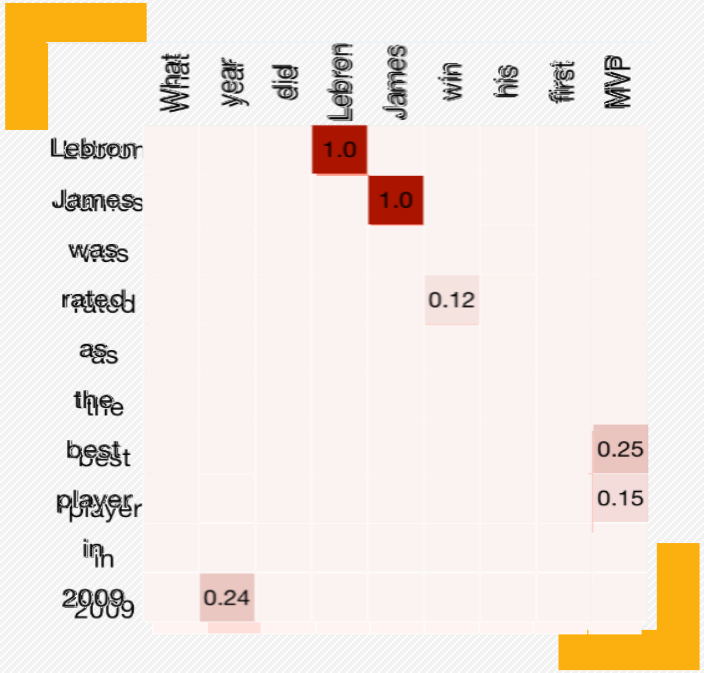
Steve Curry won his first MVP in 2014.



?

