机器学习

深 度 强 化 学 习

深度学习在检索中的应用

宋奇　　3月2日

扫钉钉群，加入我们

讲师介绍

宋奇

地图事业部　检索和推荐技术负责人

从事上下点的检索和推荐算法研发工作

在出行场景下率先实现了检索推荐系统，从人工规则模型，到传统机器学习和深度模型的技术升级，带给用户极致的发单体验。

课程目标 / 学习受益

- 了解出行场景的检索业务
- 理解检索中的关键技术问题
- 掌握Sent2Vec和Seq2Seq模型原理
- 能够在类似的业务有所借鉴和迁移

目录
Contents

01

第一章

检索业务介绍

产品形态

帮助用户**快速、准确**找到想要的起终点

起点

推荐上车点	73%
Rgeo挪点	15%
Rec推荐	7%
Sug检索	5%

终点

猜你想去	16%
家和公司	2%
Rec推荐	52%
Sug检索	30%

看几个CASE


●●●●● 中国联通 下午7:10 51%
北京 首都机场t2 取消
首都国际机场-T2航站楼
朝阳区首都机场路
国内出发
国内到达
国际港澳台出发
国际港澳台到达
首都机场T2航站楼-公交站
北京首都国际机场二号航站楼2号停车楼
朝阳区机场南路北京首都国际机场北京首都国际机场二号航...
首都机场t2航站楼停车场
朝阳区首都国际机场二号航站楼停车楼1楼
首都国际机场T2航站楼2号停车楼入口
朝阳区机场西路
国航头等舱休息室
顺义区首都国际机场中心广场东侧北京首都国际机场T3-E航...
首都机场T2航站楼国航头等舱休息室
朝阳区T2航站楼2号航站楼2层8号门对面


●●●●● 中国联通 下午1:19 11%
北京 很久 取消
很久以前烤串坊(清河店) 5.5km
海淀区西三旗清景园5号楼1层
很久以前烤串坊(天通苑店) 11.1km
昌平区天通苑五区19号楼底商铺
很久以前烤串坊(莲花桥店) 16.8km
丰台区吴家场路莲花池西里24号
很久以前只是家串店(朝阳店) 22.9km
朝阳区八里庄东里1号莱锦创意产业园A5号
很久以前烤串坊(簋街店) 16km
东城区东直门内大街209
很久以前烤串坊(望京店) 16.2km
朝阳区望京广顺北大街河萌中路
很久以前烤串坊(台湾街店) 16.1km
石景山区鲁谷路台湾街C1区远洋山水北区4号楼对面
很久以前只是家串店(双井店) 22.5km
朝阳区东三环南路甲5号
很久以前I只是家串店(方庄店) 23.3km
丰台区蒲方路甲12号
很久以前只是家串店(石园店) 32.6km
顺义区石园南区花园路南侧新华书店1层底商


●●●●● 中国联通 下午6:02 45%
北京 未名 取消
未名国际青年旅舍
海淀区成府路150号蓝润大厦
未名视通研发楼
北京市海淀区
未名楼酒店
海淀区颐和园路1-1号北京大学
未名宾馆
西城区西直门北礼士路甲54号
未名四合院酒店
东城区宝钞胡同甲38号
未名视界摄影主题酒店
西城区鼓楼西大街76号
金龙源宾馆
西城区鼓楼西大街76
未名连锁
北京市西城区
未名天日语学校(北大校区)
海淀区海淀路50号北京大学资源东楼1楼1102室
未名公寓
海淀区学院路32

