智能问答系统的探索与实践

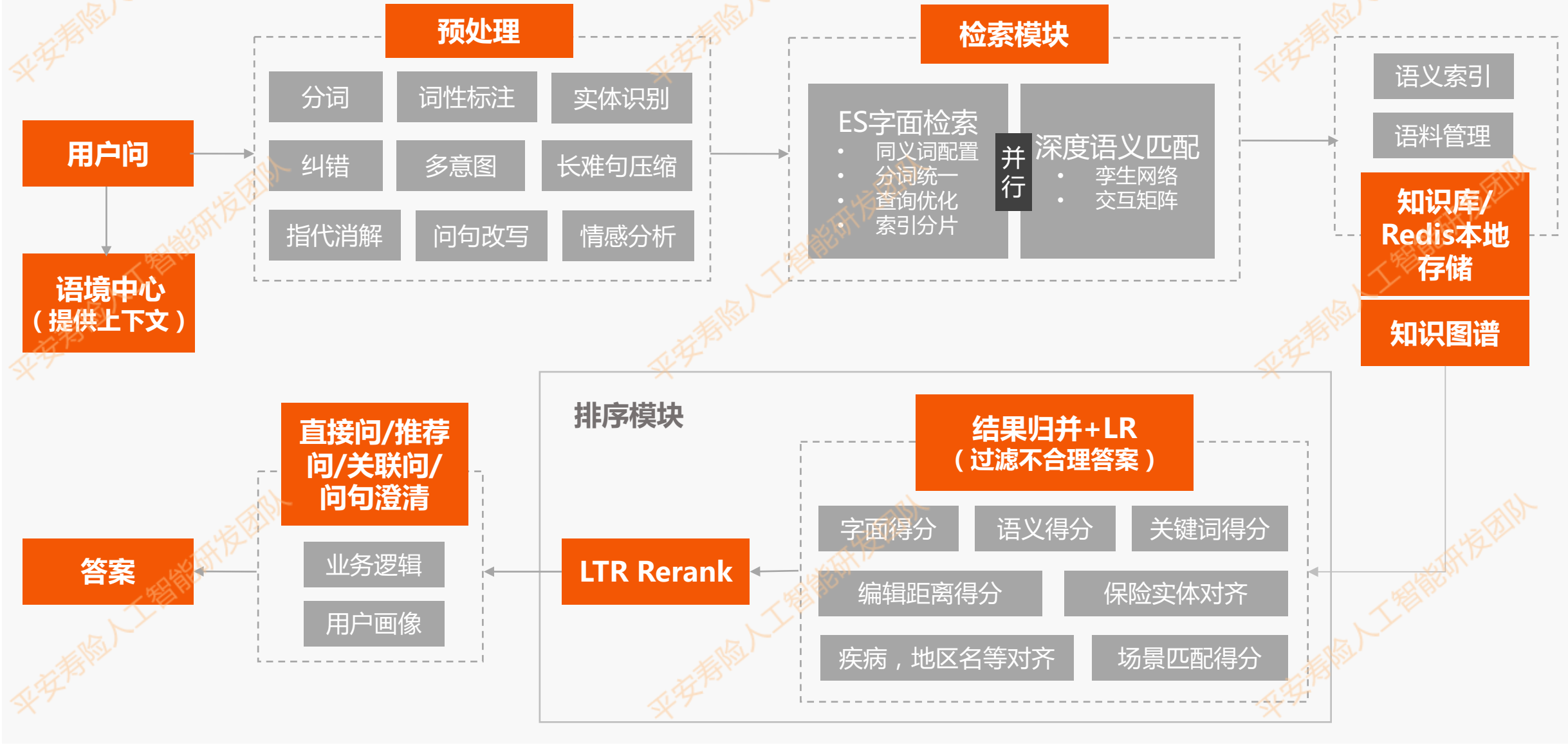
谢舒翼

平安人寿智能平台团队资深算法工程师

CONTENTS

框架介绍：*平安人寿智能问答引擎算法架构*

中国平安人寿保险



预处理：*长难句*

语法树分析+关键词典：

- **Step1: 通过标点或空格分割长句成若干个短句，然后对短句分类，去掉口水语句。**
- **Step2: 基于概率和句法分析的句子压缩方案，只保留主谓宾等核心句子成分。配合保险关键词典，确保关键词被保留。**

示例1

”嗯你好，我之前06年的时候买了一个保险，嗯一年只交了518元，然后后面我就再也没有买了，是06年的事情，然后现在我打电话给那个服务热线吧，我想退保就是。”

-> ”我之前06年的时候买了保险，我想退保。”

示例2

”一年半之前做过小腿骨折手术，现在已修复，只是固定钢板还没取，准备8月取，医生说也可以不取，不影响正常生活和工作，能否投保康寿宝？”

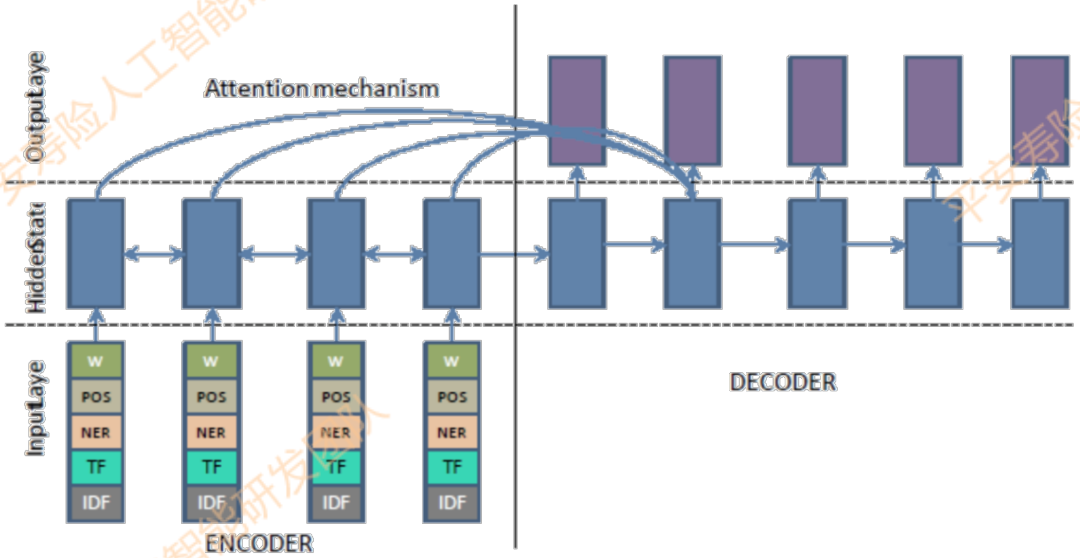
-> ”做过腿骨折手术，不影响正常生活和工作，能否投保康寿宝？”

预处理：*长难句*

中国平安人寿保险

文本摘要&句子压缩主流方法：一种是抽取式（extractive），另一种是生成式（abstractive）。

从传统的TextRank抽取式，到深度学习中采用RNN、CNN单元处理，再引入Attention、Self-Attention、机器生成摘要的方式，这些跟人类思维越来越像，都建立在对整段句子的理解之上。与此同时生成摘要的效果，也常常让我们惊艳。



Seq2Seq with Attention

- 基于NoisyChannel Model的句子压缩方法，2005
- Global Inference for Sentence Compression An Integer Linear Programming Approach，2008
- 基于概率统计和句法分析的中文语句压缩系统的研究与实现，2012
- A Neural Attention Model for Abstractive Sentence Summarization,2015,facebook,用神经网络做生成式摘要的开山之作
- Incorporating Copying Mechanism in Sequence-to-Sequence Learning,2016,华为诺亚方舟实验室,CopyNet
- LCSTS- A Large Scale Chinese Short Text Summarization Dataset,2016,一个中文文本摘要的数据集
- Get To The Point: Summarization with Pointer-Generator Networks,2017,Google

预处理：*纠错*

中国平安人寿保险

预处理：*指代消解*

中国平安人寿保险

实现思路：分词→词性标注→依存句法分析→主谓宾提取→实体替换/指代消解

举个栗子

输入1：感冒/nhd 可以/v 投保/vn 平安福/nbx 吗/y？

输入2： 那/rzv 癌症/nhd 呢/y？

输出：癌症可以投保平安福吗？

- 感冒 --(SBV)--> 投保
- 可以 --(ADV)--> 投保
- 投保 --(HED)--> ##核心##
- 平安福 --(VOB)--> 投保
- 吗 --(RAD)--> 投保
- ？ --(WP)--> 投保