Doc 2

Lebron James was rated as the best player in 2009



2. 注意力机制 & 多通道交叉 Case study

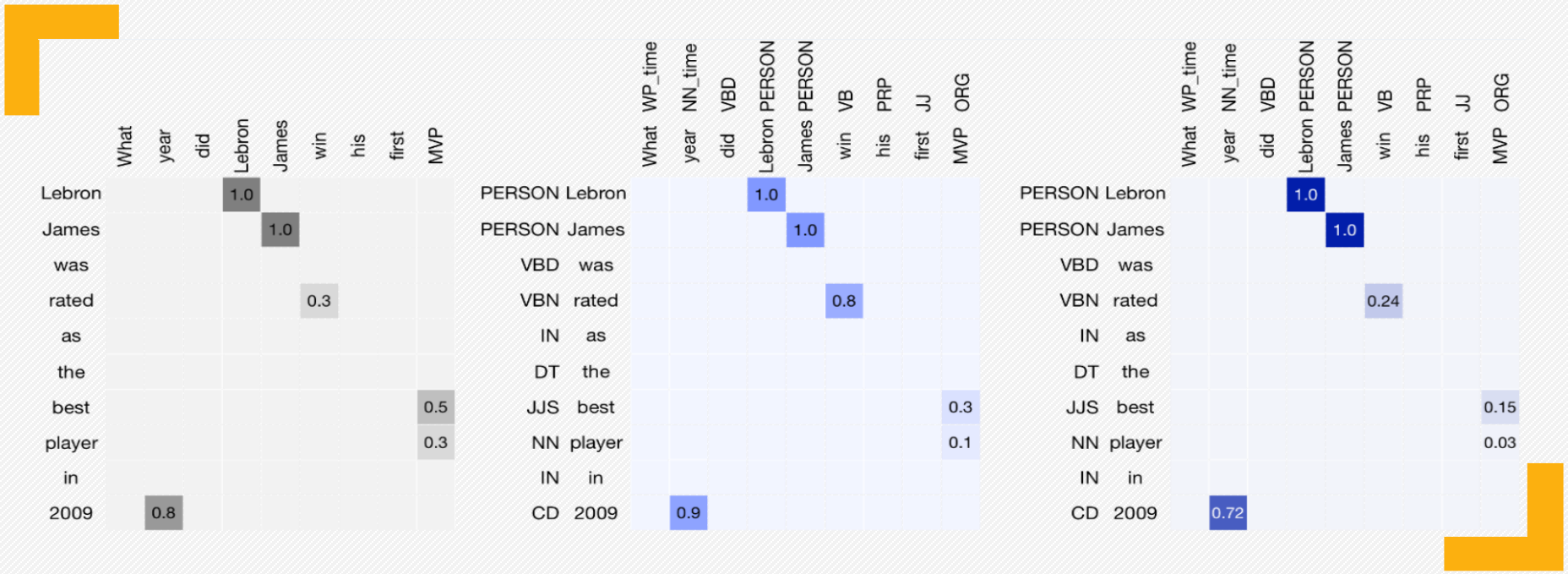
语法匹配信息对匹配的影响

Query

What year did Lebron James win his first MVP ?

