基于深度学习的FAQ问答系统

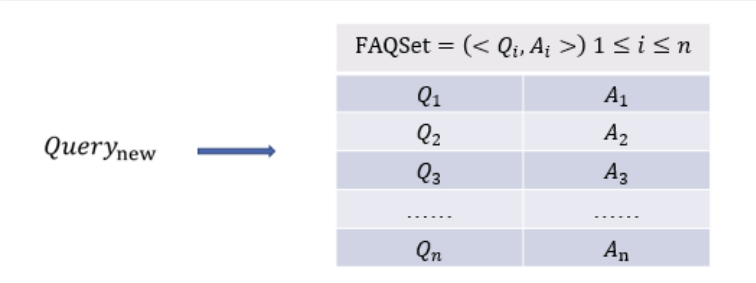
腾讯知文实验室 发表于 语言、知识与人工智能

导语 问答系统是信息检索的一种高级形式，能够更加准确地理解用户用自然语言提出的问题，并通过检索语料库、知识图谱或问答知识库返回简洁、准确的匹配答案。相较于搜索引擎，问答系统能更好地理解用户提问的真实意图，进一步能更有效地满足用户的信息需求。问答系统是目前人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向。

一、引言

问答系统处理的对象主要包括用户的问题以及答案。根据问题所属的知识领域，问答系统可分为面向限定域的问答系统、面向开放域的问答系统、以及面向常用问题集（Frequently Asked Questions, FAQ）的问答系统。依据答案来源，问答系统可分为基于结构化数据的问答系统如KBQA、基于文本的问答系统如机器阅读理解、以及基于问答对的问答系统如FAQ问答。此外，按照答案的反馈机制划分，问答系统还可以分为基于检索式的问答系统和基于生成式的问答系统。

本文主要阐述FAQBot检索型问答系统的相关研究和处理框架，以及深度学习在其中的应用。FAQ检索型问答是根据用户的新Query去FAQ知识库找到最合适的答案并反馈给用户。如图所示：



其中，Qi是知识库里的标准问，Ai是标准问对应的答案。

具体处理流程为：

- 候选集离线建好索引。采用Lucene引擎，为数万个相似问集合建立字级别倒排索引。Lucene引擎的性能能够将召回时间控制在毫秒级别，大大减轻后续模块的计算压力；
- 线上收到用户 query 后，初步召回一批候选集作为粗排结果传入下一模块进行进一步精确排序；
- 利用matching模型计算用户query和FAQ知识库中问题或答案的匹配程度；
- 利用ranking 模型对候选集做 rerank 并返回 topk个候选答案。

可以看出，FAQ问答系统的核心任务可以抽象为文本匹配任务。传统文本匹配方法如信息检索中的BM25，向量空间模型VSM等方法，主要解决字面相似度问题。然而由于中文含义的丰富性，通常很难直接根据关键字匹配或者基于机器学习的浅层模型来确定两个句子之间的语义相似度。近几年，利用神经网络，尤其是深度学习模型学习文本中深层的语义特征，对文本做语义表示后进行语义匹配的方法开始被提出并应用于检索式问答系统。基于深度学习的模型一方面能够节省人工提取特征的大量人力物力。此外，相比于传统方法，深度文本匹配模型能够从大量的样本中自动提取出词语之间的关系，并能结合短语匹配中的结构信息和文本匹配的层次化特性，发掘传统模型很难发掘的隐含在大量数据中含义不明显的特征，更精细地描述文本匹配问题。

FAQ问答系统一般有两种解决思路，一种是相似问题匹配，即对比用户问题与现有FAQ知识库中问题的相似度，返回用户问题对应的最准确的答案，这种思路类似于text paraphrase；另一种是问题答案对匹配，即对比用户问题与FAQ知识库中答案的匹配度，返回用户问题对应的最准确的答案，这种思路为答案选择，即QA匹配。这两个类型相通的地方在于都可以看作文本语义匹配，很多模型能同时在两个任务上都得到很好的效果，区别在于QA匹配存在问题与答案不同质的问题。

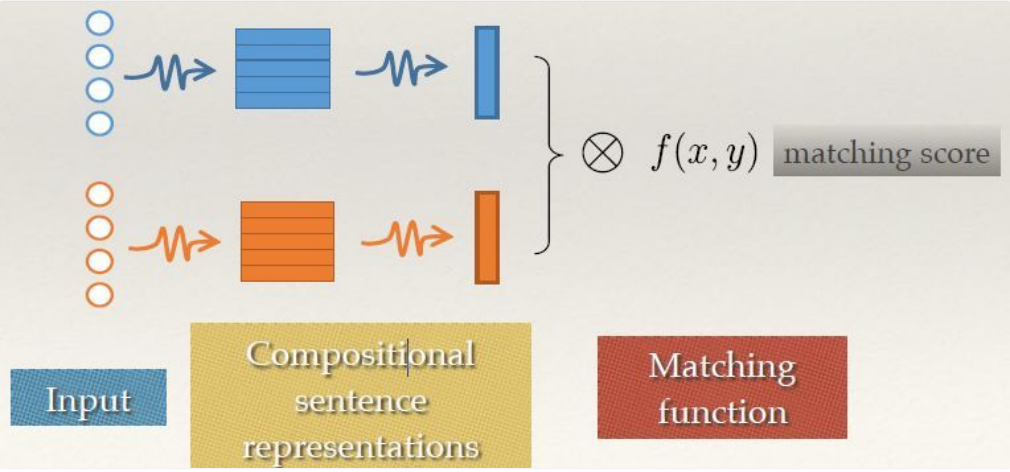
下面总结一些基于深度学习的文本匹配工作，希望能够抛砖引玉，如有遗漏或错误，欢迎补充或指出。

2.1 模型框架

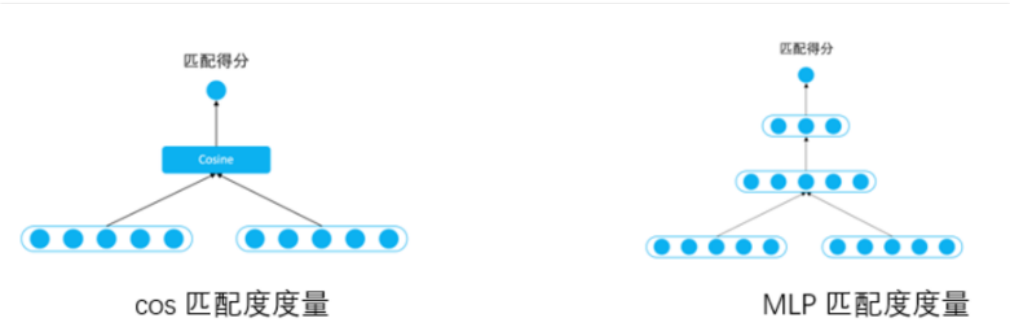
概括来讲，深度语义匹配模型可以分为两大类，分别是representation-based method 和 interaction-based method。