更多栗子

输入1：平安福很好

输入2：这个产品比福满分好在哪？

-> 平安福比福满分好在哪？

输入1：平安福、福满分都是医疗险吗

输入2：前者能报销啥

-> 平安福能报销啥

输入1：平安福、福满分都是医疗险吗

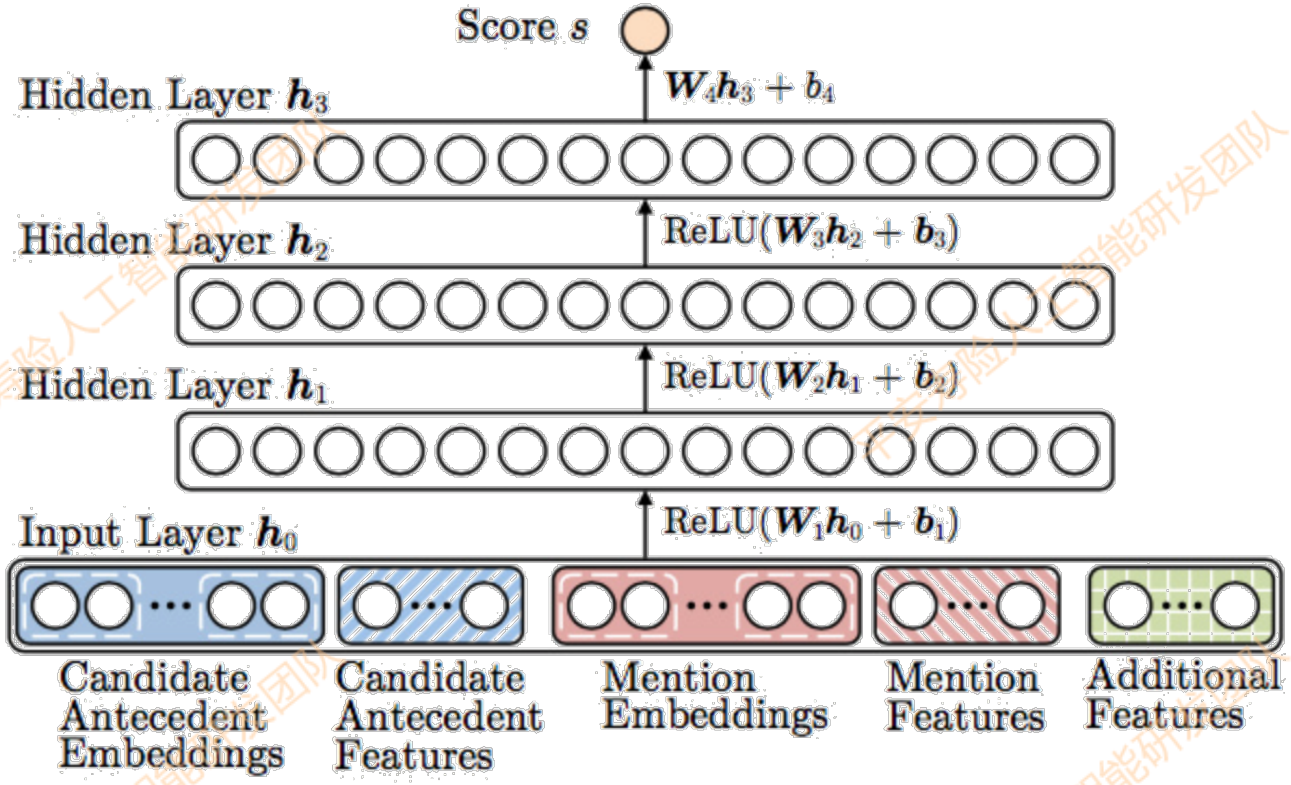
输入2：第一种能报销啥 -> 平安福能报销啥

输出：第二种呢 -> 福满分能报销啥

预处理：*指代消解*

待消解项，先行语可通过句法分析找出。

- Mention Pair models：将所有的指代词（短语）与所有被指代的词（短语）视作一系列pair，对每个pair二分类决策成立与否
- Mention ranking models：显式地将mention作为query，对所有candidate做rank
- Entity-Mention models：一种更优雅的模型，找出所有的entity及其对话上下文。根据对话上下文聚类，在同一个类中的mention消解为同一个entity。但这种方法其实也用得不多

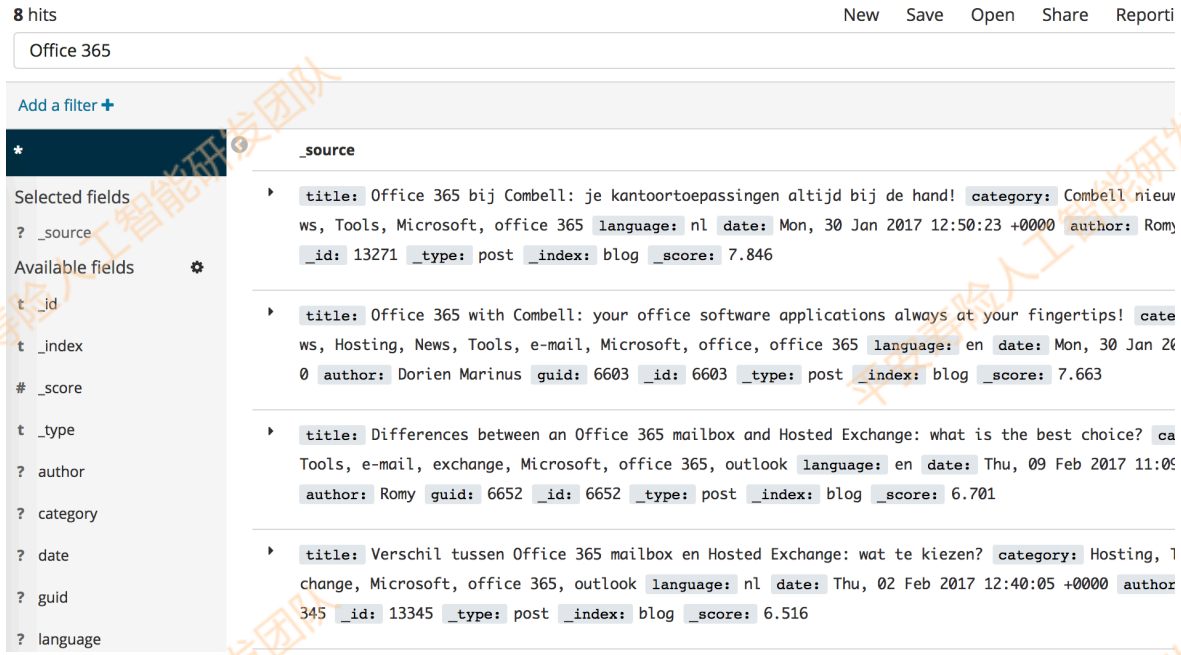


深度增强学习用于mention-ranking指代消解

检索和深度语义匹配：*ElasticSearch字面检索*

中国平安人寿保险

- 根据知识库数据建立索引
- 支持保险分类，机构查询过滤
- 分词统一，配置保险专用名词
- 同义词配置
- 可插拔的相似度算法
- 得分归一化
- 分片优化



Kibana可视化

检索和深度语义匹配：*孪生网络*

中国平安人寿保险

Siamese CBOW

预训练好词向量，用求和取平均的方式来表征句向量，对标准问和相似问进行训练，添加负采样，损失函数为Contrastive Loss。让正样本之间的句向量表征尽量相似。预先算出语料的所有句向量表征，将用户问题通过模型转化成句向量，搜索语料里最相似的若干个句向量作为候选答案列表。

- Siamese CBOW- Optimizing Word Embeddings for Sentence Representations，2016，Yandex
- Learning Text Similarity with Siamese Recurrent Networks，2016
- Siamese Recurrent Architectures for Learning Sentence Similarity，2016，MIT

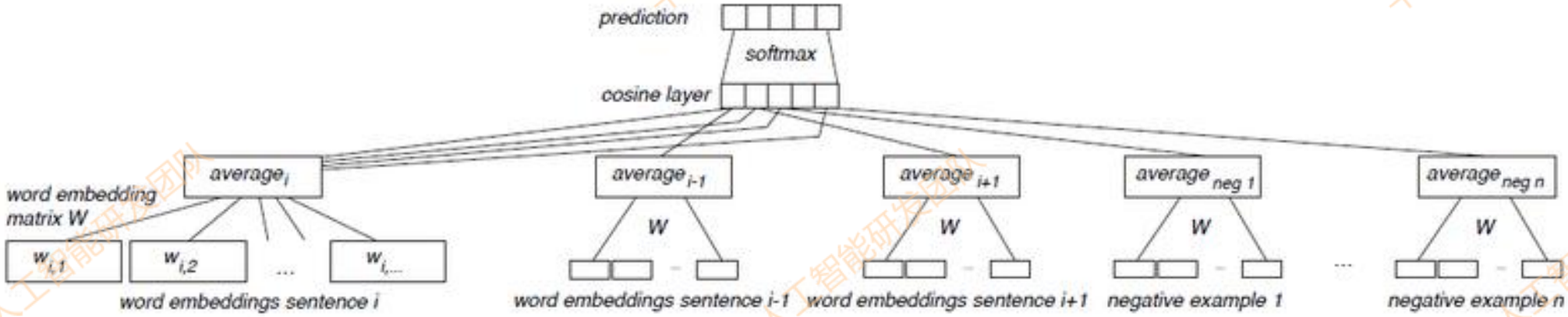


Figure 1: Siamese CBOW network architecture. (Input projection layer omitted.)

检索和深度语义匹配：*孪生网络优化点*

中国平安人寿保险

Embedding：词向量，字向量，上下文词向量

- Character-based Neural Networks for Sentence Pair Modeling,2018, character-based ngram