检索怎么做

问题的本质是什么？

Query->POI->（Longitude，Latitude）

把人脑中的语义描述转化为机器能理解的经纬度位置信息

哪些关键问题？

用户想要什么：Query分析

理解POI数据 ：名称地址的标注分析

结果是否相关：Query-POI语义匹配

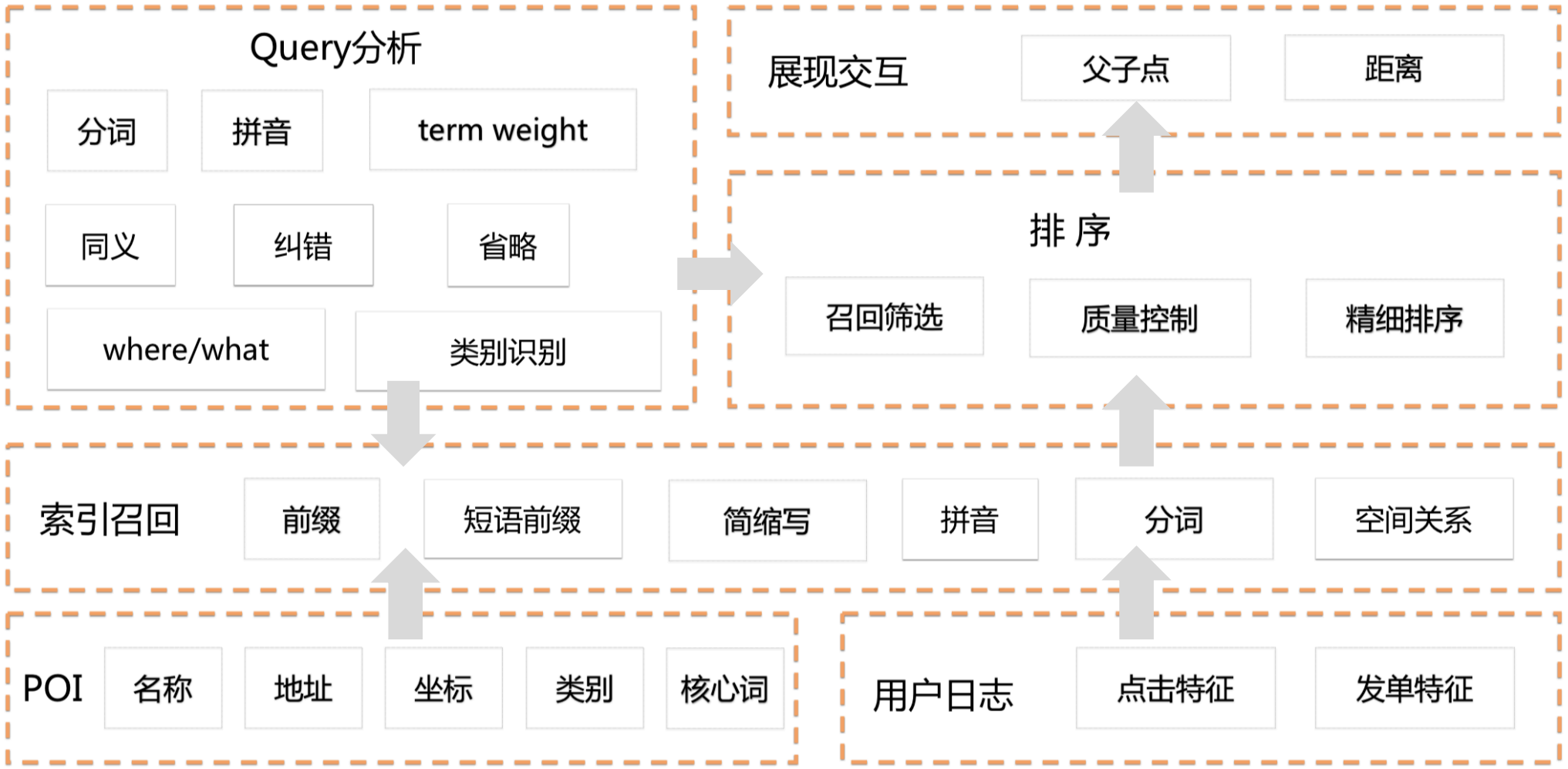
排序是否合理：基于用户行为的复杂排序模型（Learning to Rank）

如何评估？

客观指标：前三点击率，字长，无结果率

人工指标：NDCG，满意度评估

检索系统


Query分析
分词
拼音
term weight
同义
纠错
省略
where/what
类别识别
展现交互
父子点
距离
排 序
召回筛选
质量控制
精细排序
索引召回
前缀
短语前缀
简缩写
拼音
分词
空间关系
POI
名称
地址
坐标
类别
核心词
用户日志
点击特征
发单特征

02 第二章

Sent2Vec：短文本匹配的应用

匹配模型比较

发展阶段	模型	特点
概率统计	**BM25**	tf-idf特征，难以解决字面不匹配的问题
主题模型	**PLSA/LDA**	隐语义分析，只利用了文档级别的特征
翻译模型	**WMT/PMT**	统计翻译概率，利用query-click-title训练
深度模型	**Sent2Vec**	提取深度语义特征，利用query-click-title训练

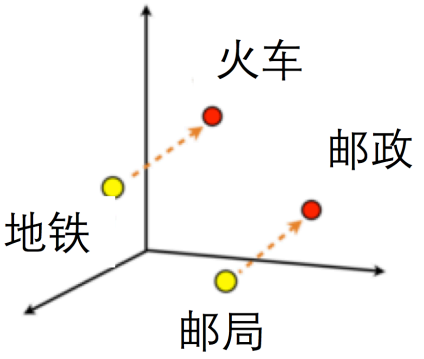
从点击日志中学习语义向量

Click	Query	POI Title
1	一考	一烤成名
0	一考	玉马科目一考场
0	一考	艺考路上
0	一考	北京卓桥艺考画室
0	一考	京都府驾校科目一考场
0	一考	艺考画室
0	一考	北京声乐培训
0	一考	科目一考试场
0	一考	艺考文化交流中心有限公司
0	一考	艺考联盟