1) Represention-based Method

框架图如下：

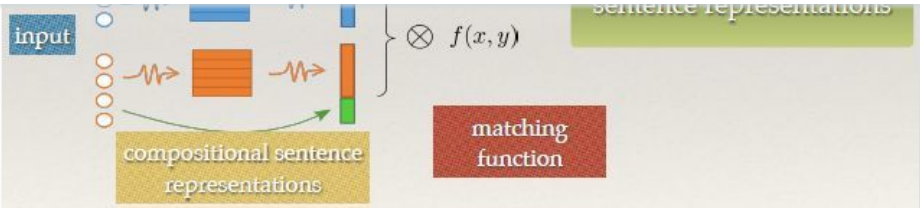


这类算法首先将待匹配的两个对象通过深度学习模型进行表示，之后计算这两个表示之间的相似度便可输出两个对象的匹配度。这种方式下，更加侧重对表示层的构建，使其尽可能充分地将待匹配的两个对象都转换成等长的语义表示向量。然后在两个对象对应的两个语义表示向量基础上，进行匹配度的计算。针对匹配度函数f(x,y)的计算，通常有两种方法，如下图所示：一种是通过相似度度量函数进行计算，实际使用过程中最常用的就是 cosine 函数，这种方式简单高效，并且得分区间可控意义明确；另一种方法是将两个向量再接一个多层感知器网络（MLP），通过数据去训练拟合出一个匹配度得分，更加灵活拟合能力更强，但对训练的要求也更高。



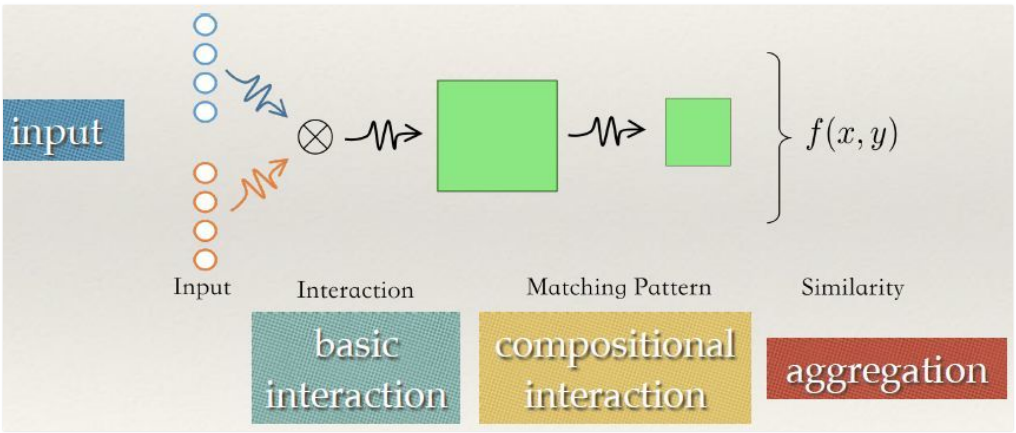
Represention-based Extended

上述的representation-based method存在的问题是直接基于句子的表示太粗糙，无法准确进行文本匹配任务。受信息检索领域的启发，结合主题级别和单词级别的匹配信息通常可以获得更好的表现。于是进一步对句子表示进行扩展，加入细粒度匹配信息。框架图如下：



2) Interaction-based Method

框架图如下：



基于交互的方法是通过Interaction来对文本相似性建模。该方式更强调待匹配的两个句子得到更充分的交互，以及交互后的匹配。在表示层不会将句子转换成一个整体表示向量，一般情况下会保留和词位置相对应的一组表示向量。首先基于表示层采用DNN或直接由word embedding得到的句子表示，和词位置对应的每个向量体现了以本词语为核心的一定的全局信息；然后对两个句子按词对应交互，由此构建两段文本之间的 matching pattern，这里面包括了更细致更局部的文本交互信息；基于该匹配矩阵，可以进一步使用DNN等来提取更高层次的匹配特征，最后计算得到最终匹配得分。Interaction-based 方法匹配建模更加细致、充分，一般来说效果更好，但计算成本增加，更加适合一些效果精度要求高但对计算性能要求不高的场景。

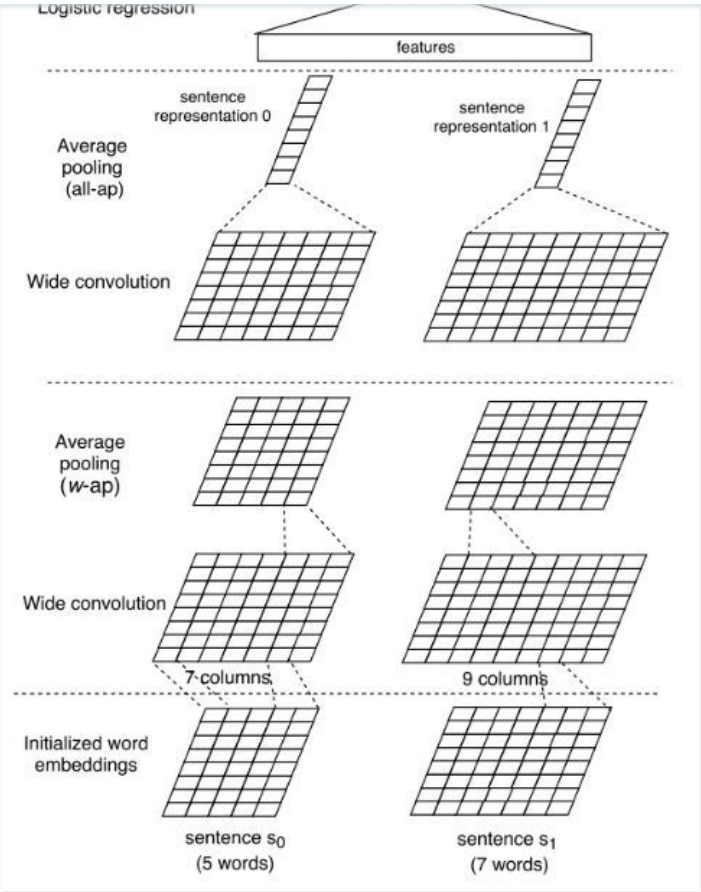
下面总结了不同类型的深度学习文本匹配模型。可以看出，深度文本匹配现有工作很多，本文将对近几年的部分工作进行详细介绍，其他可参考对应文献进行深入阅读。

- representation-based:DSSM[1]; CDSSM[2]; ARC I[3]; CNTN[4]; LSTM-RNN[5]
- representation-based extension:MultiGranCNN[6]; MV-LSTM[7]
- interaction-based:ARC II[8]; MatchPyramid[9]; Match-SRNN[10]; DeepMatch[11]; ABCNN[12]; QA-LSTM/CNN-attention[13,14]; AP[15]; AICNN[16]; MVFNN[17]; BiMPM[18]; DQI[22]; DIIN[23]