知识清洗

- 标准问/相似问清洗

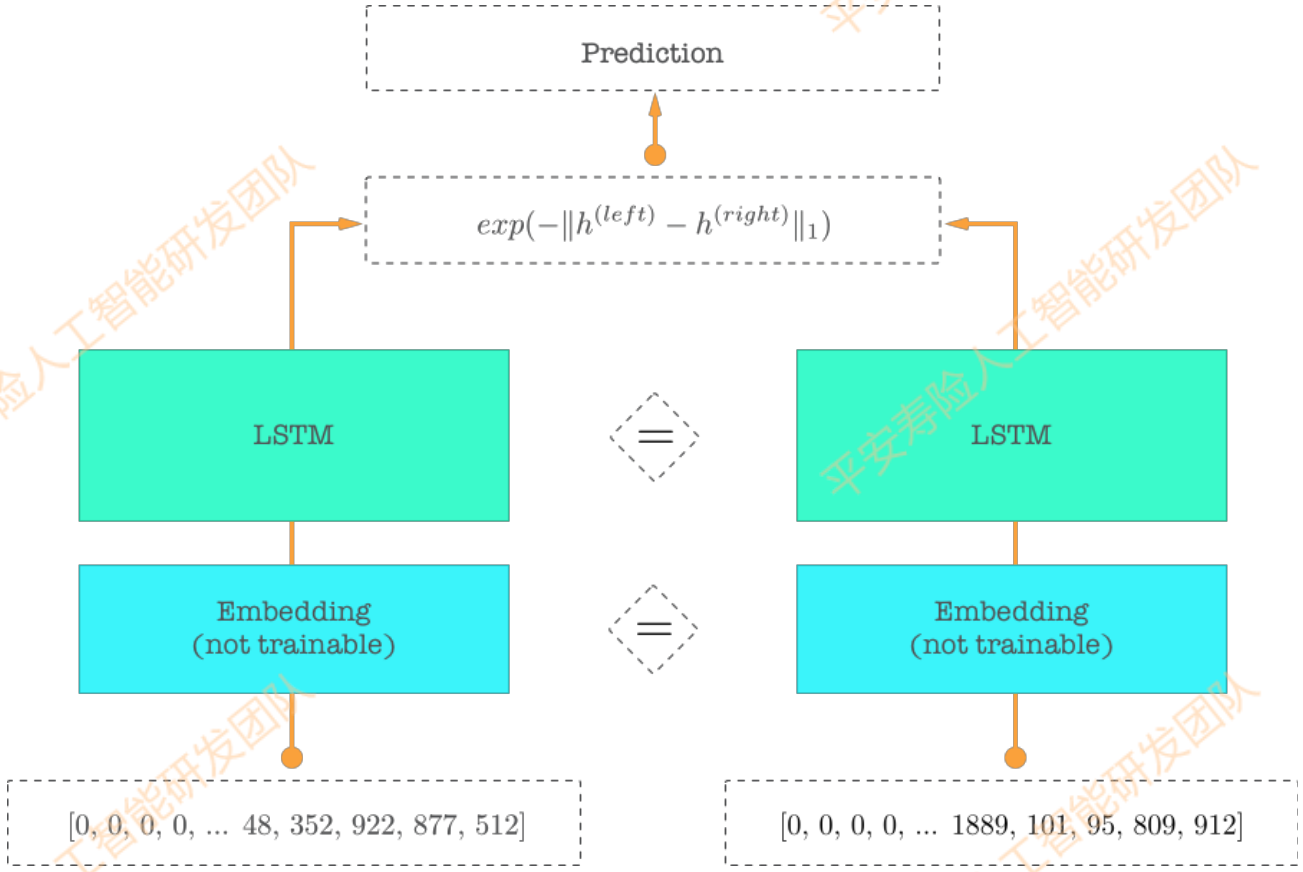
数据扩充

- 增大正负训练样本，扩大词向量维度，引入预训练模型bert，更强表征

不同模型支持

- CBOW/LSTM/RNN/CNN, transfer learning

对网络output的位置使用attention而不是直接对每个时刻的输出求均值



2017 Kaggle冠军解决方案：Siamese MaLSTM

检索和深度语义匹配：*BERT for QA*

中国平安人寿保险

NLP四大任务

1）序列标注，分词/POS Tag/NER/语义标注等；**2）分类任务**，文本分类/情感计算等；**3）句子关系判断**，Entailment/QA/自然语言推理等；**4）生成式任务**，机器翻译/文本摘要等。

Bert as a service: Mapping a variable-length sentence to a fixed-length vector using BERT model

https://github.com/hanxiao/bert-as-service#building-a-qa-semantic-search-engine-in-3-minutes

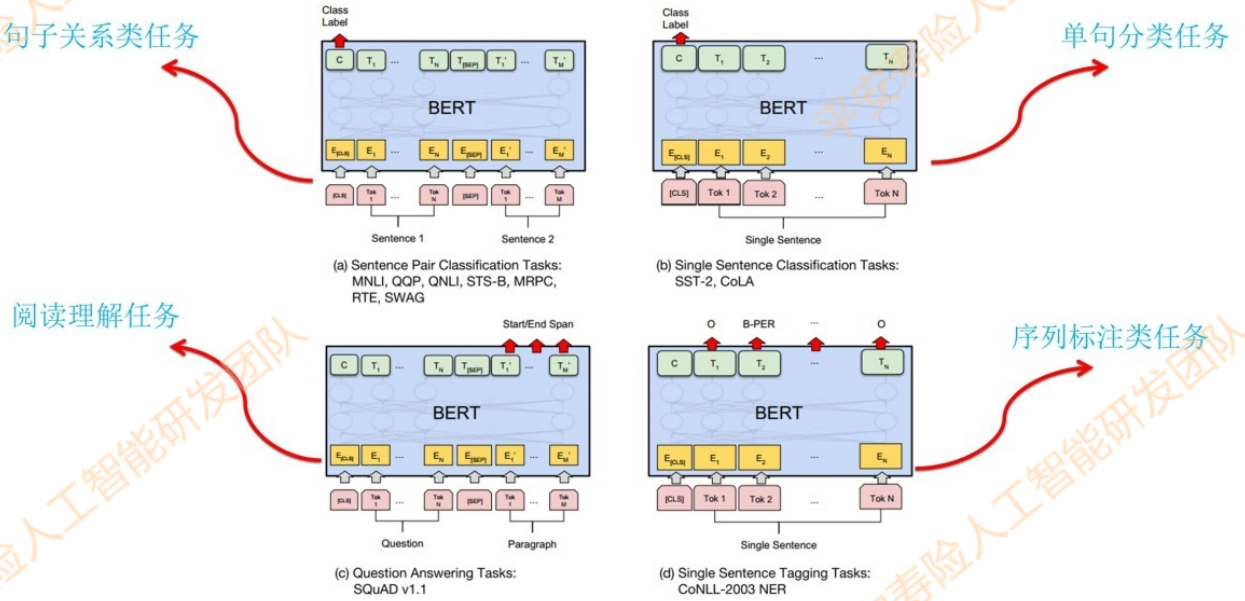
BERT- Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, Google

Fine-tune支持问答任务:

新增自定义Processor

并重载获取label的get_labels和获取单个输入的

get_train_examples,get_dev_examples和get_test_examples函数。其分别会在main函数的FLAGS.do_train、FLAGS.do_eval和FLAGS.do_predict阶段被调用。



Bert：如何改造下游任务？

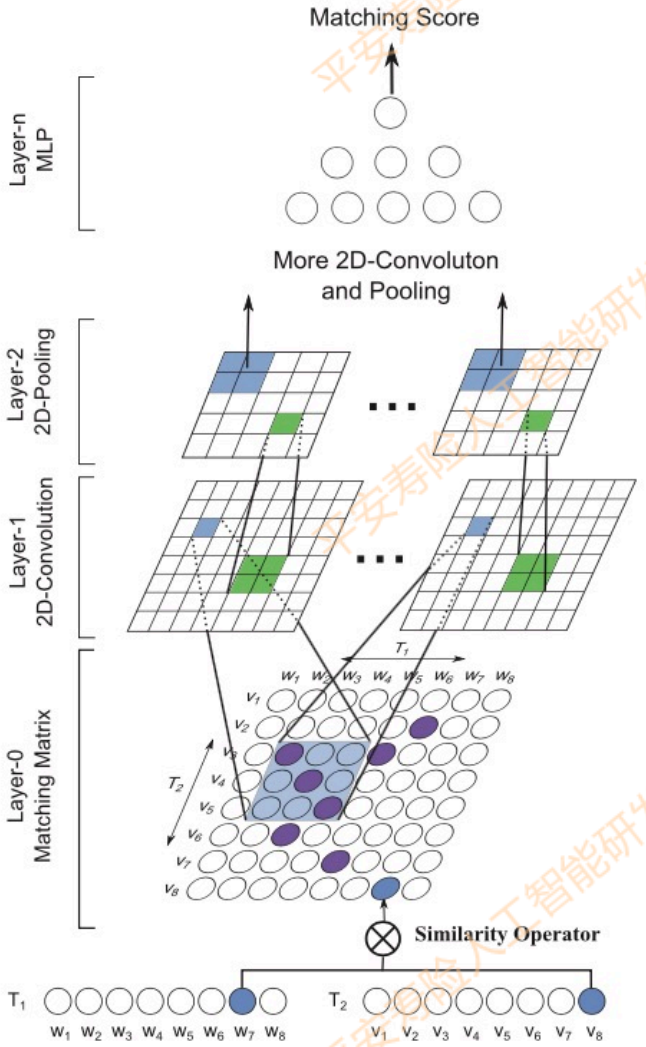
检索和深度语义匹配：*交互矩阵*

中国平安人寿保险

MatchPyramid

- 借鉴了cnn在处理图像时的原理，因为cnn就是在提取像素，区域之间的相关性，进而提取图像的特征。那么，就可以将单词与单词之间的相似度看成是像素，那么对于两个单词数为M,N的句子，其相似度矩阵就是M*N，然后！就可以用卷积搞事情了。
- 衡量单词之间的相似度：
  - ✓ 如果单词相同就是1，否则为0
  - ✓ 计算词向量的余弦距离
  - ✓ 计算词向量的内积
- 卷积提取ngram特征，MLP拟合相似度得分

Text Matching as Image Recognition_MatchPyramid，2016



检索和深度语义匹配：*MatchZoo*

中国平安人寿保险

MatchZoo，A Toolkit for Deep Text Matching

- 开源的Python 环境下基于 TensorFlow 开发的文本匹配工具。
- https://github.com/NTMC-Community/MatchZoo

```
# code example
>>> model = MVLSTM()
>>> model.params['lstm_units'] = 32
>>> model.params['top_k'] = 50
>>> model.params['mlp_num_layers'] = 2
>>> model.params['mlp_num_units'] = 20
>>> model.params['mlp_num_fan_out'] = 10
>>> model.params['mlp_activation_func'] = 'relu'
>>> model.params['dropout_rate'] = 0.5
>>>model.guess_and_fill_missing_params(verbose=0)
 >>> model.build()
 >>> model.compile()
 >>> model.backend.summary()
 >>> model.fit(x, y, batch_size=32, epochs=5)
 >>> model.predict(test_x)
```