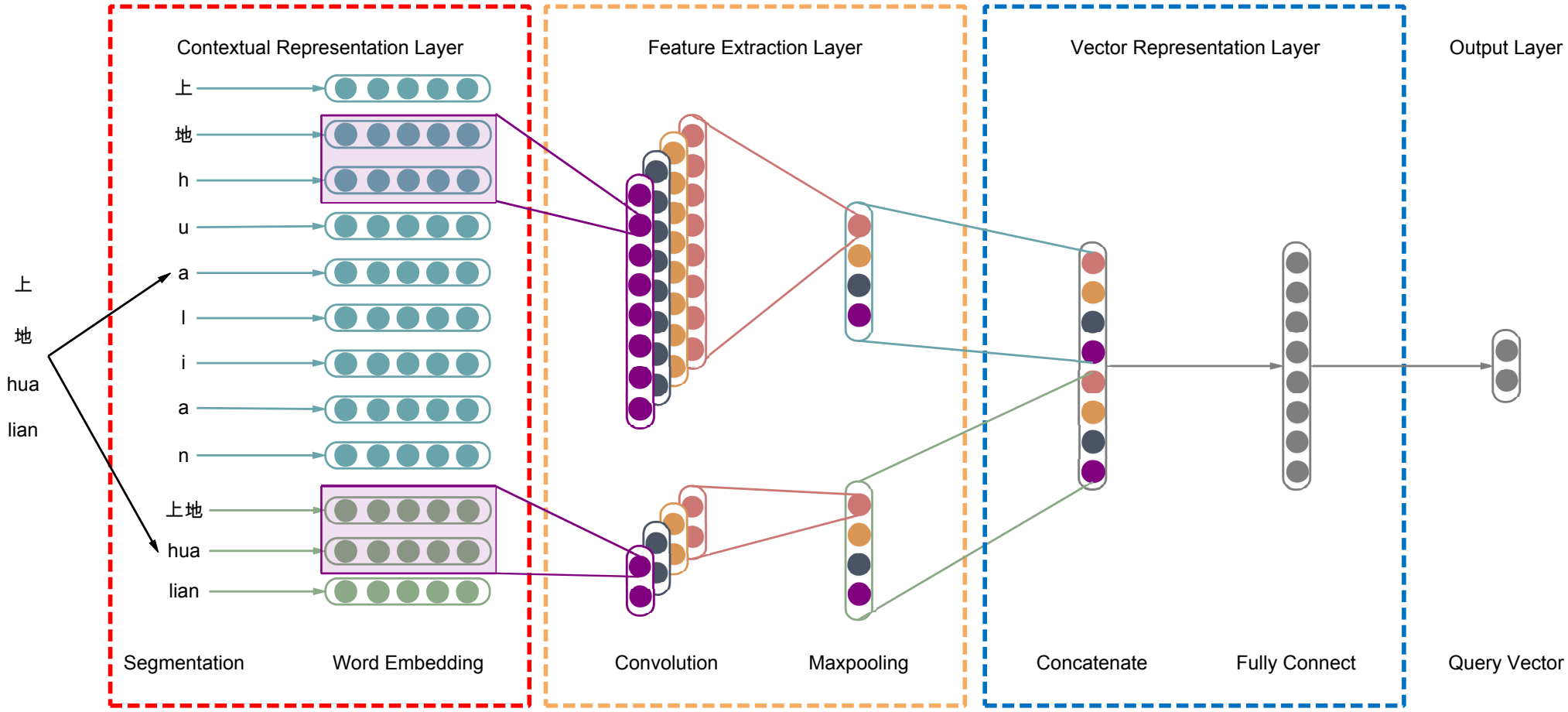
Query vector *Title vector*

{0.01224 0.875 -0.42.. -0.03 2.45 3.14}

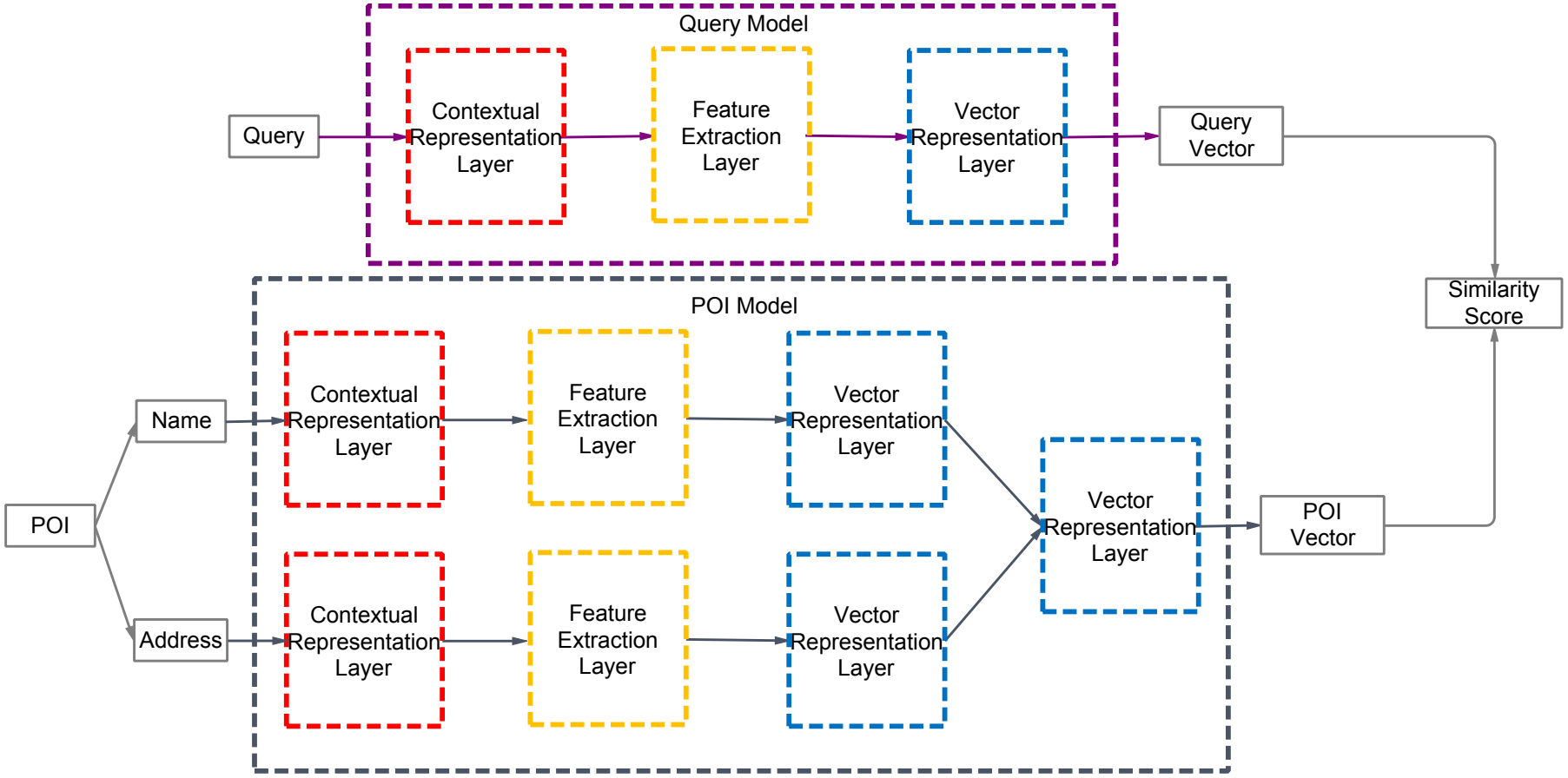
$$\text{cosine}(q, d) = \frac{q^T d}{\|q\|\|d\|}$$