2.2 模型介绍

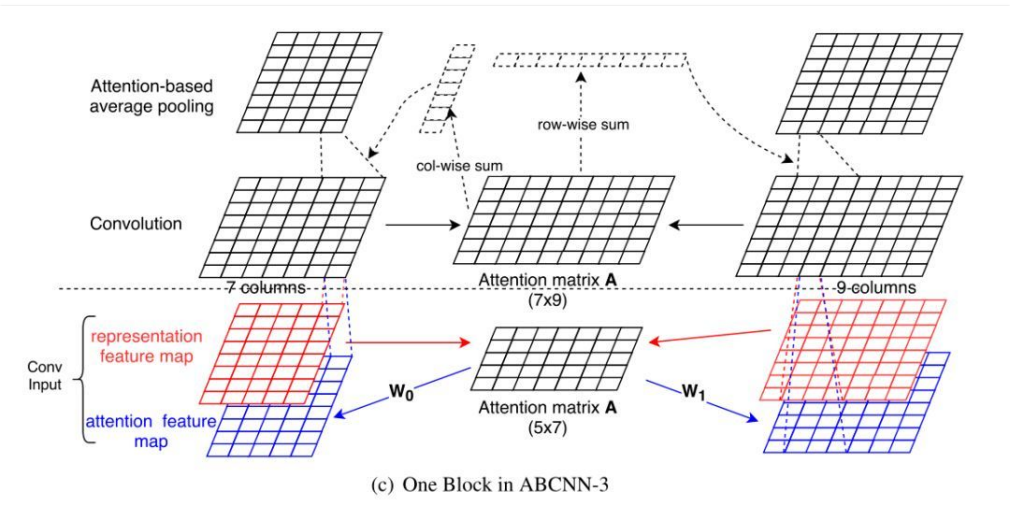
2.2.1 ABCNN[12]

首先介绍BCNN，它是ABCNN模型的基础，即未添加Attention的模型。模型结构如图所示：



输入层：将输入句子进行padding后转化成词向量即可；
卷积层：对句子表示进行卷积，使用wide conv的方式；
pooling层：论文中使用了两种pooling方式，一种是最后一个pooling层为all-ap，还有一种是中间pooling层为w-ap。区别就是池化时的窗口大小不同；
输出层：接logistic 回归层做2分类。

ABCNN是在BCNN的基础上加了两种attention机制。模型结果如下图：



(c) One Block in ABCNN-3

Figure 3: Three ABCNN architectures

(1)在输入层加入attention

其原理为将输入拓展成双通道。新添加的通道是attention feature map，即上图中的蓝色部分。首先计算attention matrix A，其每个元素Aij代表句子1中第i个单词对句子二中第j个单词的match_score，这里使用了Euclidean距离计算。然后再分别计算两个句子的attention feature map。使用两个矩阵W0,W1分别和A还有 A的转置相乘，得到与原本feature尺寸相同的feature

(2)在pooling层加入attention

Attention matrix A的计算方法与上述相同，得到A后需要为两个句子分别计算attention权重向量，如上图中的两个虚线部分col-wise sum和row-wise sum。这两个向量中的每个元素分别代表了相应单词在做Average Pooling时的权重。相当于pooling不再是简单的Average Pooling，而是根据计算出的Attention权重向量得到的pooling。

2.2.2LSTM/CNN,attention[13,14]

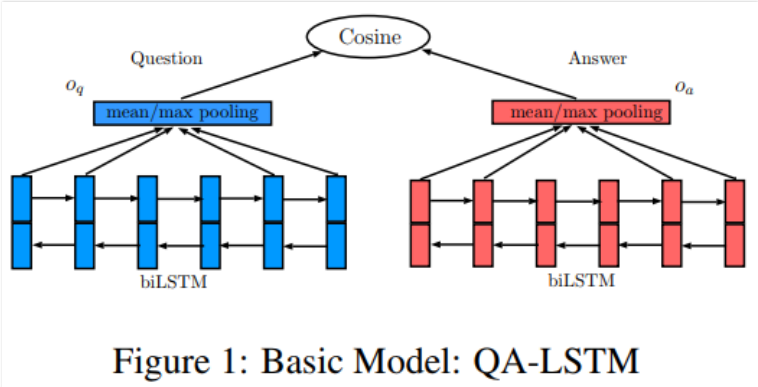


Figure 1: Basic Model: QA-LSTM

给定一个(q,a)pair，q是问题，a是候选答案。首先得到它们的词向量，再使用biLSTM进行encoder，生成问题和答案的分布式表示，然后利用余弦相似度来衡量它们的距离。训练目标是hinge loss。