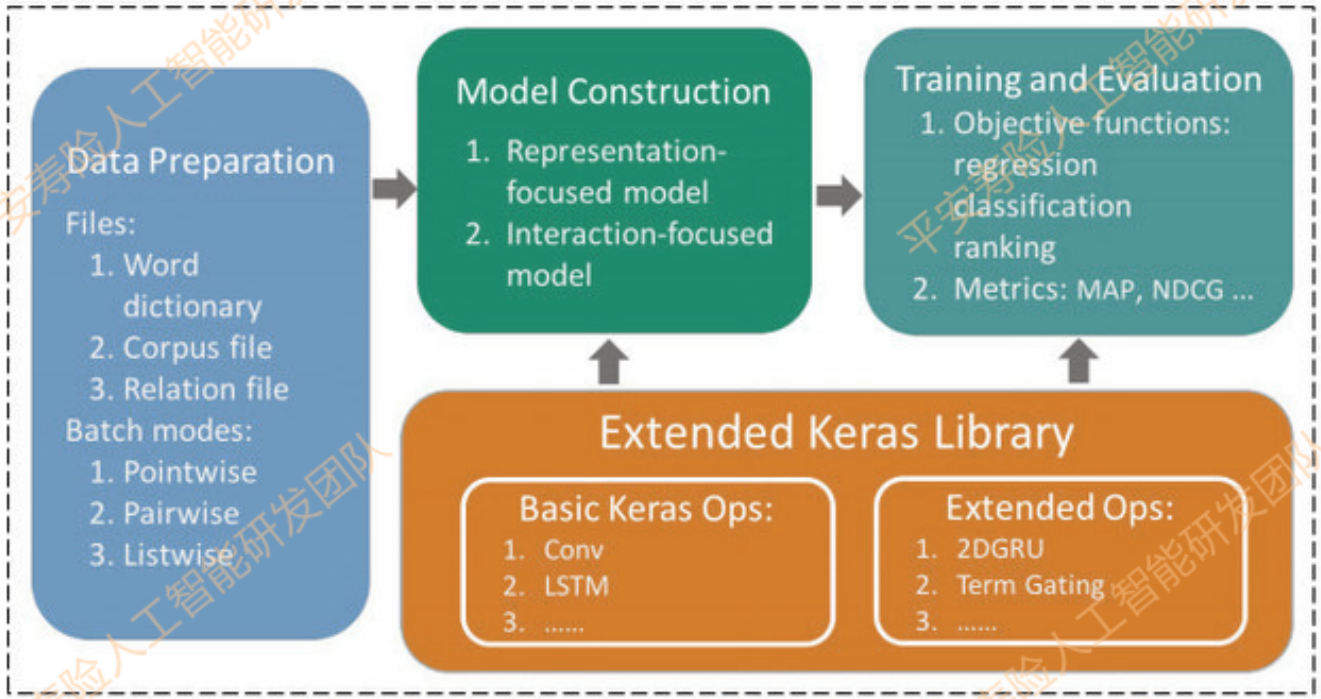


Figure 1: The Architecture of the MatchZoo toolkit.

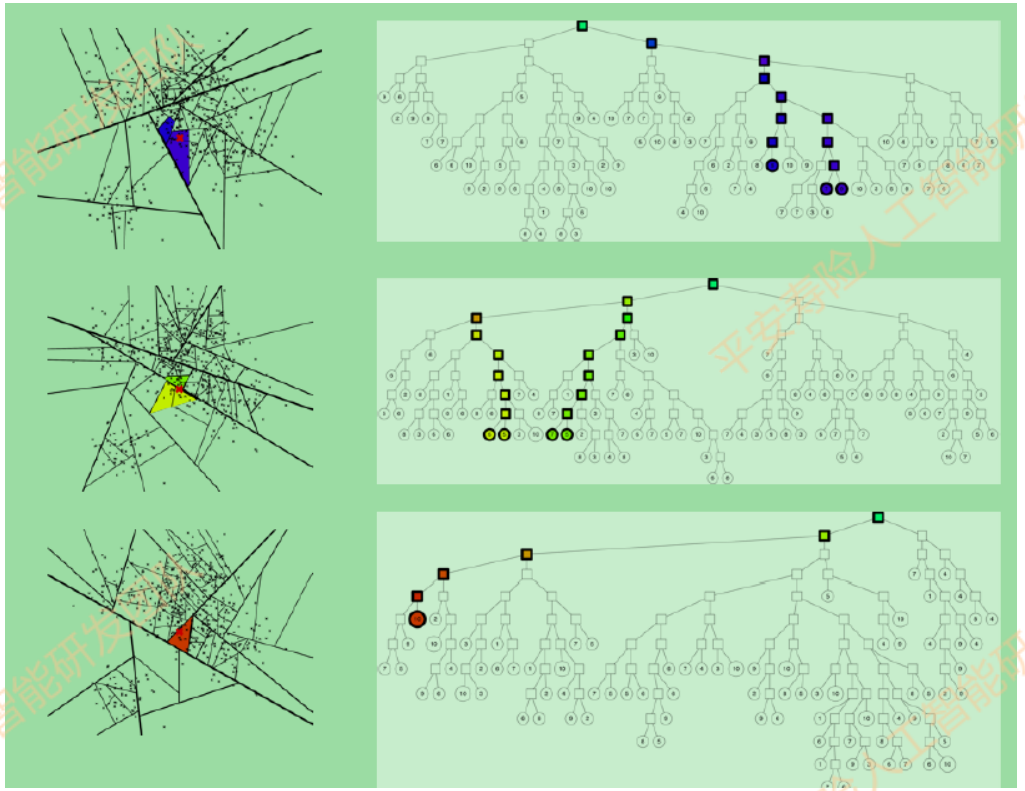
检索和深度语义匹配：*知识库和知识指引*

中国平安人寿保险

- 在知识库里录入知识，自动生成语义索引和字面索引。
- 语料更新无需发版。

✓ Annoy搜索算法：建立一个数据结构，使得查询一个向量的最近邻向量的时间复杂度是次线性。二叉树，随机选2个点聚类，超平面分割。同一批数据建立多棵树，检索答案合并排序。

✓ 小问题：第一次查询的速度比较慢；近似算法，准确率逼近100%。



排序：*Deeprank*

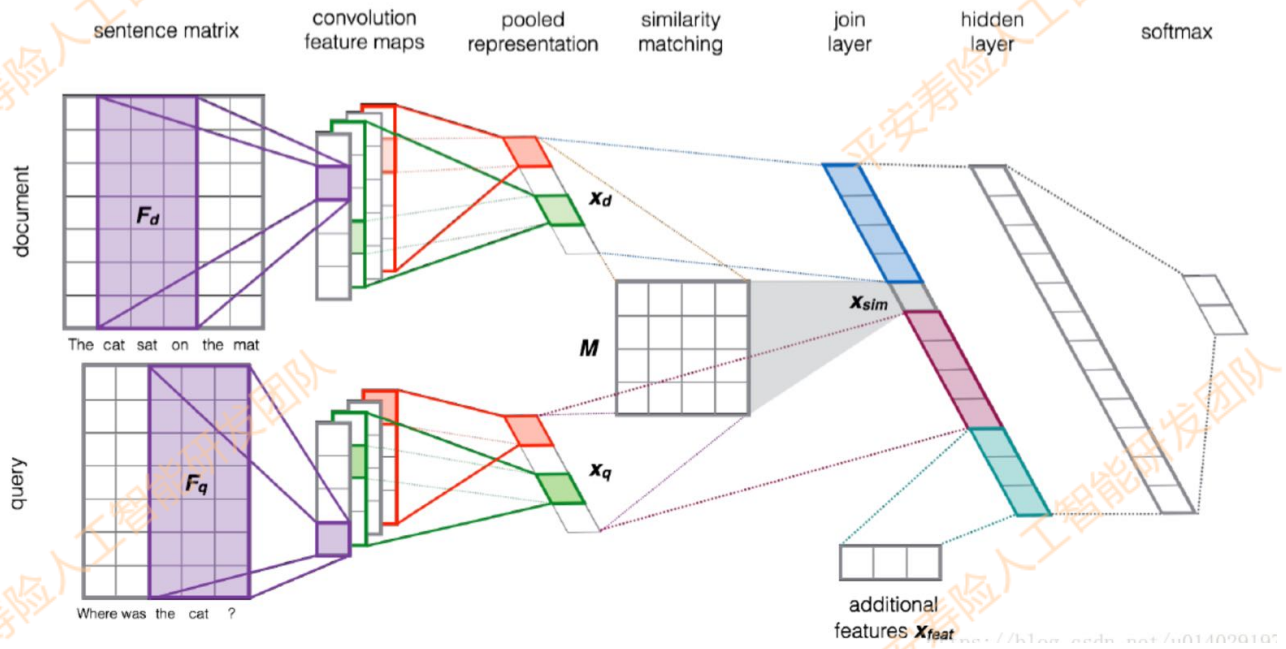
中国平安人寿保险

- 使用 n-gram 窗口，可以捕捉更长的上下文语义
- 将 query 和 document 的语义向量及其相似度拼接成新的特征向量输入 MLP 层进行 learning to rank
- 可以在 learng2rank 模型的输入向量中方便地融入外部特征
- 支持 end-to-end 的 matching + ranking 任务