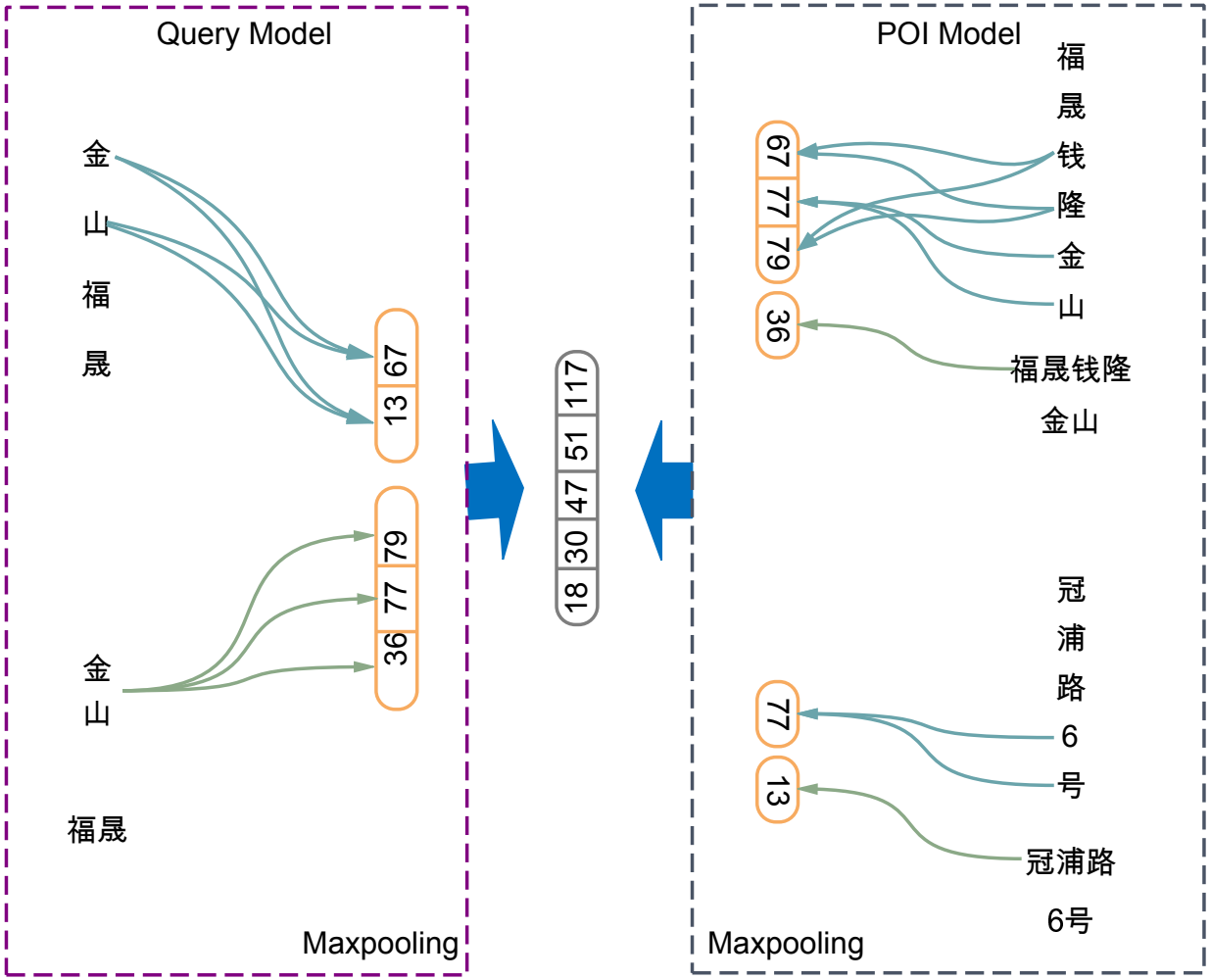
DPSM-Query Model



DPSM-Architecture

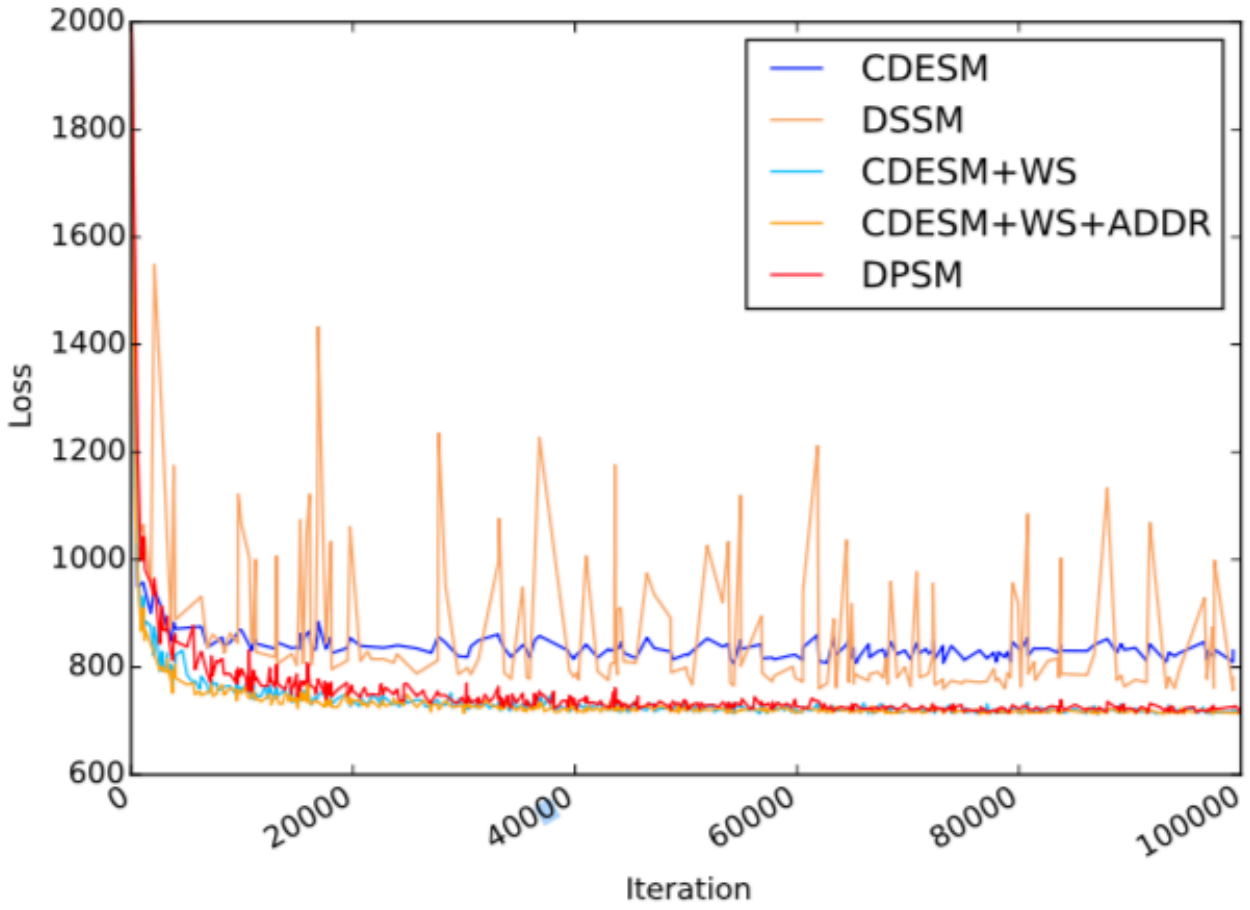


DPSM-Analysis



- Maxpooling 提取句子级别的特征
- Query 高激活神经元：36, 77, 79, 13, 67
- POI name 高激活神经元：36, 79, 77, 67
- POI address 高激活神经元：77, 13
- 语义匹配程度依赖相同高激活神经元个数

Experimental Results



Offline Click Experimental Results

Models	NDCG@3	NDCG@10
BM25	0.6891	0.8537
DSSM	0.7319	0.8726
CDESM	0.7355	0.8743
CDESM+WS	0.7435	0.8776
CDESM+WS+ADDR	0.7444	0.8776
DPSM	0.7524	0.8812