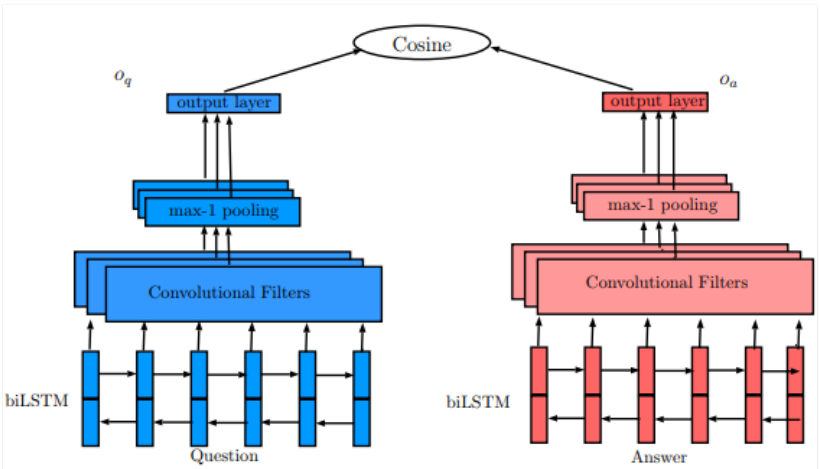


Figure 2: QA-LSTM/CNN

在biLSTM表示输出的基础上进一步使用CNN，CNN可以获取biLSTM输出的向量之间的局部信息。从而给出问题和答案的更多复合表示。

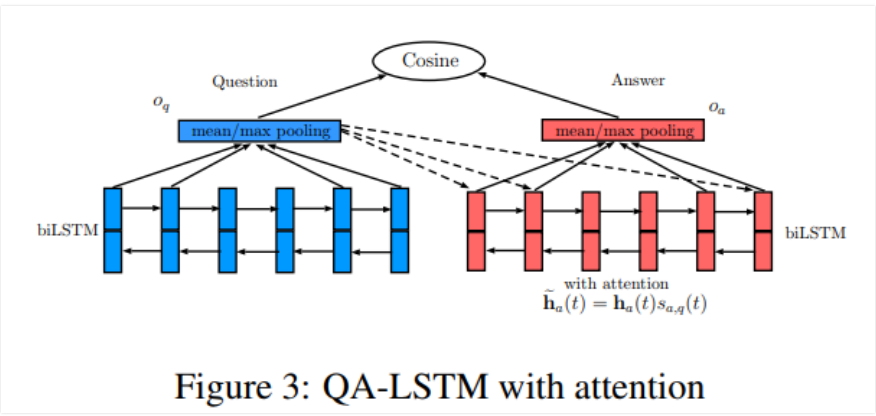
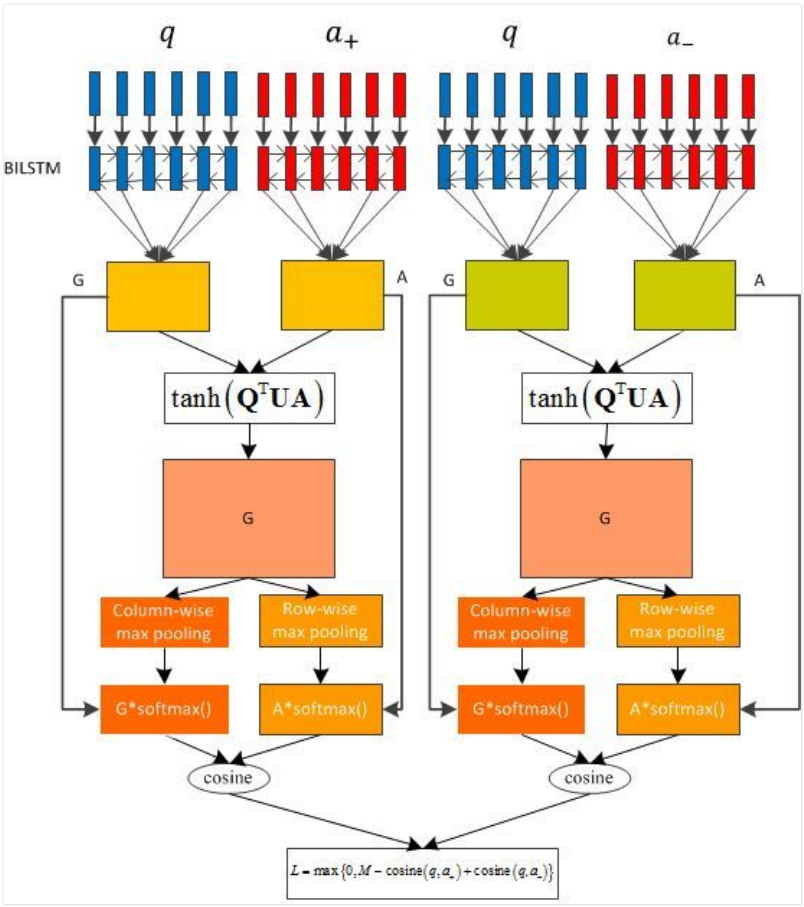


Figure 3: QA-LSTM with attention

biLSTM输出向量将乘以softmax权重，该权重由biLSTM的问题嵌入得到。

2.2.3 Attentive Pooling Networks[15]

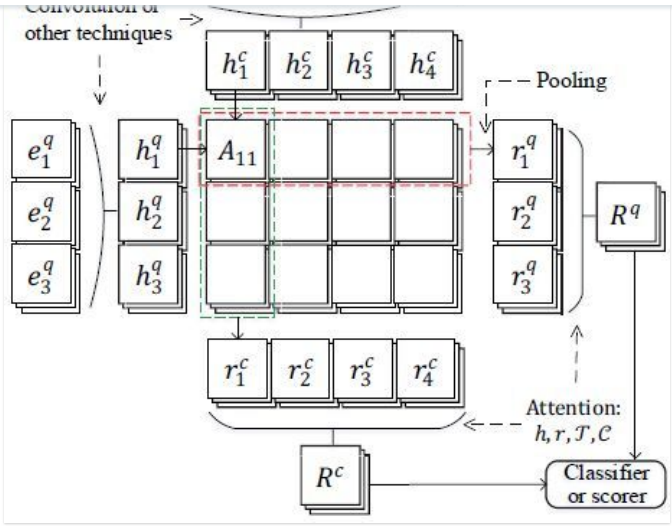
QA_LSTM with attention中attention的设计是通过问题对答案的影响进行特征加权，但是它忽略了答案对问题的影响。Attentive pooling networks同时将attention应用到问题和答案，提高算法的准确率。通过同时学习两种输入的表示以及它们之间的相似性测量，其创新点在于将Q和A这两个输入通过参数矩阵U投射到一个共同的表示空间，用Q和A的representation构造了一个矩阵G，分别对G的row和column做max pooling，这样就能分别能得到Q和A的attention vector。AP_BILSTM模型框架图如下：



AP_BILSTM模型的设计首先将问题和答案经过BILSTM抽取特征，然后通过两者的特征计算soft alignment，得到的G矩阵表示了问题和答案相互作用的结果。对该矩阵的列取最大，即为答案对问题的重要性得分，同理对该矩阵行取最大即为问题对答案的重要性得分。这两个向量再作为attention向量分别和问题和答案表示相乘后得到问题和答案新的表示，最后再做匹配。

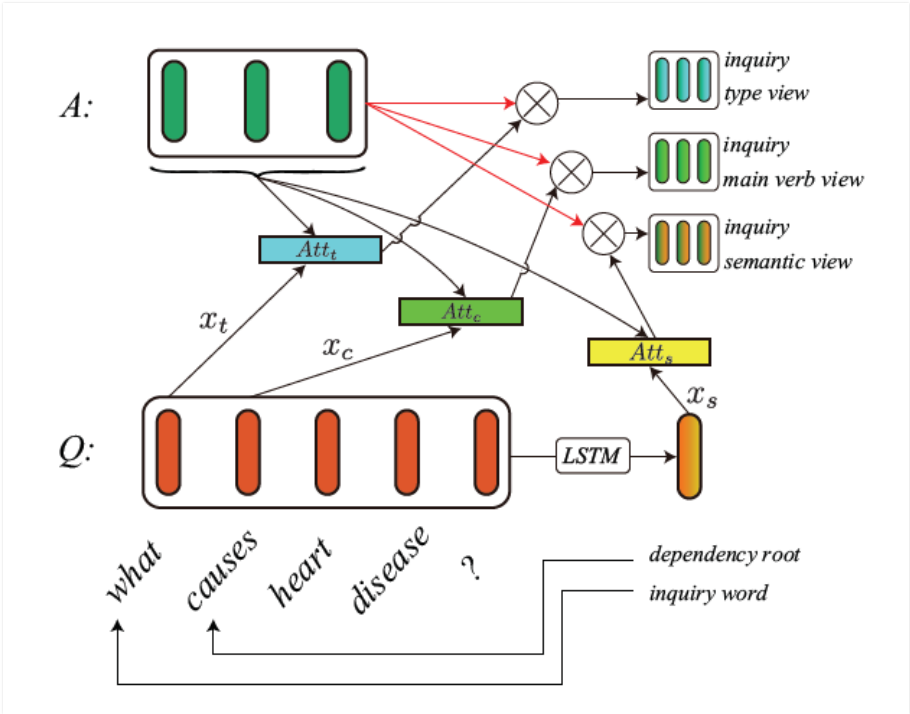
2.2.4 AICNN[16]

之前关于答案选择的研究通常忽略了数据中普遍存在的冗余和噪声问题。 在本文中，设计一种新颖的注意力交互式神经网络（AI-NN），以专注于那些有助于回答选择的文本片段。 问题答案的表示首先通过卷积神经网络（CNN）或其他神经网络架构来学习。然后AI-NN学习两个文本的每个配对片段的相互作用。 之后使用逐行和逐列池化来收集交互信息。之后采用注意机制来衡量每个细分的重要性，并结合相互作用来获得问答的固定长度表示。 模型框架图如下：