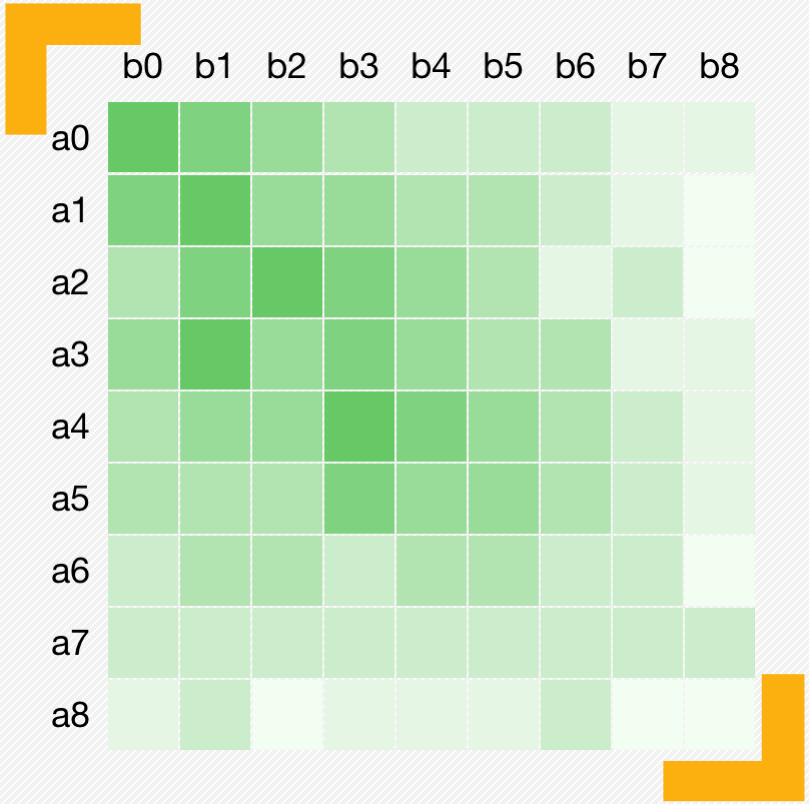
Doc

Lebron James was rated as the best player in 2009

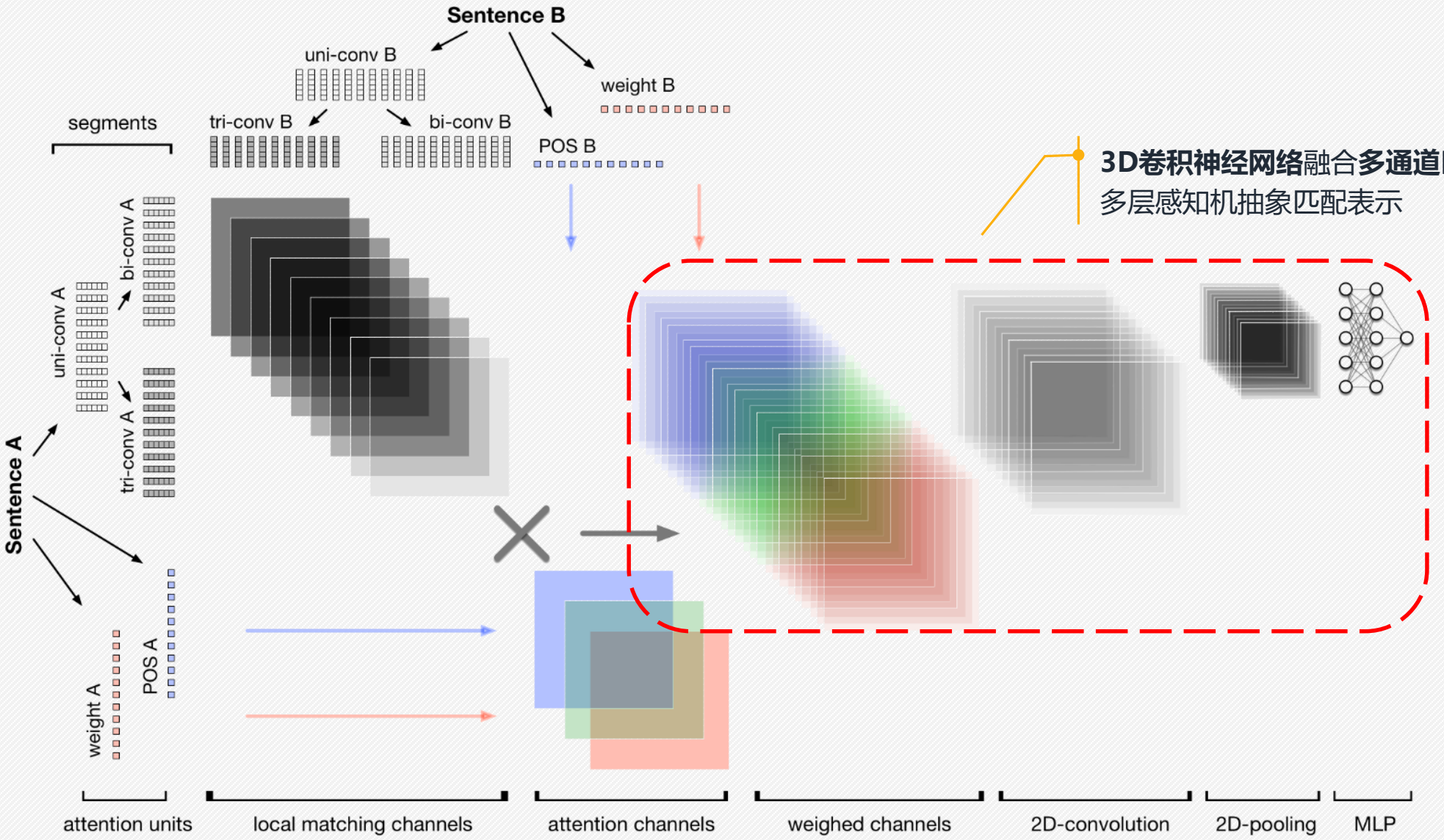


2. 注意力机制 & 多通道交叉 Case study

位置信息对匹配的影响



3. 交叉通道融合 张量分类问题

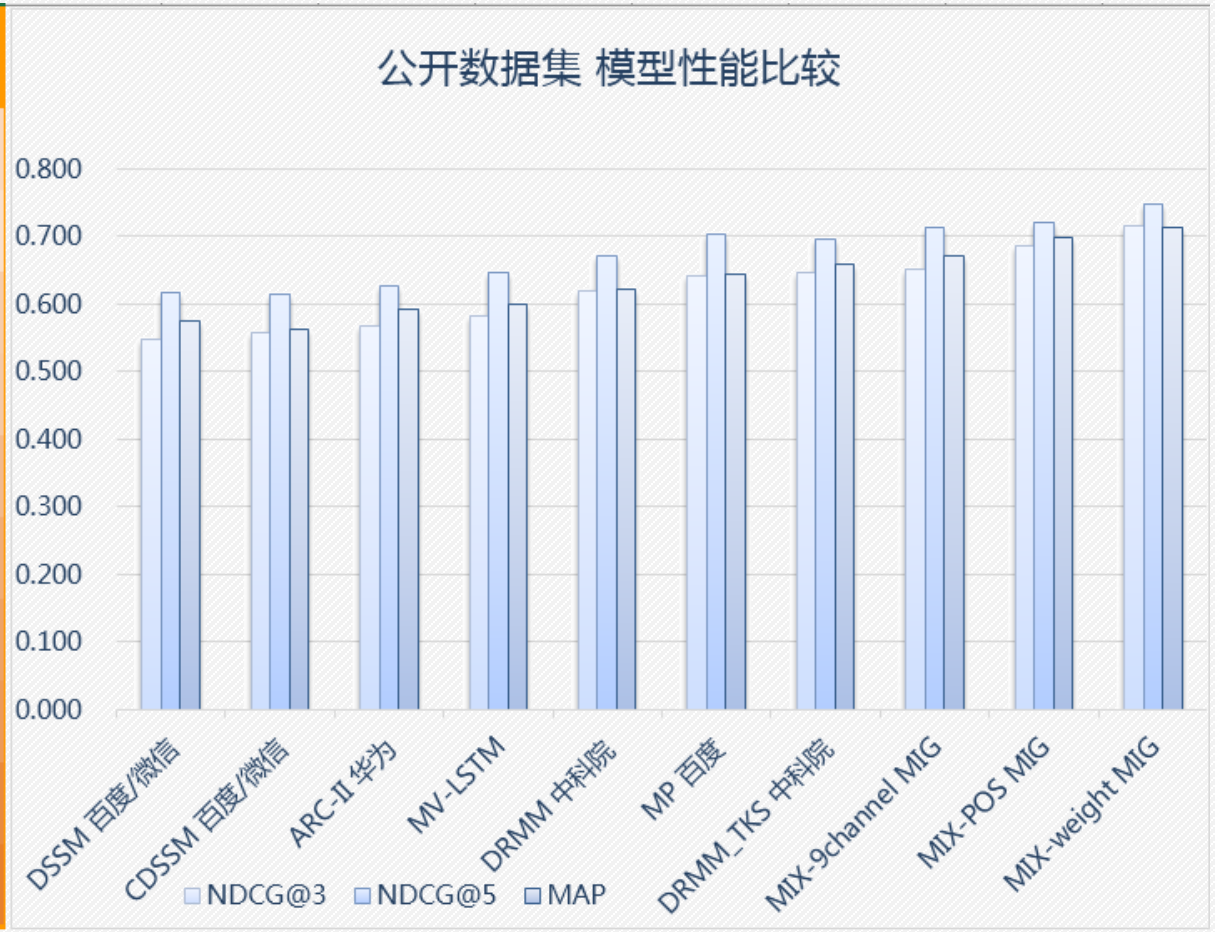