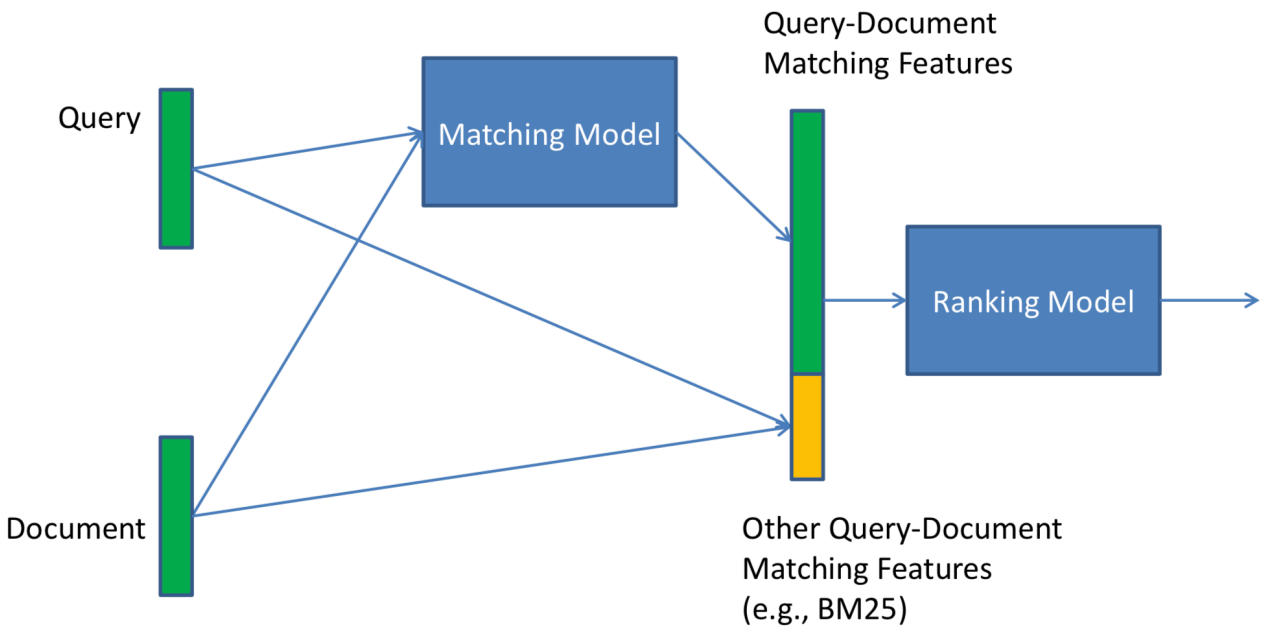
Offline Relevance Experimental Results

Models	NDCG@3	NDCG@10
BM25	0.5735	0.5781
DSSM	0.6021	0.6099
CDESM	0.6080	0.6188
CDESM+WS	0.6198	0.6256
CDESM+WS+ADDR	0.6203	0.6212
DPSM	0.6231	0.6333

Online Experimental Results

Exp	NDCG@10
online	0.8208
online+DPSM	0.8548

Future work



- 方法：
  同时学习匹配模型和排序模型
- 匹配模型：
  CNN，提取Query和Doc语义向量
- 排序模型：
  DNN，输入匹配模型特征和其他特征

03 第三章

Seq2Seq：纠错中的应用

问题转化

query改写行为

纠错
- 橘子酒店 → 桔子酒店
- 三里tt → 三里屯

同义
- 地铁口 → 地铁站
- 工体北门 → 工人体育场北门

query省略
- 物美超市 → 物美

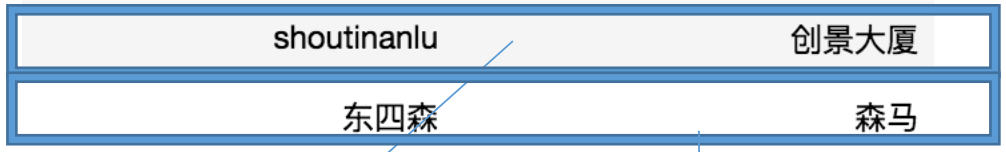
翻译问题（Seq2Seq）

原序列
- 用户输入的原始query

目标序列
- 用户点击的poi title
- session中用户主动改变的query

纠错语料处理

query	title
招商银行北京市朝阳区	招商银行
北京haite	北京海腾时代科技有限公司
东单小报房	小报房胡同
龙湖北京大兴购物	龙湖北京大兴天街购物中心
怀柔huan	怀柔区环保局
怀柔huan	怀柔环球国际影城
怀柔huan	黄花城水长城旅游区
家和事	家和事兴文化传播中心
北京市huojian	中国人民解放军火箭军总医院
和平门菜市场西	和平门菜市场
宝利欣苑	保利欣苑东门
tanxiaoxi	坛笑香