排序打分方案Example

Query：医疗保险516只有住院才赔吗？

- S，我买的住院日额516如果住院可以申请理赔吗？
- B，如果被保险人生病住院，住院日额516能赔多少钱？
- D，平安福门诊医疗可以报销？

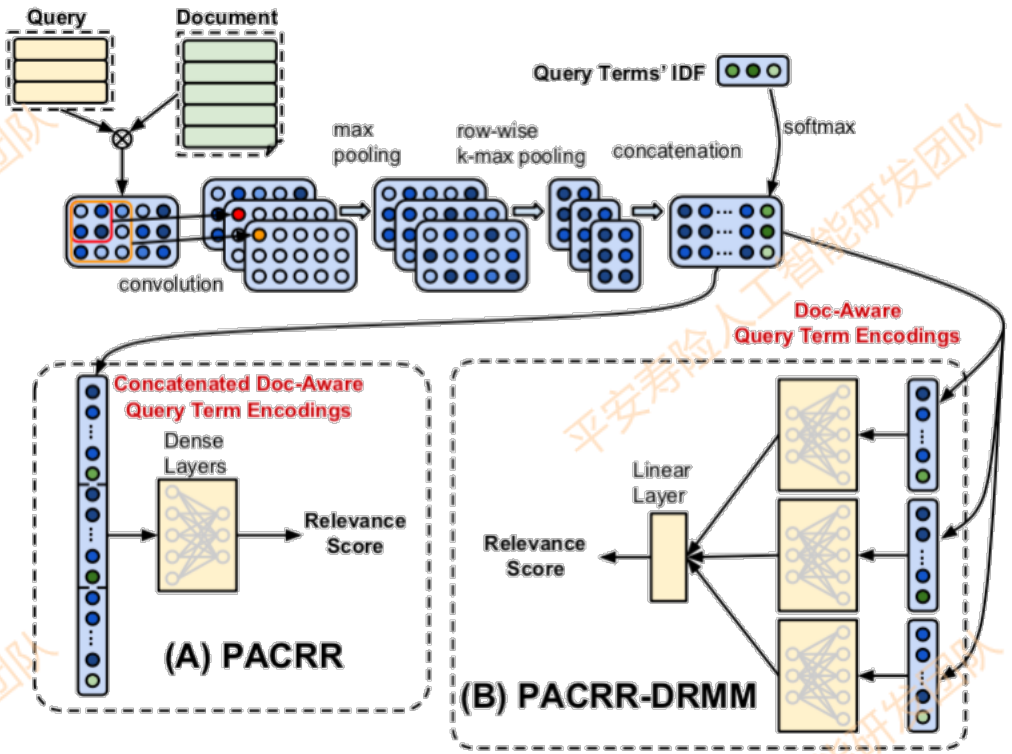


排序：*Deeprank*

中国平安人寿保险

DRMM+PACRR

- Context-sensitive Term Encodings构造qd相似度矩阵，卷积提取ngram信息，firstk，kwindow
- 两层max-pooling获取最强相似信息，拼接
- 使用相同的MLP网络独立地计算每一个q-term encoding（矩阵的每一行）的分数，再通过一个线性层得到query与doc的相关性得分 。

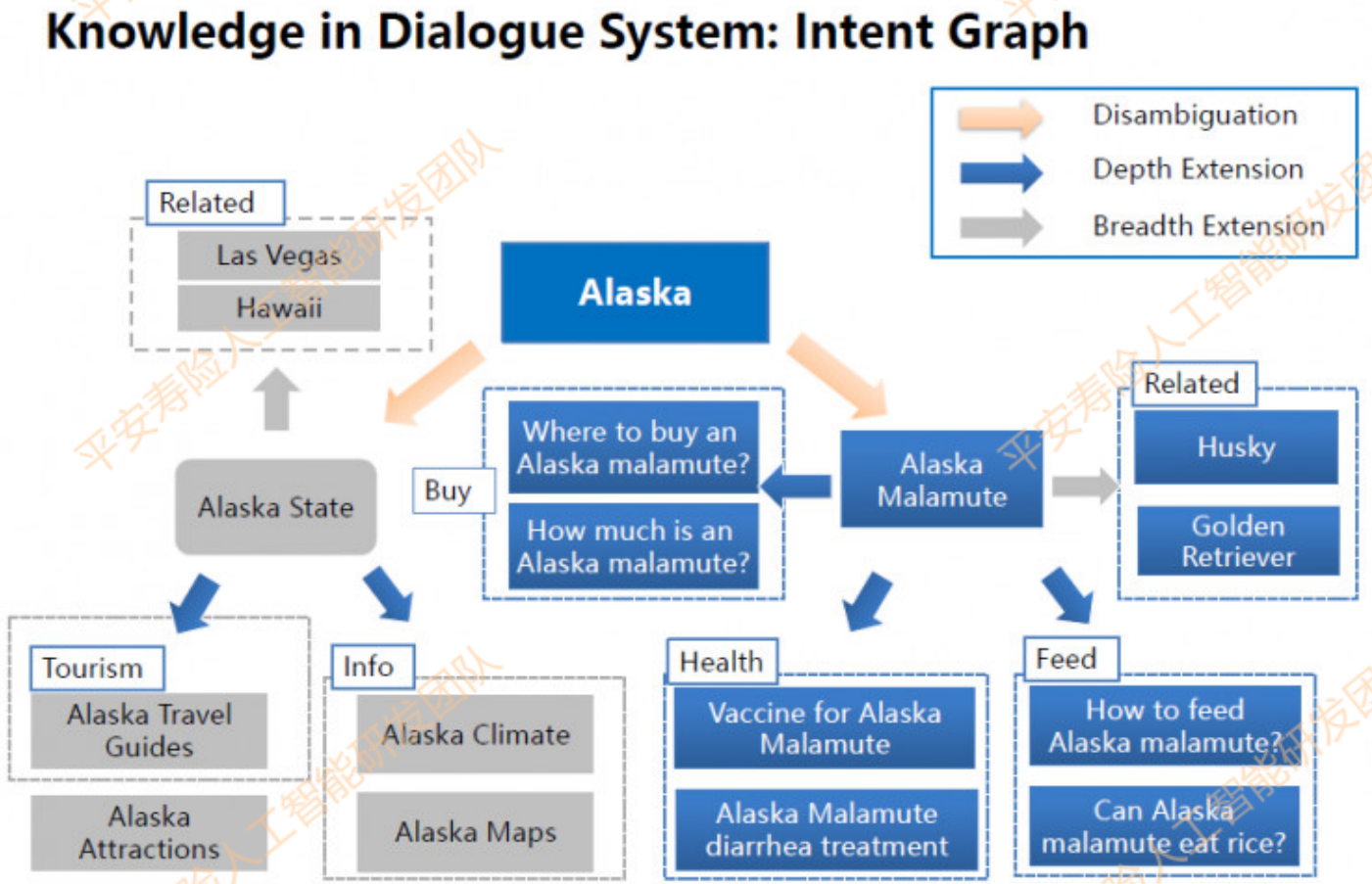


A Deep Relevance Matching Model for Ad-hoc Retrieval,2017,DRMM
PACRR- A Position-Aware Neural IR Model for Relevance Matching,2017,PACRR
Deep Relevance Ranking Using Enhanced Document-Query Interactions,2018,DRMM+PACRR

效果评估：*话术澄清*

中国平安人寿保险

- 问题有歧义或者匹配答案有置信度但不够高的时候触发。
- Example: 苹果多少钱？
- 意图图谱的节点代表一个个意图节点。这些“意图”之间的关系包括需求澄清（disambiguation）、需求细化（depth extension）、需求横向延展（breadth extension）等



效果评估：*直接回答/推荐问/关联问/搜索问答*

中国平安人寿保险

- 直接回答：Top1答案
- 推荐问：Top3答案
- 关联问：基于大数据的推荐算法挖掘
- 搜索式问答：非保险类问题，答案可以来自网络wiki
- 闲聊：检索式+生成式模型

类型	指标	备注
问答评估指标	有效问题数量	日志随机抽样统计可得出
	Top1准确率	日志随机抽样统计可得出
	Top3准确率	日志随机抽样统计可得出
	有效问题响应准确率	日志随机抽样统计可得出
	知识覆盖率	日志随机抽样统计可得出