2.2.5 MVFNN[17]

上述基于神经网络的方法通过计算注意力来考虑信息的几个不同方面。 这些不同类型的注意力总是简单地总结并且可以被视为“单一视图”，不能从多个方面来审视问题和候选答案，导致严重的信息丢失。 要克服这个问题，此模型提出了一种多视图融合神经网络，其中每个关注组件生成QA对的不同“视图”，并且融合QA本身的特征表示以形成更整体的表示。模型框架图如下：



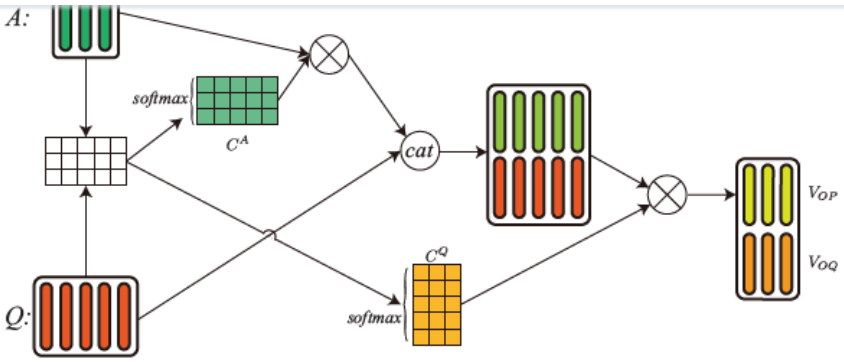
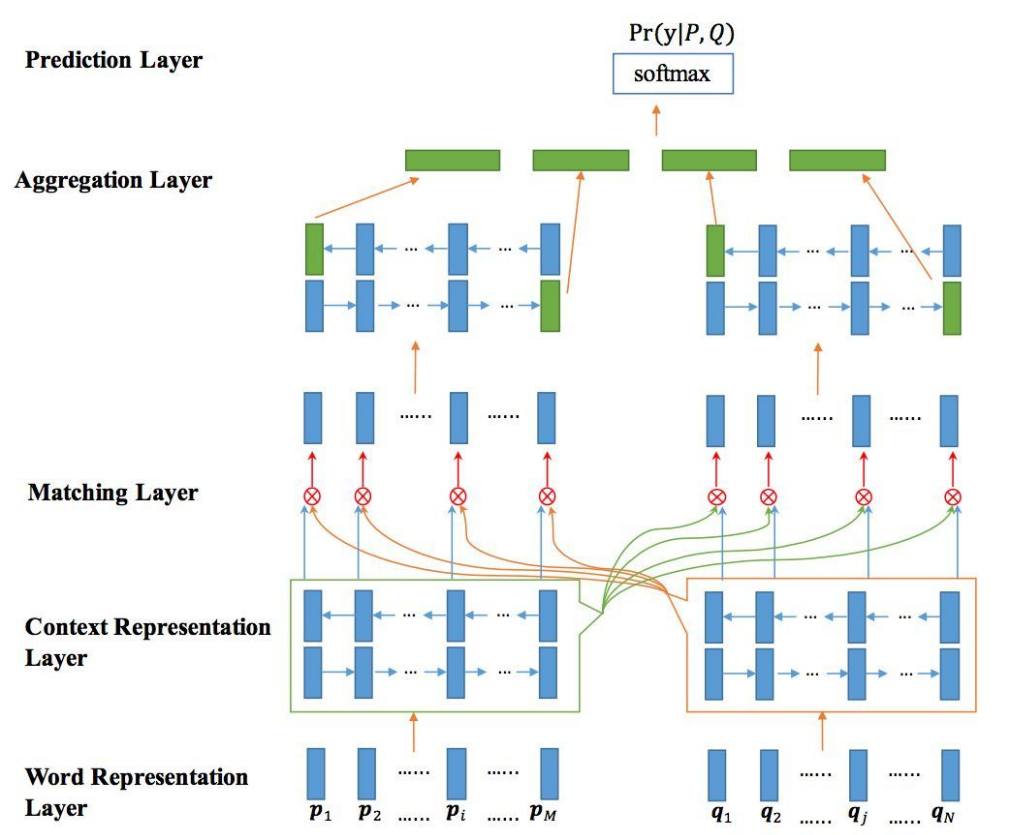


Figure 3: The generating process of co-attention view.

对于一个问题，可能会有一堆视图来模拟其相应的答案。 在此模型中，根据直觉构建了四个视图。这四个视图被命名为查询类型视图，查询主动词视图，查询语义视图和co-attention视图。最后使用fusion RNN模型来对这些视图进行融合。通过不同视图的融合，能对两个对象进行更准确的建模。

2.2.6 BiMPM[18]

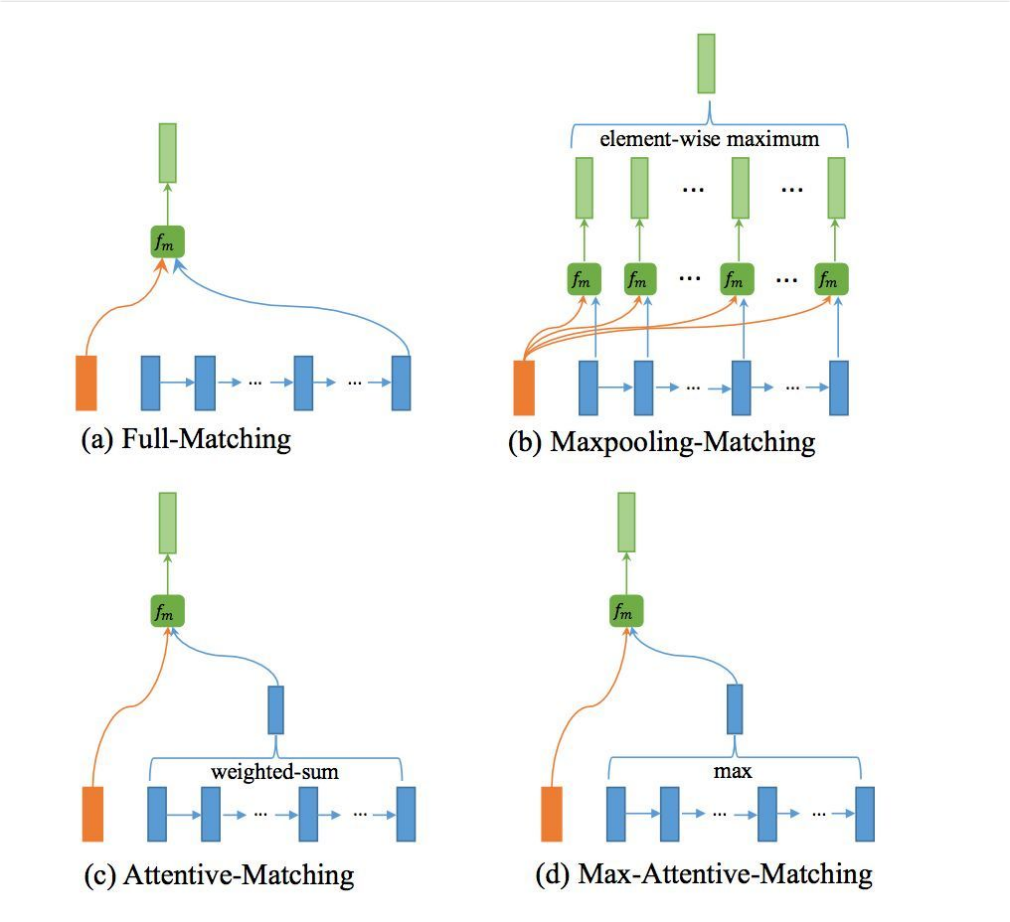
针对基于交互这一类方法，一般是先对两个句子的单元相互匹配，之后再聚集为一个向量后做匹配。这种方式可以捕捉到两个句子之间的交互特征，但是之前的方式只是基于词级别的匹配但是忽略了其他层级的信息。此外，匹配只是基于一个方向忽略了相反的方向。一种双向的多角度匹配模型bilateral multi-perspective matching（BiMPM)解决了这方面的不足。模型框架如下图：