3D卷积神经网络融合**多通道**匹配结果，
多层感知机抽象匹配表示

效果评测

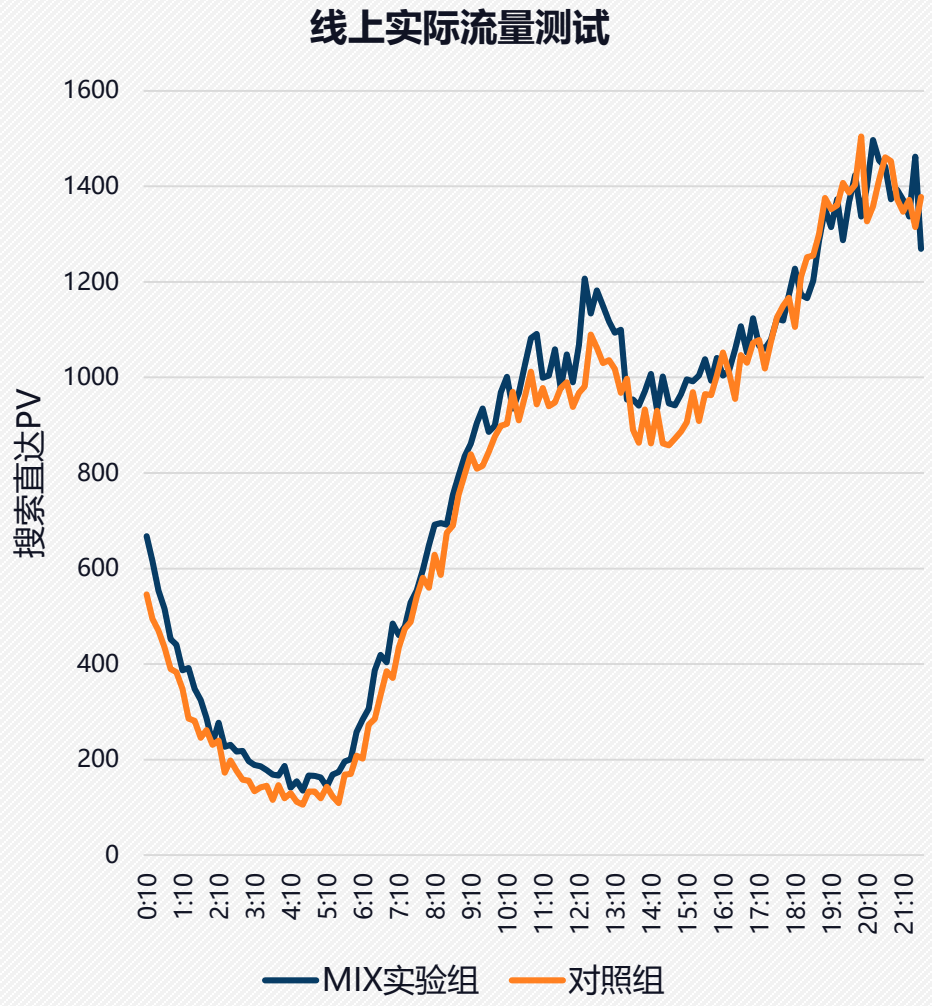
11.1%

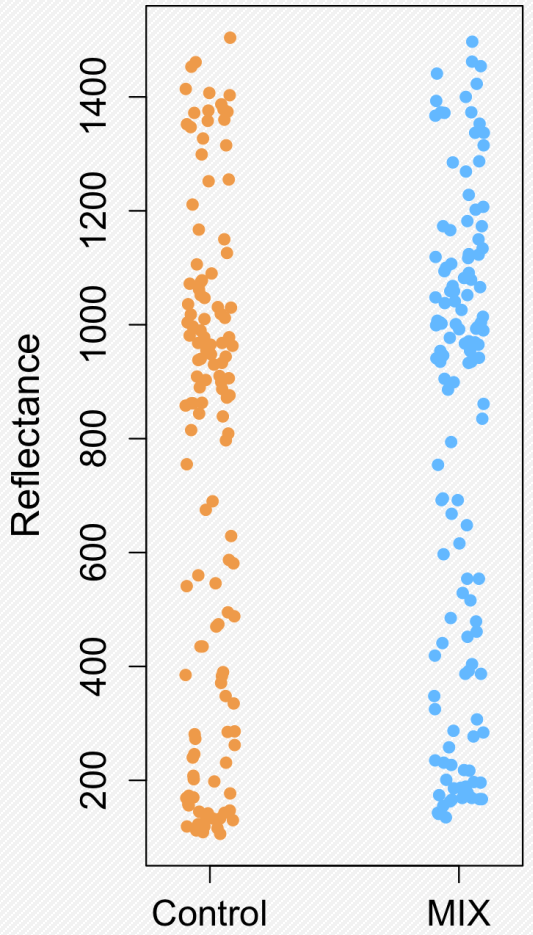
模型	NDCG@3	NDCG@5	MAP
DSSM 百度/微信	0.547	0.617	0.575
CDSSM 百度/微信	0.559	0.615	0.562
ARC-II 华为	0.568	0.626	0.592
MV-LSTM	0.582	0.645	0.599
DRMM 中科院	0.619	0.670	0.622
MP 百度	0.642	0.704	0.643
DRMM_TKS 中科院	0.646	0.696	0.659
MIX-9channel MIG	0.651	0.714	0.672
MIX-POS MIG	0.686	0.721	0.697
MIX-weight MIG	**0.715**	**0.748**	**0.713**