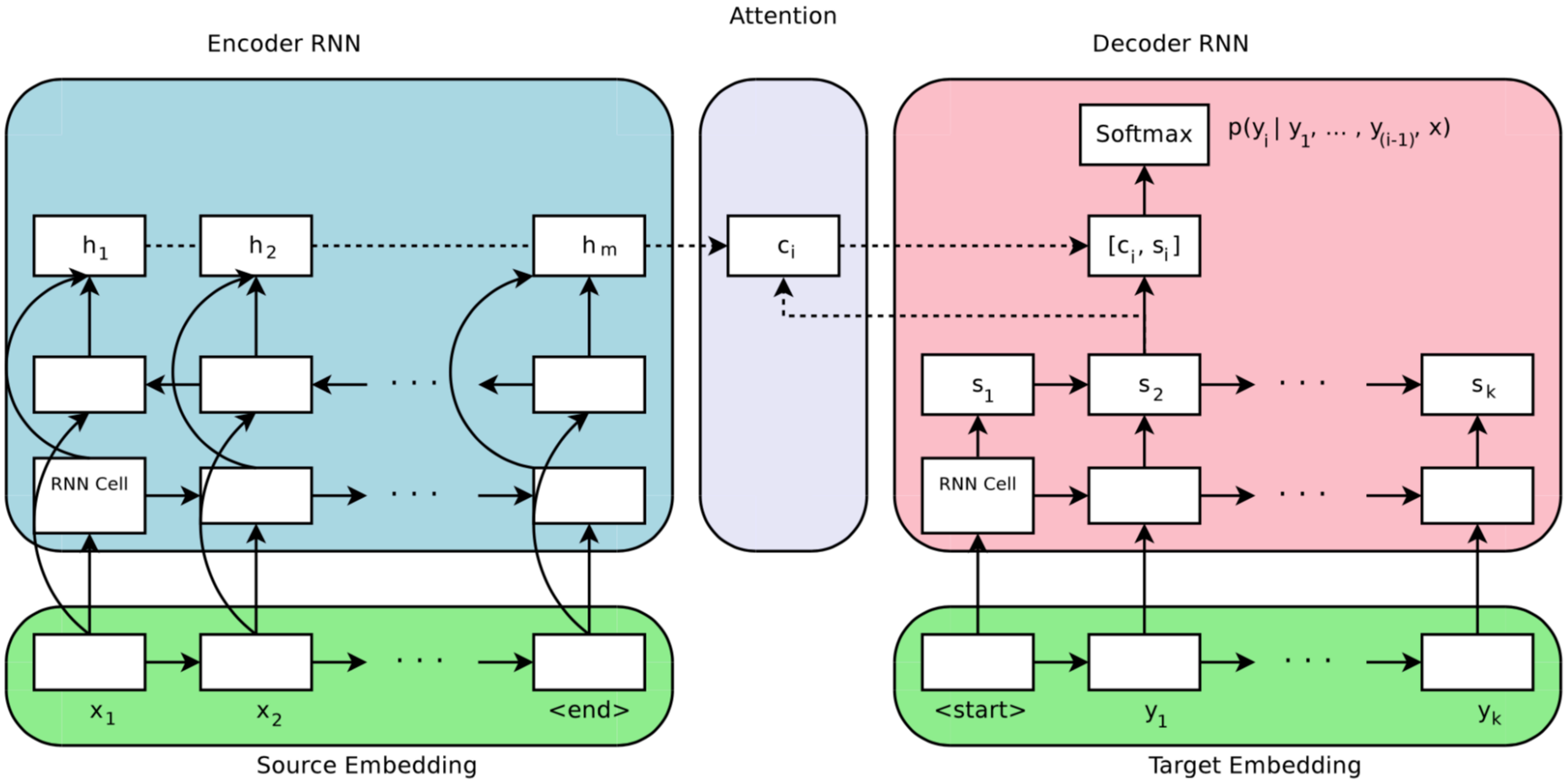
拼音编辑距离较大，丢弃

相对编辑距离较大，丢弃

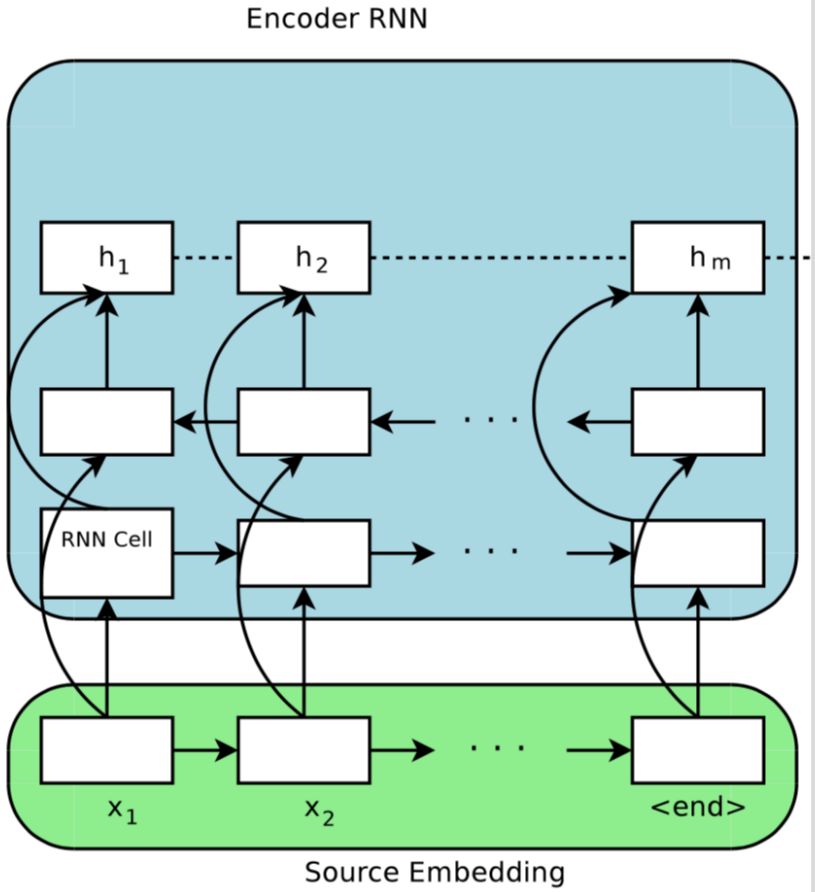
长度差别较大，截断成相同字数

拼音编辑距离较小，保留

Seq2Seq Encoder-Decoder architcture



Seq2Seq-Encoder



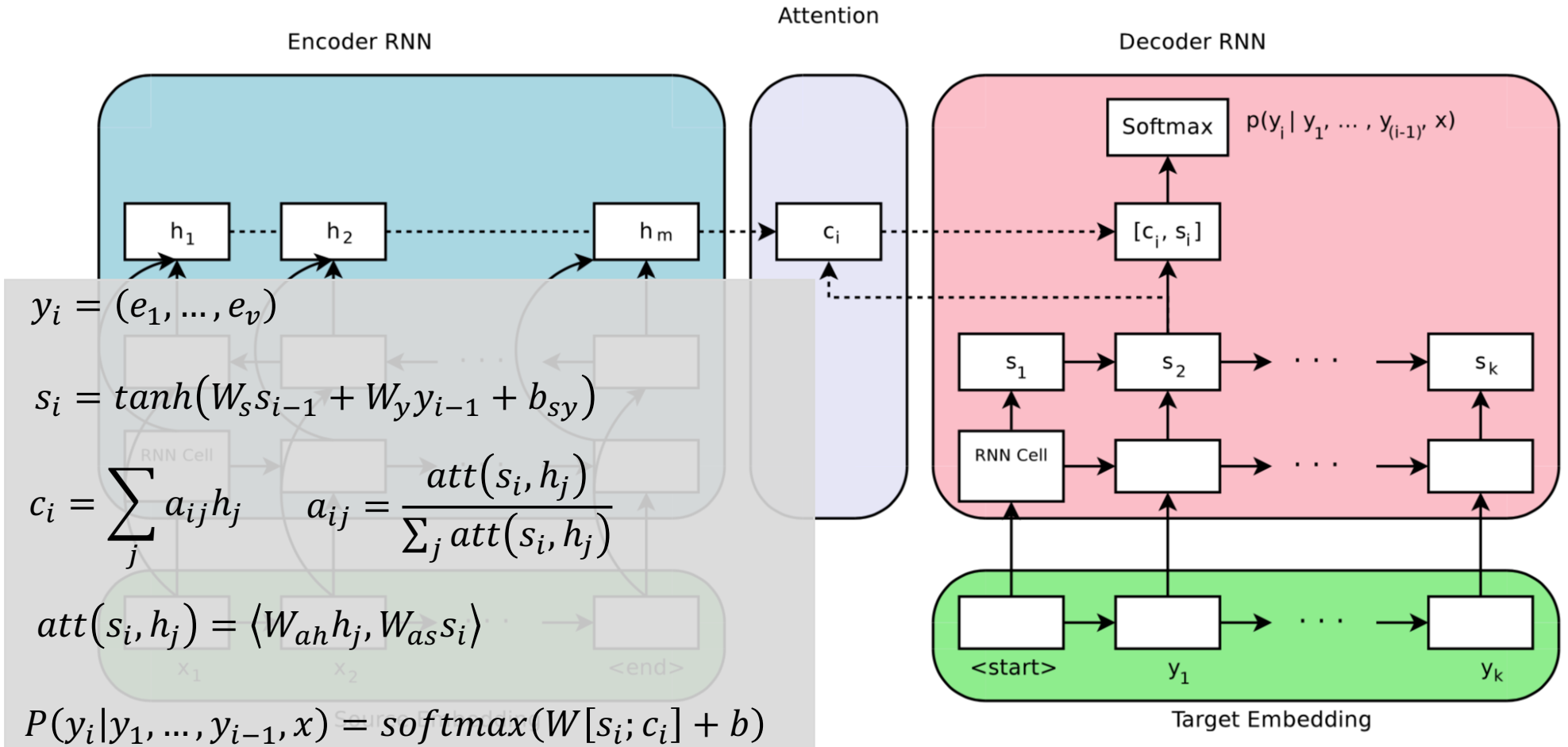
$$x_i = (e_1, \dots, e_v)$$

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$$

$$\overrightarrow{h_i} = tanh\left(W_{\overrightarrow{h}}\overrightarrow{h_{i-1}} + W_x x_i + b_{\overrightarrow{hx}}\right)$$