模型自下而上一共包含五层，分别为单词表示层、上下文表示层、匹配层、聚合层和预测层，其中匹配层为模型的核心，共提出了四种匹配策略，这里的匹配可以看成是attention机制。

单词表示层：使用GloVe模型训练向量，对字符embedding进行随机初始化，单词中的字符组成单词的向量表示作为LSTM网络的输入。

匹配层：模型的核心层，包含四种匹配策略，分别是：Full-Matching、Maxpooling-Matching、Attentive-Matching和 Max-Attentive-Matching。四种匹配策略如下图：



聚合层：利用BiLSTM对匹配层的输出向量进行处理，取得p、q前向和后向最后一个time step的输出进行连接后输入到预测层。

预测层：softmax层，softmax函数分类。

上述是对近几年部分深度文本匹配模型的总结，接下来则介绍基于深度模型的FAQBot。

三、基于深度学习的FAQBot实现

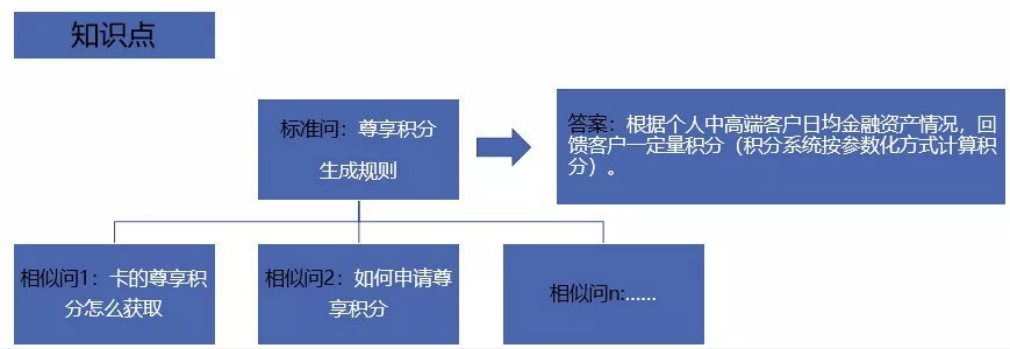
3.1 模型化流程

3.2 数据获取及构造

3.2.1 数据获取

对于有大量问答记录的场景例如智能客服，这些记录里面有很多高频的知识点(知识点包括问题和答案)。这些高频的知识点对应的问法通常并不唯一。即知识库的结构为一个问题集合对应同一个答案。针对FAQ数据有以下三种数据类型：

1. 标准问q：FAQ中问题的标准用户query
2. 答案A： FAQ中标准问对应的的标准回答



其中，标准问q、对应答案A以及该标准问q对应的所有相似问q1,q2,...，一起组成一个知识点。一个知识点的样例见下图：

3.2.2 数据构造

数据构造包含了两个方面：

（1）训练集测试集构造

测试集：将相似问中的第一条相似问q1作为query，从FAQ知识库的所有知识点中通过Lucene召回30个知识点作为候选集

训练集：包含两部分，一部分是正例的构造，另一部分是负例的构造，这两部分数据的构造方式将直接影响到最终的效果。在正例的构造中，因为每个知识点的第一个相似问是作为测试集中出现的，所以在构造训练集的时候排除掉所有知识点中的第一条相似问q1。这样的话，有多于2个相似问的知识点还有多于的其他相似问可以用来构造训练集。将这些识点中的标准问和从相似问的第二条开始（即[q2,q3,...,qn]）可以按照不同方式构造出正例和负例。

训练集正例的构造：去除所有知识点中的第一条相似问q1，其他相似问及标准问两两组合成正例pair对；对于相似问多的知识点进行剪切。

训练集负例的构造的方式包括：

- 按Jaccard距离召回；
- 按Lucene召回；
- 从其他知识点中随机选择；
- 按照正例中各问题出现的比例从其他知识点中采样选择；
- 每个句子和句子中的名词/动词构成pair对；
- 针对知识点分布不均衡的问题，对相似问很多的知识点进行相似问剪切。