效果评估：*自动化测试框架+验证集+效果评估*

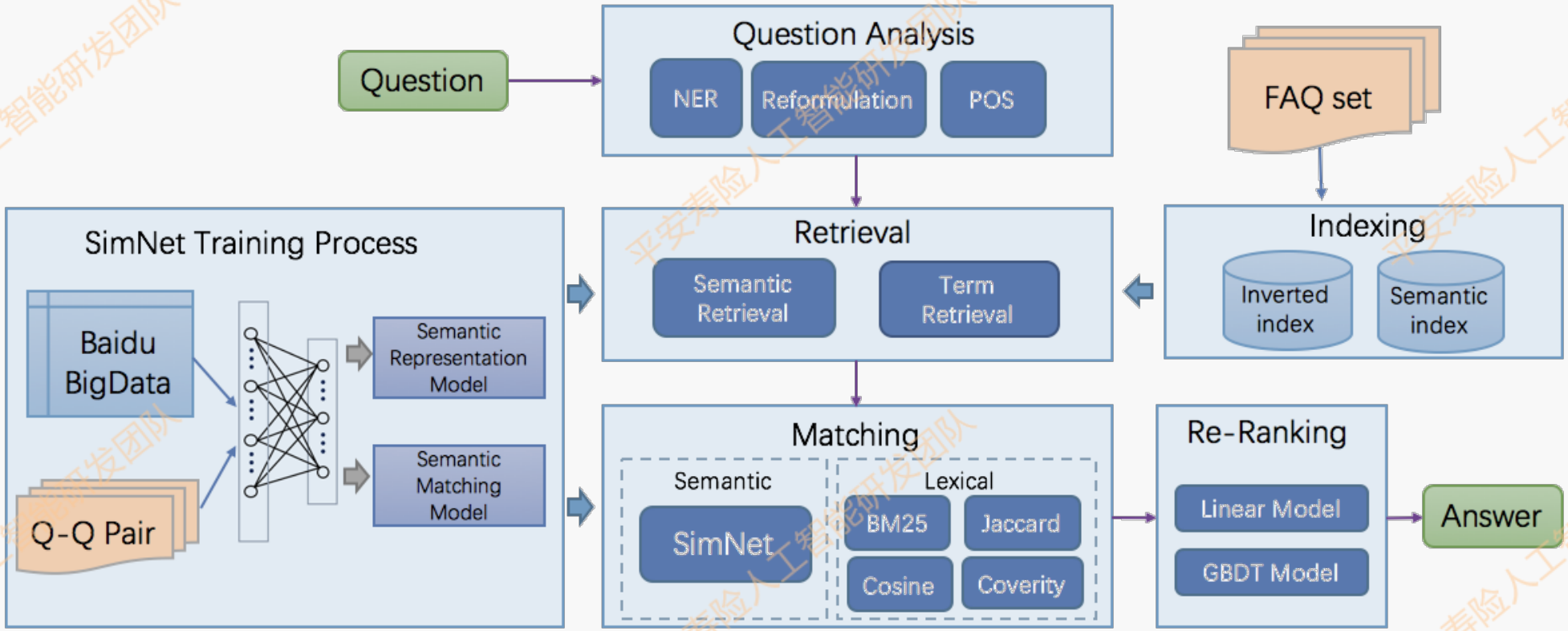
中国平安人寿保险

- 原始语料测试
- 五大验证集/测试集进行模型迭代优化和效果评估
- 语义验证集，根据业务需求编写+线上日志抽样标注
- 字面验证集，对语料进行删除非关键词，增加噪音，同义词转写等十余种方法进行生成
- 标注：直接匹配问，推荐问，top10标知识缺失

附录1：*AnyQ架构*

中国平安人寿保险

- 统一的训练，测试数据格式定义；
- 通过json配置文件可以灵活的选择网络类型，数据类型，损失函数以及其他超参数。



附录2：*MatchZoo Models*

中国平安人寿保险

Model List

Representation-focused model:

- **DSSM:** Learning Deep Structured Semantic Models for Web Search using Click through Data.
- **CDSSM:** Learning Semantic Representations Using Convolutional Neural Networks for Web Search.
- **ARC:** Convolutional Neural Network Architectures for Matching Natural Language Sentences.
- **MV-LSTM:** A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations.
- **Siamese-LSTM :** Siamese Recurrent Architectures for Learning Sentence Similarity.
- **DUET:** Learning to Match using Local and Distributed Representations of Text for Web Search.
- **ABCNN:** Attention-Based Convolutional Neural Network for Modeling Sentence Pairs.

Interaction-focused model:

- **GLM:** A Word Embedding based Generalized Language Model for Information Retrieval.
- **DESM:** A Dual Embedding Space Model for Document Ranking.
- **MultiGranCNN:** An Architecture for General Matching of Text Chunks on Multiple Levels of Granularity.
- **DRMM:** A Deep Relevance Matching Model for Ad-hoc Retrieval.
- **CONV-KNRM**: Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search.
- **match-LSTM:** Learning Natural Language Inference with LSTM.
- **DRMM-TKS:** A Deep Relevance Matching Model for Ad-hoc Retrieval (*A variation of DRMM).
- **BiMPM:** Bilateral Multi-Perspective Matching for Natural Language Sentences.
- **DLCM:** Learning a Deep Listwise Context Model for Ranking Refinement. *NEW, TO BE APPEARED IN SIGIR 2018*
- **ESIM:** Enhanced LSTM for Natural Language Inference * ACL 2017.

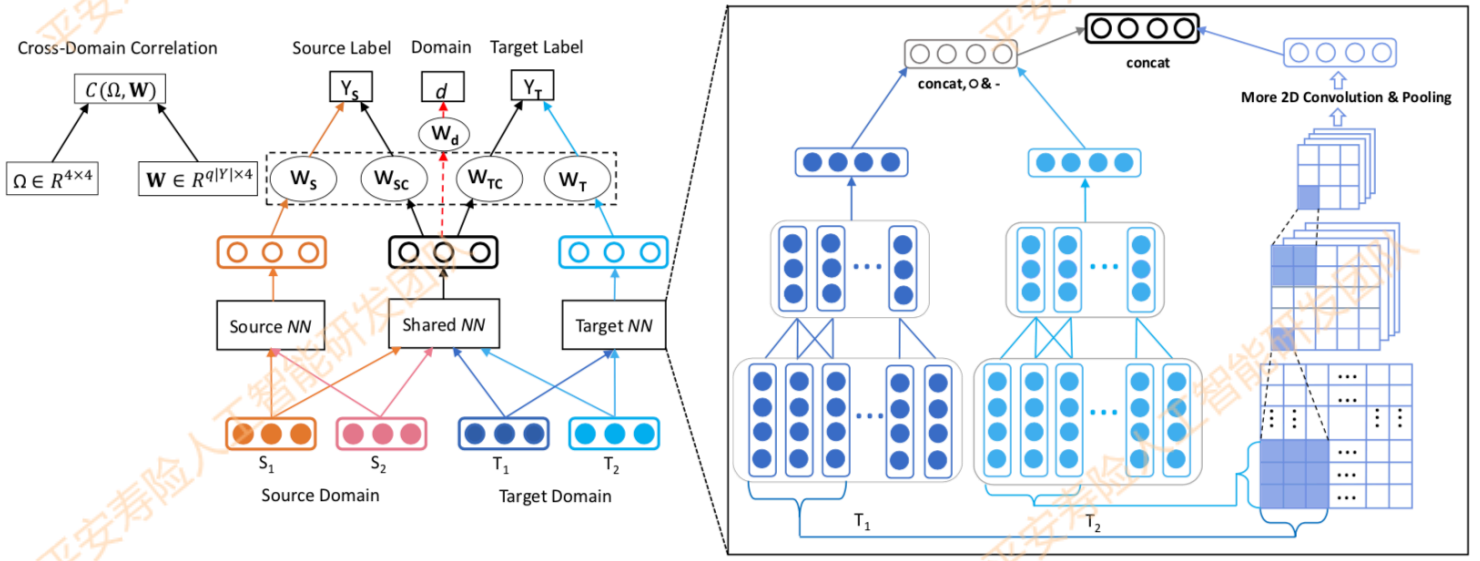
……

附录3：*迁移学习*

中国平安人寿保险

- sentence representation：考虑响应时间的需求，采用了cnn-based method，采用 Representation model和Interaction model相结合的方式，对表示模型做了些更改，并将最后的embedding 进行 concatenate；
- 在传统迁移学习的框架上，引入了半正定协方差矩阵，对输出层的域内以及域间信息权重进行建模；
- 引入对抗损失，增强 shared 层的抗噪能力。

Modelling Domain Relationships for Transfer Learning on Retrieval-based Question Answering Systems in E-commerce，2018，Alibaba. 迁移学习在问答系统上的创新，已在AliExpress上线.准确率比state of art模型略低但比单纯的表示模型高，QPS高，支持大规模线上系统。



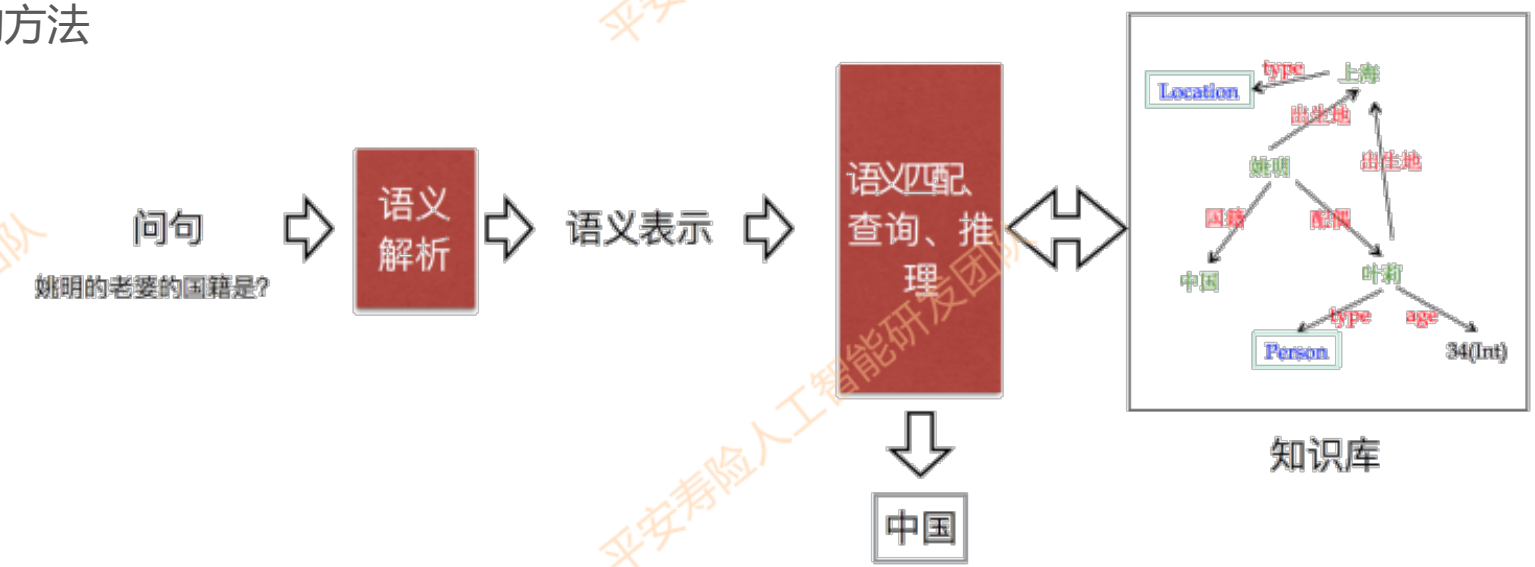
附录4：*基于知识图谱的问答——KBQA*

中国平安人寿保险

解决事实类问题和简单的推理

基于知识图谱的问答模块，需要解决两个核心问题：（1）如何理解问题语义，并用计算机可以接受的形式进行表示（问题的理解和表示）；（2）以及如何将该问题表示关联到知识图谱的结构化查询中（语义关联）。