$$\overleftarrow{h_i} = tanh\left(W_{\overleftarrow{h_i}}\overleftarrow{h_{i-1}} + W_x x_i + b_{\overleftarrow{hx}}\right)$$

Seq2Seq-Decoder



$$y_i = (e_1, \dots, e_v)$$

$$s_i = tanh\left(W_s s_{i-1} + W_y y_{i-1} + b_{sy}\right)$$

$$c_i = \sum_j a_{ij} h_j \qquad a_{ij} = \frac{att(s_i, h_j)}{\sum_j att(s_i, h_j)}$$

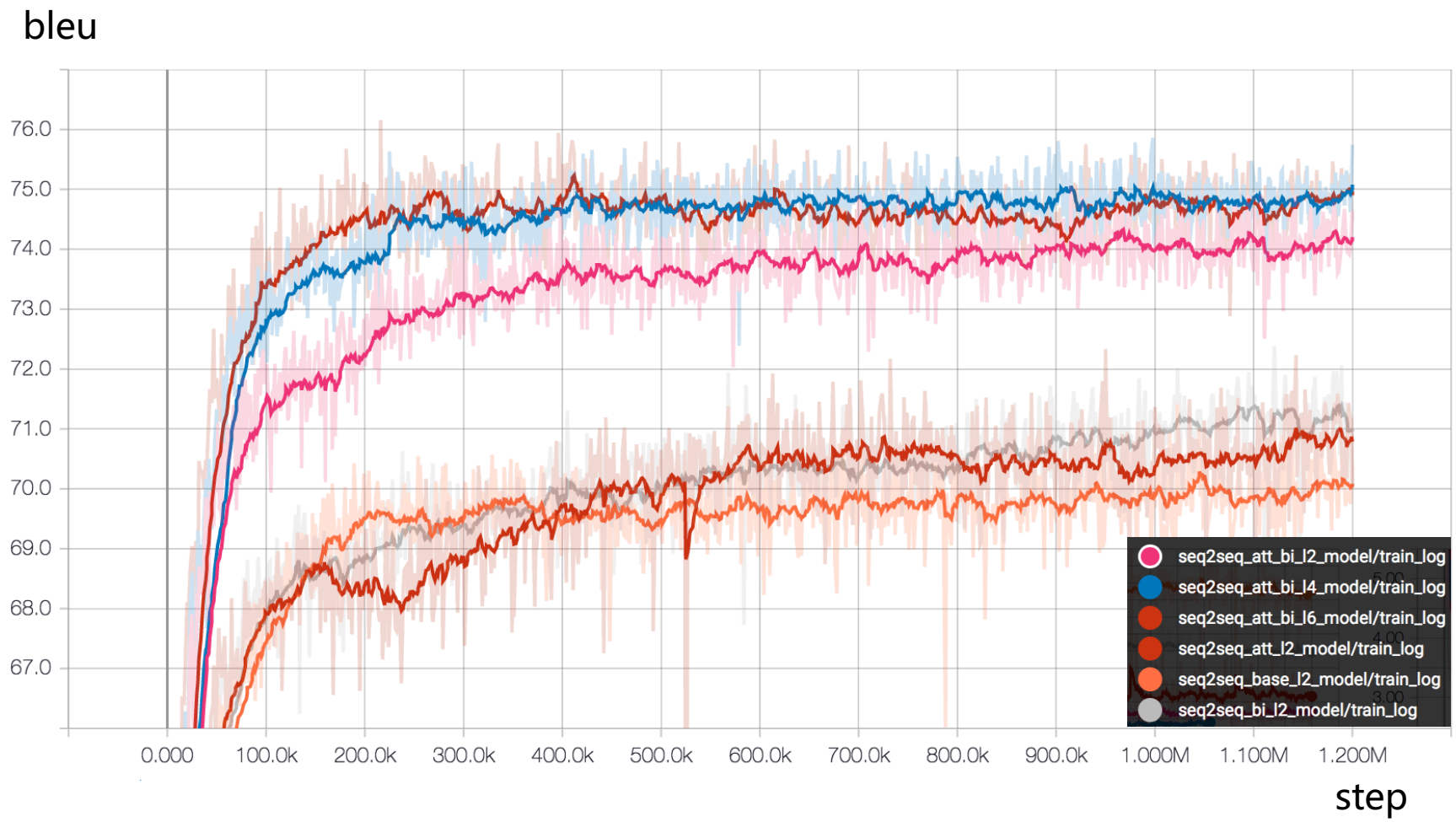
$$att(s_i, h_j) = \langle W_{ah} h_j, W_{as} s_i \rangle$$