（2）数据增强策略

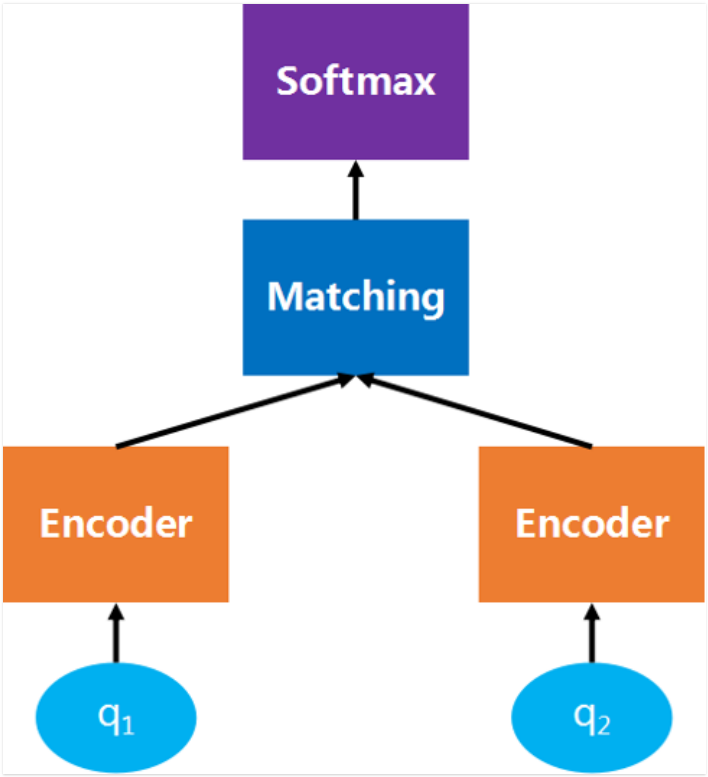
由于深度学习需要较多的数据，为了增强数据，我们采用了以下策略：

- 交换两个句子之间的顺序；
- 对句子进行分词，重新组合生成新的句子；
- 打乱句子的顺序，随机抽取句子。

3.3 模型建立

3.3.1 模型框架

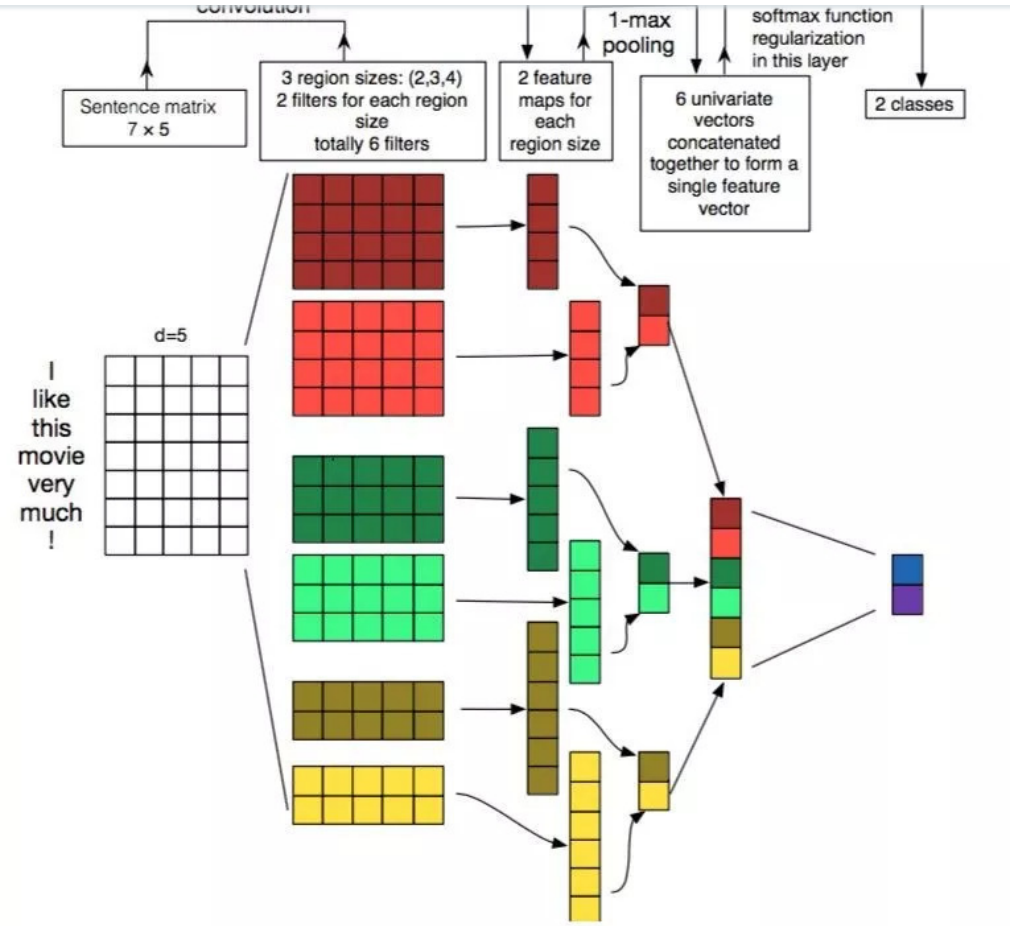
的传统文本特征，将所有这些特征进行concat。最后接上softmax层，做最终的分类。模型的框架如下图所示：



3.3.2 模型建立及迭代优化

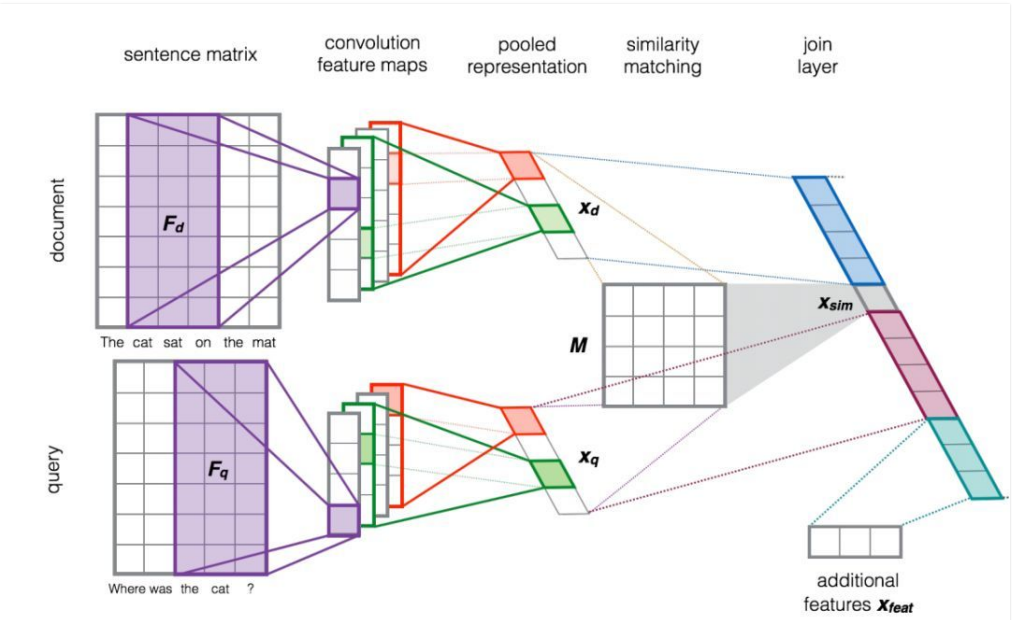
Embedding层：使用word2vec和fasttext训练词向量和字符向量。

Encoder层：卷积具有局部特征提取的功能，所以可用 CNN 来提取句子中类似 n-gram 的关键信息，考虑文本的上下文信息。于是我们采用textCNN[19]来对句子进行编码表示，encoder过程见下图：

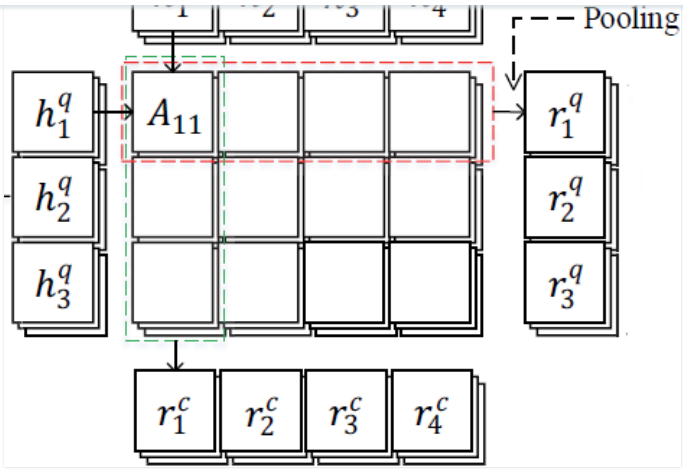


Matching层：在得到两个句子的表示后，要针对两个句子的表示进行matching操作。可以根据需要构造出很多种类型的matching方式如下图[20]，我们采用相对比较简单的element-wise相加和相乘的方式来进行matching。

join层：在matching层之后得到的两个句子的共同表示之后，进一步引入额外的传统特征进行join操作，类似于下图[21]。



引入interaction：上述步骤对两个句子encoder时没有考虑两个句子之间的关联。于是进一步引入更细致更局部的句子交互信息，从而能捕捉到两个句子之间的交互特征，根据交互得到的矩阵获取两个句子新的表示。如图：



引入attention机制：采用注意机制使用权重向量来衡量句子不同部分重要性的不同。attention的计算主要思想沿用了AICNN和ABCNN中的几种attention，分别是feature的attention，interaction后新的表示和句子原表示之间的attention。