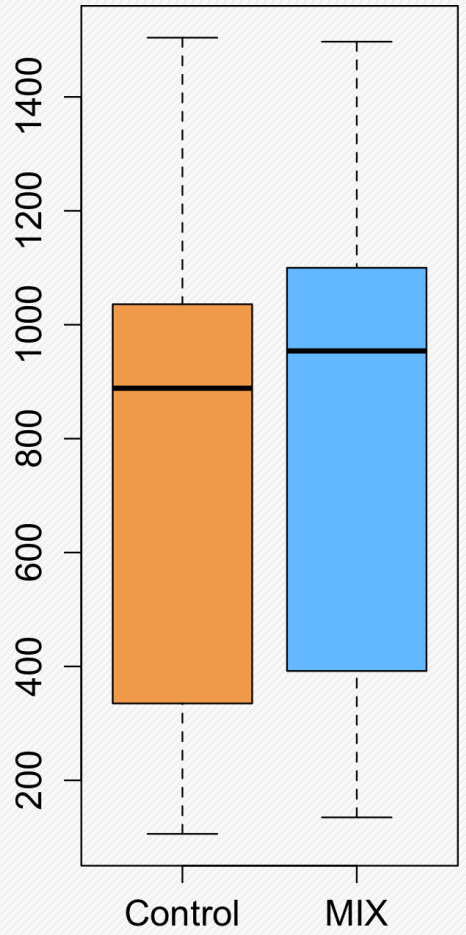


性能评测

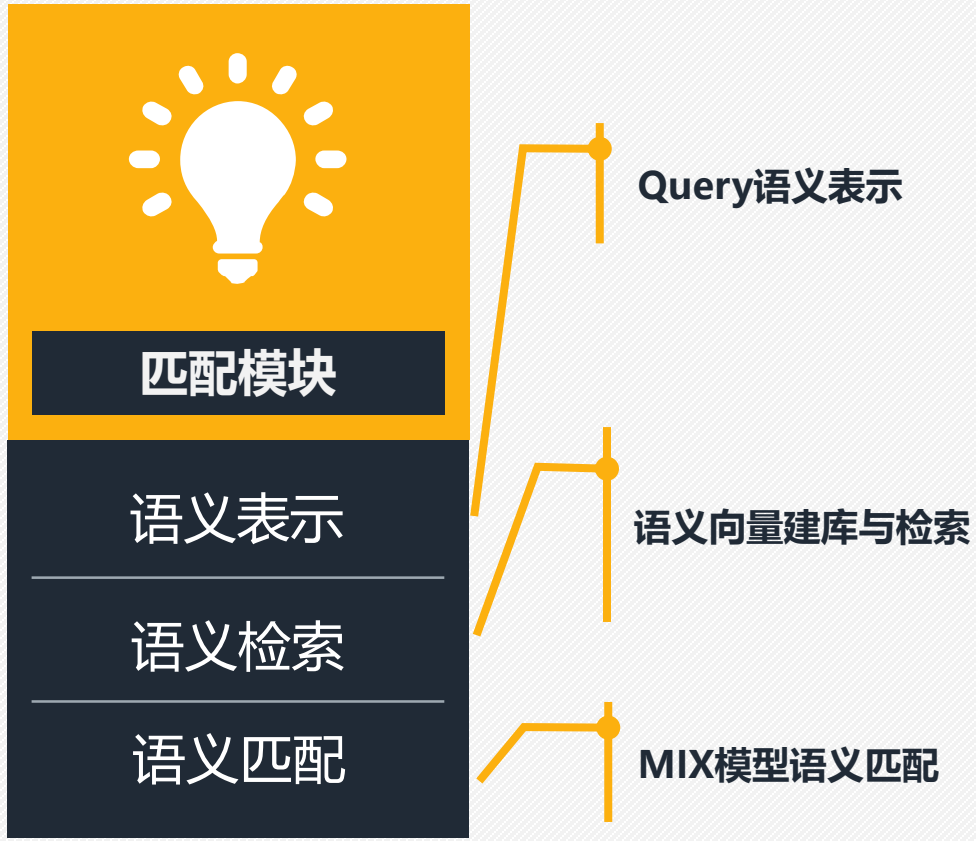
5.7%







语义搜索业务落地



未来工作 - 关于搜索直达的研究

有趣的问题

- 语义搜索
- 多媒体搜索
- 需求全面满足
- 需求预先满足

End2end语义召回
语义辅助排序

视频细分意图识别
视频理解
语义匹配

主动需求持续满足
主动需求多维度满足

主动需求个性化预测
主动需求场景化预测

诚邀合作

移动QQ浏览器的进化

感谢聆听

移动浏览产品部 大数据开发组 自然语言处理组

分享人：陈浩蓝