$$P(y_i | y_1, \dots, y_{i-1}, x) = softmax(W[s_i; c_i] + b)$$

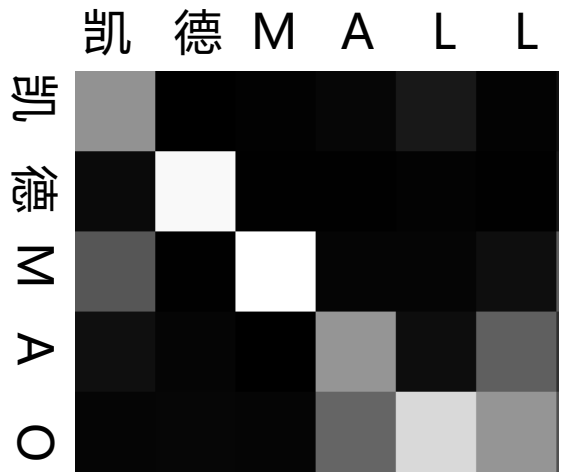
Seq2Seq –Hyperparameters

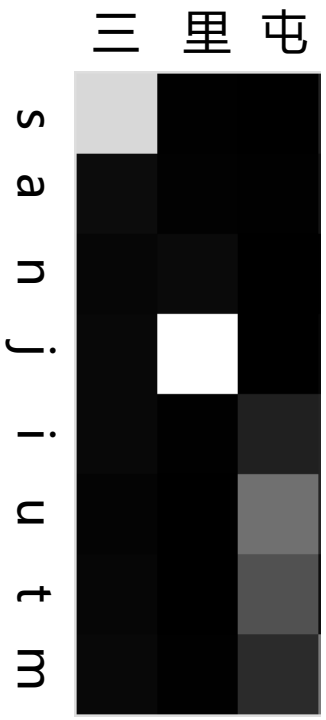
- Embedding Dimensionality
  - 128~2048，一般128就够用了
- RNN Cell Variant
  - LSTM 比 GRU效果好
- Bidirectional Encoder
  - 双向比单向提升明显
- Attention Mechanism
  - 使用Attention提升明显
  - 配合双向Encoder使用提升更明显
- Encoder and Decoder Depth
  - 4层encoder/decoder效果稳定

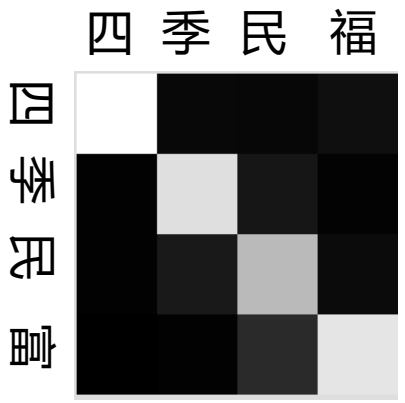
纠错实验效果-BLEU

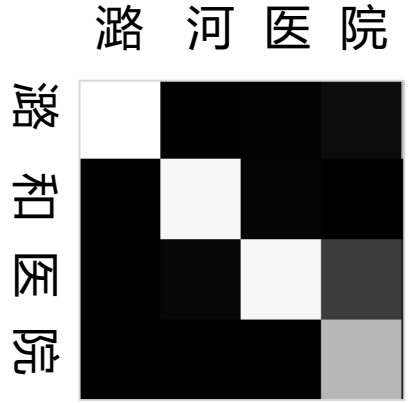

bleu
76.0
75.0
74.0
73.0
72.0
71.0
70.0
69.0
68.0
67.0
0.000
100.0k
200.0k
300.0k
400.0k
500.0k
600.0k
700.0k
800.0k
900.0k
1.000M
1.100M
1.200M
step
seq2seq_att_bi_l2_model/train_log
seq2seq_att_bi_l4_model/train_log
seq2seq_att_bi_l6_model/train_log
seq2seq_att_l2_model/train_log
seq2seq_base_l2_model/train_log
seq2seq_bi_l2_model/train_log

纠错实验效果-Attention









纠错实验效果-准确率

模型	TOP1准确率	TOP3准确率
HMM	86.0%	95.0%
UI-Enc-Dec-2Layers_160W	79.0%	90.7%
BI-Enc-2Layers_160W	81.0%	92.0%
Att-2Layers_160W	80.7%	91.2%
BI-Enc-Att-2Layers_160W	84.3%	91.8%
BI-Enc-Att-4Layers_160W	85.0%	92.0%
BI-Enc-Att-6Layers_160W	84.9%	92.5%
BI-Enc-Att-4Layers_450W	88.9%	93.0%

后续效果改进和应用

- 语料扩展
  - session query pair
  - hmm model result
- 地域特征
  - 增加城市维度的上下文
  - add city embeddings
- 场景扩展
  - 国际化多语言
  - 同义/省略

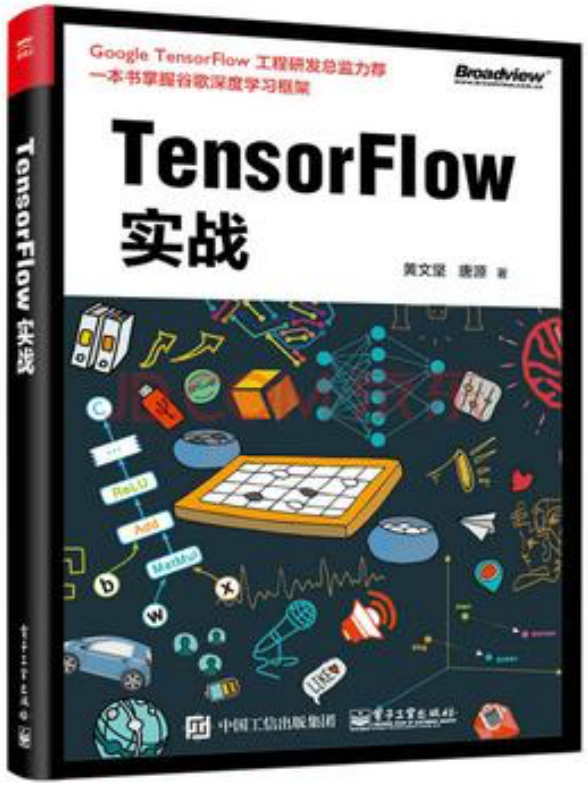
小结

1、出行场景下检索业务形态

2、Sent2Vec如何解决短文本匹配问题

3、Seq2Seq如何解决纠错问题

推荐书籍 / 拓展阅读材料



http://ml.intra.xiaojukeji.com

https://github.com/faneshion/MatchZoo

https://github.com/tensorflow/mnt

https://github.com/tensorflow/tensor2tensor

感谢大家的反馈！

Q&A