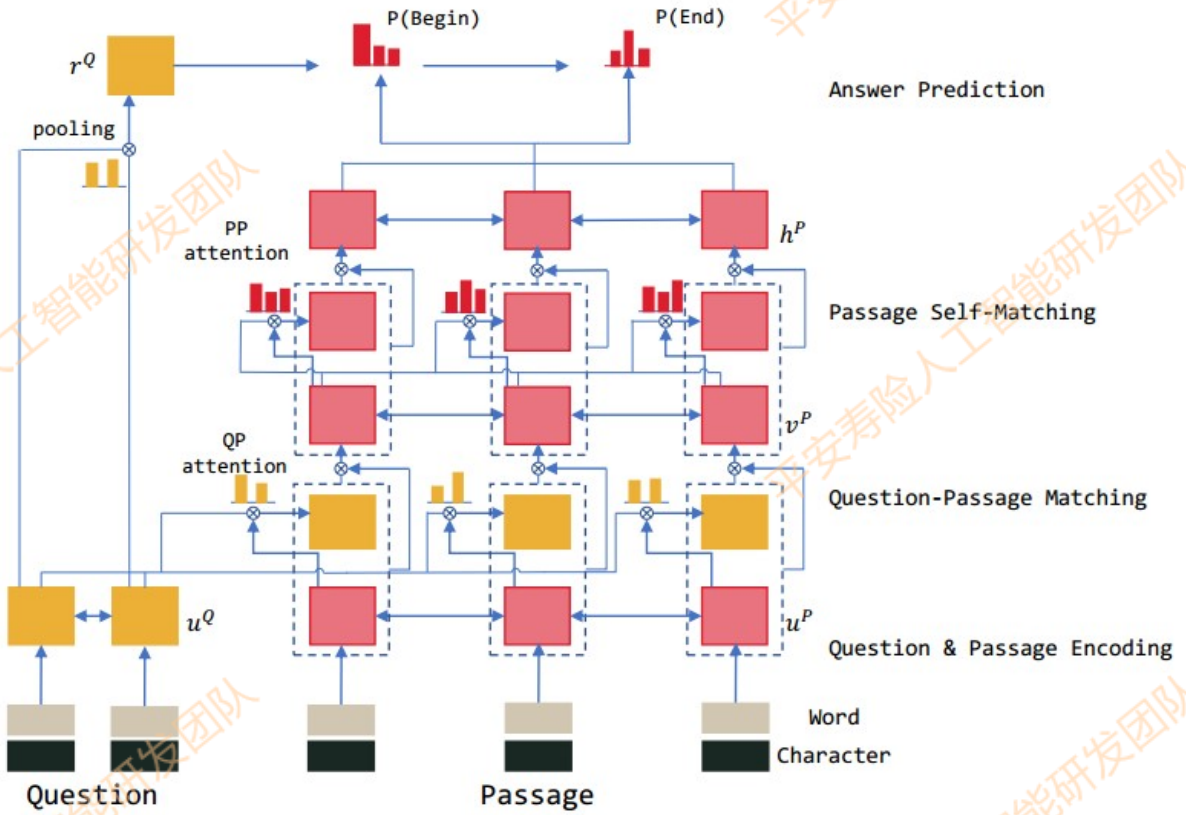
- 基于模版的方法：自然语言查询-->意图识别(Intention Recognition)-->实体链指(Entity Linking)+关系识别(Relation Detection) -->查询语句拼装(Query Construction)-->返回结果选择(Answering Selection)
- 基于语义解析的方法
- 基于神经网络的方法



附录5：*阅读理解式问答*

中国平安人寿保险

- 基于阅读理解的问答:适用数据类型为(**给定一个问题Q和一个与Q相关的文档D,自动得到Q对应的答案A**)非结构化文本,主要的方法有匹配式,抽取式和生成式
- 匹配式:给出文章,问题和答案集,从答案集中选出最高得分的答案,像选择题. 例如Attentive-reader, Impatient-reader
- 抽取式:顾名思义就是从文档中抽取出答案,前提是文档中包括问题答案. 抽取式的一般框架是,Embedder+Encoder+Interaction-layer+Answer. 主要模型有:Match-LSTM,R-NET,BiDAF
- 生成式:答案形式是这样的：1)答案完全在某篇原文.2)答案分别出现在多篇文章中.3)答案一部分出现在原文，一部分出现在问题中.4)答案的一部分出现在原文，另一部分是生成的新词.5)答案完全不在原文出现（Yes / No 类型). 常见模型：改进的R-Net,S-NET,R3-NET



Proximate State of Art: R-NET structure

附录6：*Learning to Rank*

中国平安人寿保险

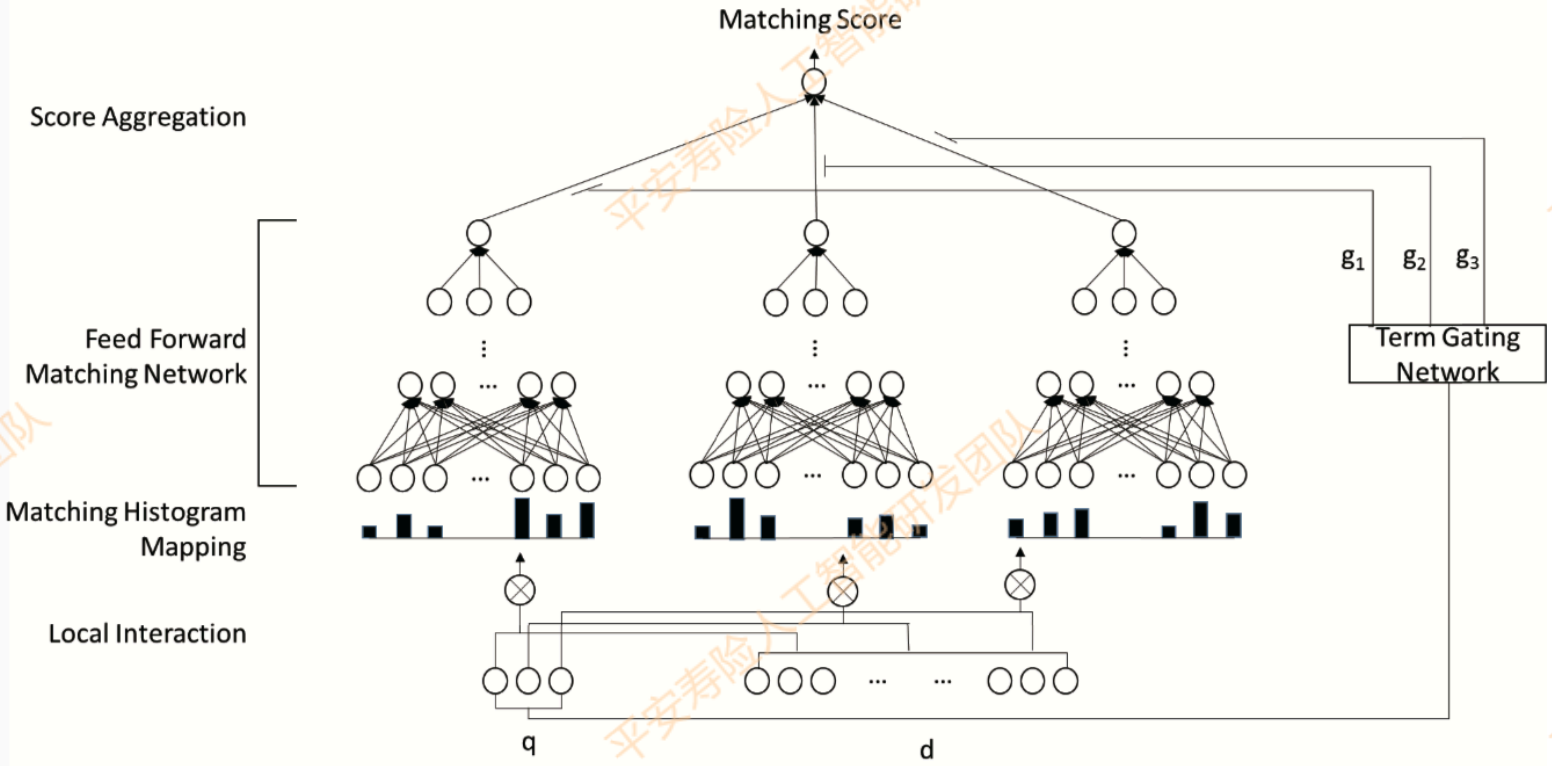
- **用机器学习的方法去排序**

类别	输入数据	样本复杂度	所转化的主要问题	代表算法	特点
pointwise	单个文档	O(n)	分类、回归	Prank,RankProp,OAP-BPM	考虑单个文档之间的排序特征, 不考虑同一查询下文档间的关 系信息, 偏离了排序问题的实质. 模型较简单, 训练时间较短.
pairwise	具有偏序关 系的文档对	O(n2)	二分类问题	Ranking SVM,RankNet,RankBoost	考虑文档对之间的偏序关系, 部分保留了同一查询下文档间的 关系信息, 并不考虑文档在文档列表上的位置, 接近排序问题 的实质. 模型较复杂, 训练时间较长, 需较高效的学习算法.
listwise	所有相关联 的整个文档 列表	O(n!)	最优化问题等	LambdaRank, ListNet, AdaRank	考虑同一查询下不同文档的序列关系, 完全符合排序问题的实质. 模型的复杂度以及训练时间的长短很大程度上依赖于文 档列表的损失函数或优化目标的定义.