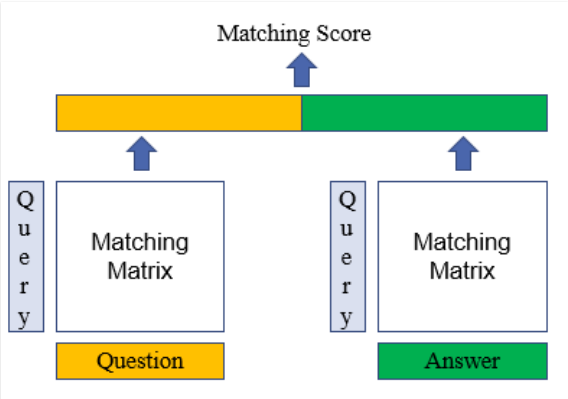
四、总结与展望

4.1 数据层面

- 建立更加合理的知识库：每个知识点只包含一个意图，且知识点之间没有交叉，歧义，冗余等容易造成混淆的因素
- 标注：为每个FAQ积累一定数量的有代表性的相似问
- 后期的持续维护：包括新FAQ发现，原FAQ的合并、拆分、纠正等

4.2 模型层面

- 进一步捕捉syntactic level和semantic level的知识如语义角色标注（SRL, semantic role labelling）和词性标注（POS, part of speech tagging）等，引入到文本的表示之中，提高文本语义匹配的效果
- 目前大部分检索行问答的工作做的是问题和问题匹配，或是问题和答案匹配。后续可以同时引入问题和答案的信息进行建模，如图：



参考文献

[1] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]// ACM International Conference on Conference on Information &

[2] Shen Y, He X, Gao J, et al. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval[C]// Acm International Conference on Conference on Information & Knowledge Management. ACM, 2014:101-110.

[3] Hu B, Lu Z, Li H, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[J]. Advances in Neural Information Processing Systems, 2015, 3:2042-2050.

[4] Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:1305-1311.

[5] Palangi H, Deng L, Shen Y, et al. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(4):694-707.

[6] Yin W, Schütze H. MultiGranCNN: An Architecture for General Matching of Text Chunks on Multiple Levels of Granularity[C]// Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing. 2015:63-73.

[7] Wan S, Lan Y, Guo J, et al. A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations[J]. 2015:2835-2841.

[8] Hu B, Lu Z, Li H, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[J]. Advances in Neural Information Processing Systems, 2015, 3:2042-2050.

[9] Pang L, Lan Y, Guo J, et al. Text Matching as Image Recognition[J]. 2016.

[10] Wan S, Lan Y, Xu J, et al. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN[J]. Computers & Graphics, 2016, 28(5):731-745.

[11] Lu Z, Li H. A deep architecture for matching short texts[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:1367-1375.

[12] Yin W, Schütze H, Xiang B, et al. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs[J]. Computer Science, 2015.

[13] Tan M, Santos C D, Xiang B, et al. LSTM-based Deep Learning Models for Non-factoid Answer Selection[J]. Computer Science, 2015.

[14] Tan M, Santos C D, Xiang B, et al. Improved Representation Learning for Question Answer Matching[C]// Meeting of the Association for Computational Linguistics. 2016:464-473.

[15] Santos C D, Tan M, Xiang B, et al. Attentive Pooling Networks[J]. 2016.

[16] X Zhang， S Li， L Sha， H Wang. Attentive Interactive Neural Networks for Answer Selection in Community Question Answering[C]// International Conference on Artificial Intelligence.

[17] L Sha， X Zhang， F Qian， B Chang， Z Sui. A Multi-View Fusion Neural Network for Answer Selection[C]// International Conference on Artificial Intelligence.

[18] Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence.

[19] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[20] Wang S, Jiang J. A Compare-Aggregate Model for Matching Text Sequences[J]. 2016.

[21] Severyn A, Moschitti A. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks[C]// The International ACM SIGIR Conference. ACM, 2015:373-382.

[22] Xiaodong Zhang, Xu Sun, Houfeng Wang. Duplicate Question Identification by Integrating FrameNet with Neural Networks[C]//In the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)

[23] Gong Y, Luo H, Zhang J. Natural Language Inference over Interaction Space[J]. 2018.

原文发布于微信公众号 - 腾讯知文（tencent_wisdom）
原文发表时间：2018-08-27
本文参与腾讯云自媒体分享计划，欢迎正在阅读的你也加入，一起分享。

发表于 2018-08-28

深度学习 自然语言 搜索引擎 AI 人工智能

上一篇：中间件安全加固之Apache

下一篇：美参议员要求FB、谷歌出席听证会 回答平台安全问题

相关文章

来自专栏 marsggbo

Andrew Ng机器学习课程笔记--week9(上)(异常检测&推荐系...

本周内容较多，故分为上下两篇文章。 一、内容概要 1. Anomaly Detection Density Estimation Problem Motiva...

217 9

来自专栏 用户2442861的专栏

深度卷积网络CNN与图像语义分割

转载请注明出处: http://xiahouzuoxin.github.io/notes/

71 1

来自专栏 ATYUN订阅号

Python机器学习的练习二：多元线性回归

在第1部分中，我们用线性回归来预测新的食品交易的利润，它基于城市的人口数量。对于第2部分，我们有了一个新任务——预测房子的售价。这次的不同之处在于我们有多个因变...

432 6

来自专栏 机器之心

学界 | 详解指针生成网络：自动生成长段文本的抽象摘要

作者：Abigail See 机器之心编译 参与：Nurhachu Null 这篇博文是斯坦福大学计算机科学在读博士 Abigail See 对最近自己和其他研...

523 6

来自专栏 杨熹的专栏

Kaggle 神器 xgboost

在 Kaggle 的很多比赛中，我们可以看到很多 winner 喜欢用 xgboost，而且获得非常好的表现，今天就来看看 xgboost 到底是什么以及如何应...

402 4

来自专栏 一心无二用，本人只专注于基础图...

《Single Image Haze Removal Using Dark Channel Prior》...

最新的效果见 ： http://video.sina.com.cn/v/b/124538950-1254492273.html 可处理...

517 10

来自专栏 用户2442861的专栏

循环神经网络——实现LSTM

119 2

10

分享

来自专栏 专知

教你使用Keras一步步构建深度神经网络：以情感分析任务为例

【导读】Keras是深度学习领域一个非常流行的库，通过它可以使用简单的代码构建强大的神经网络。本文介绍基于Keras构建神经网络的基本过程，包括加载数据、分析数...

664 7

来自专栏 崔庆才的专栏

自然语言处理全家福：纵览当前NLP中的任务、数据、模型与...

组合范畴语法（CCG; Steedman, 2000）是一种高度词汇化的形式主义。Clark 和 Curran 2007 年提出的标准解析模型使用了超过 400...

192 0

来自专栏 机器之心

深度 | 从数据结构到Python实现：如何使用深度学习分析医学...

选自Medium 作者：Taposh Dutta-Roy 机器之心编译 运用深度学习技术进行图像和视频分析，并将它们用于自动驾驶汽车、无人机等多种应用场景中已成...

516 8