附录7：*DeepRank*

DRMM：Deep relevance matching model. Relevance matching is different from semantic matching!

1. Matching Histogram Mapping
2. Feed forward Matching Network
3. Term Gating Network



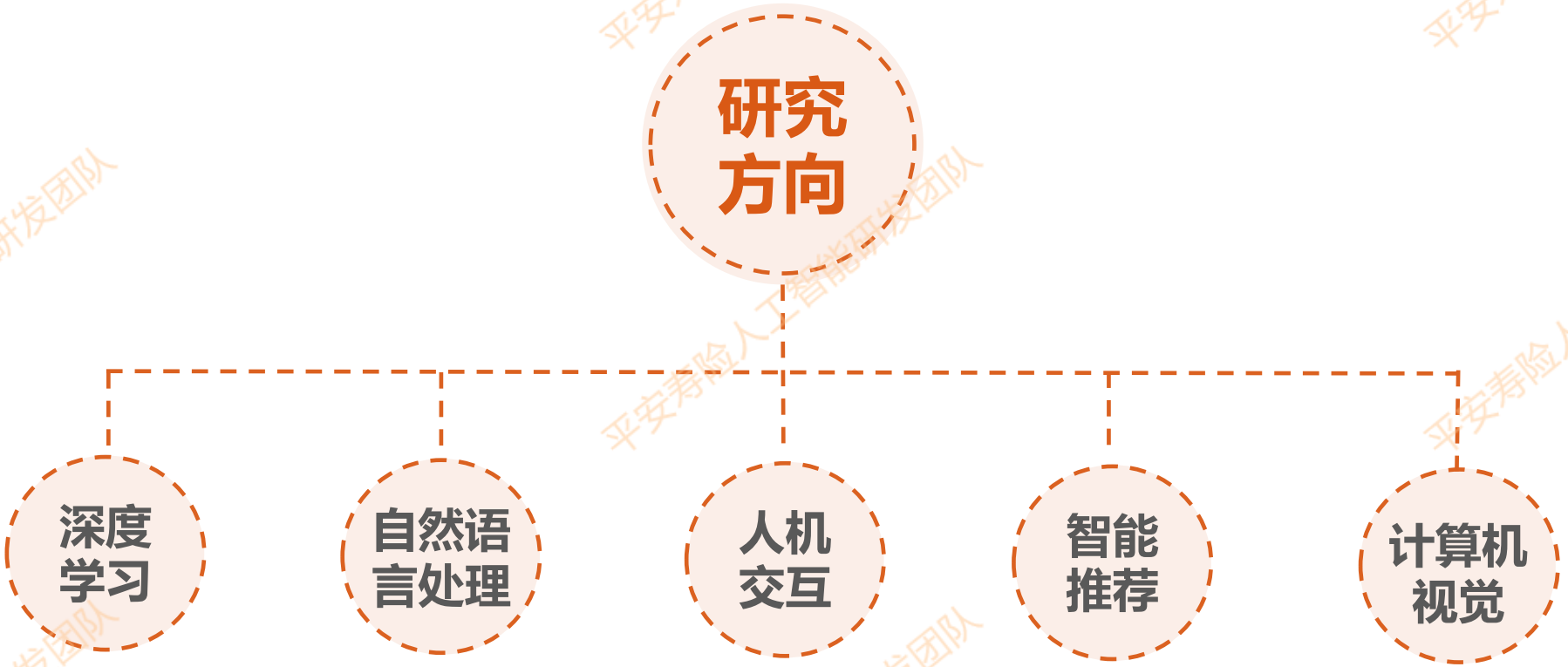
开放，共赢

优秀的问答系统有两个关键点：

- 精确的问题理解，高质量的知识来源。

平安人寿智能平台团队

中国平安人寿保险



应用场景


招聘


培训


销售


服务


风控

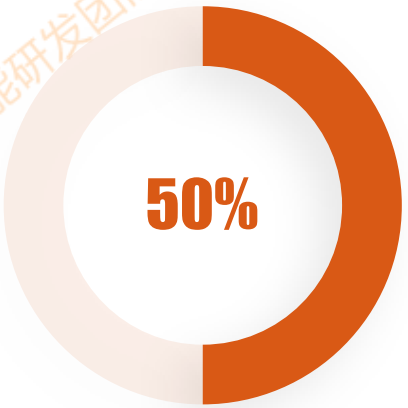
平安人寿智能平台团队

中国平安人寿保险

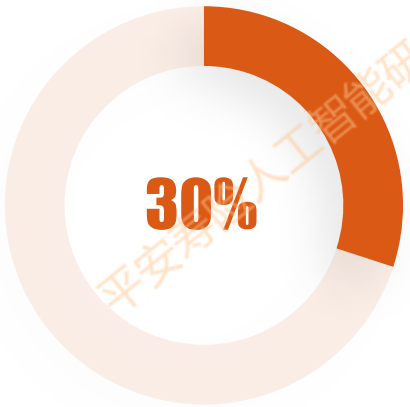
团队成员

硕士及以上

博士及以上



BATJ

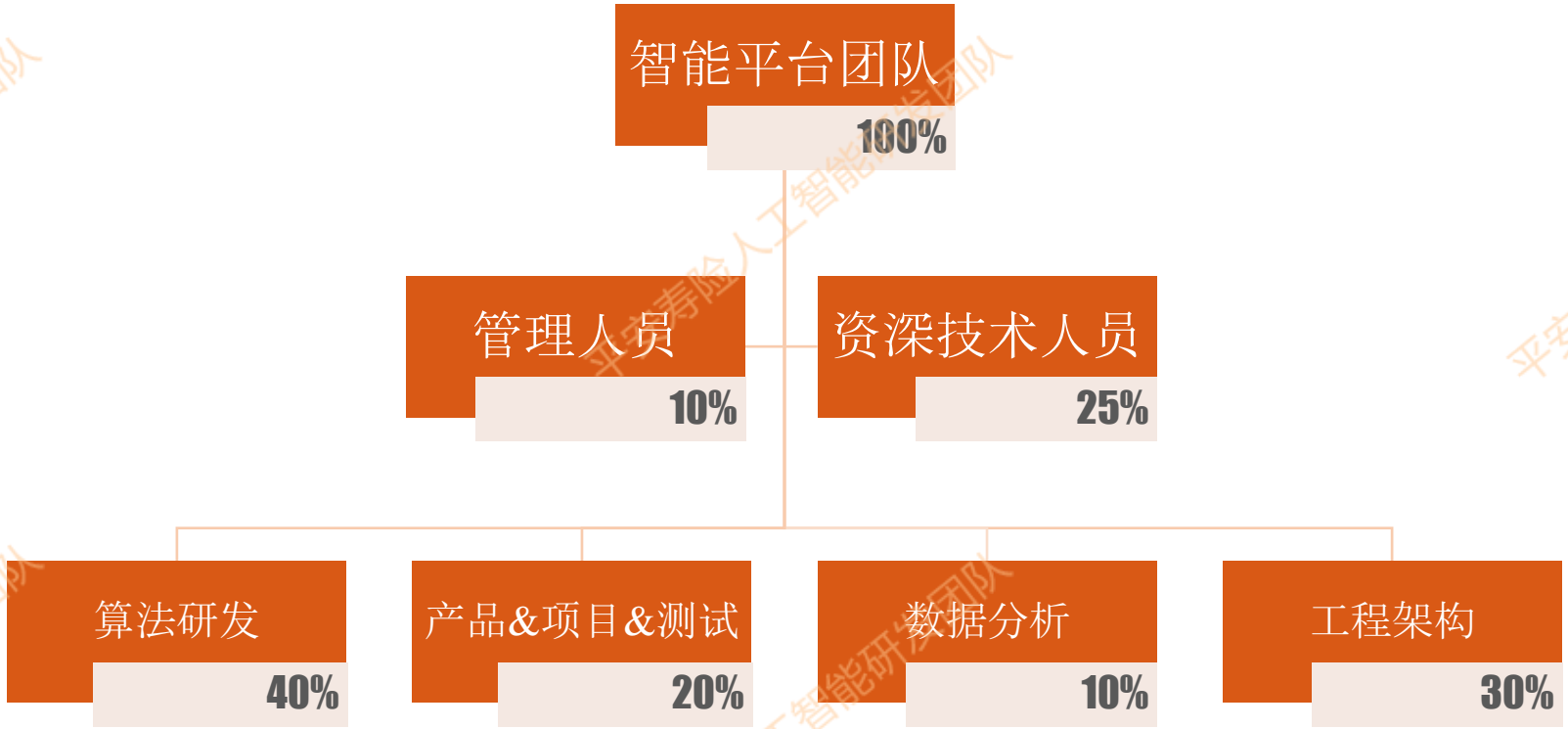


境外背景

平安人寿智能平台团队

中国平安人寿保险

团队架构



平安人寿智能平台团队

中国平安人寿保险

已有成果

1. 首个大规模应用的智能面试机器人，节省大量人力成本
2. 首个可实现业务办理的智能客服机器人，全程办理保单贷款等复杂业务，为客户带来“服务+推荐“一站式体验
3. 首个针对业务员的智能助理机器人，为业务员提供全方位、多元化、智能化服务
4. 行业领先的分布式深度学习平台，拥有超大集群规模和计算能力，大幅提升研发效率
5. 行业领先的人机交互应用平台，为各个高价值人机交互应用落地提供有力的底层支持

more. 智能保顾、个性化培训、千人千面推荐、核保理赔、无人门店、各类新技术应用....

Thank You for Listening

谢谢您的聆听