教育数据挖掘文献调研报告

——读 Question Difficulty Prediction for READING Problems in Standard Tests

PB12345678 **李四**

文献内容综述

1 简介

本文的研究动机来源于在标准化考试中，不同的试卷如果的难度不一，则会导致考试的不公平性。目前对于难度的评估都是教育专家进行评估，这种方法可能会非常主观，并且十分耗费人力，因此本文提出了一种全新的采用数据驱动的方法，通过建模来进行难度的评估。

本文研究了标准化测试中的 READING 问题，不同于其他的文本理解问题，本文所聚焦的问题是：**Question Difficulty Prediction (QDP)**，这类问题具有一些独有的数据特征。第一，文本分为三部分，文章（TD），问题（TQ）和选项（TO）。在处理中需要去建模这三者之间的关联。第二，文章由很多句子组成，而对于一个特定的问题，文章中不同的句子与该问题的关联并不同，通常通过文章中的一些关键句就可以得到这个问题的答案。第三，问题的难度是通过应试者的准确率来表示的，但这其中的问题在于，不同考试的应试者不同，能力也不同，故可能某题的正确率高于另一题是因为其应试者能力更强，而不能代表题目真实的难度。

针对上述三点QDP问题的挑战，本文提出了Test-aware Attention-based Convolutional Neural Network (TACNN) 模型。具体的，对于上述第一点，本文采用了一种基于CNN的结构来提取与表示句子信息，将TD，TQ，TO映射到同一向量空间。对于上述的第二点，本文采用注意力机制来给予与问题相关度更高的信息更高的权重。对于上述第三点，本文采用了一种pairwise的策略，利用同一考试中题目准确率与真实难度正相关的特点，消除了不同考试应试者不同所带来的偏差。

2 相关工作

在教育心理学领域和NLP领域有一些相关的研究。

其中，教育心理学中关于题目难度的估计方法都是需要一些人为的介入，需要耗费人力与专家知识。

而NLP领域中的相关研究主要聚焦于如何让机器可以选出正确的答案，而没有相关的题目难度估计的研究。

3 模型细节

TACNN模型的输入为经过预处理的TD，TQ和TO信息，输出为预测的难度值。模型由神经网络搭建而成，可以分为四个子模块：Input Layer, Sentence CNN Layer, Attention Layer and Prediction Layer。如下图所示：

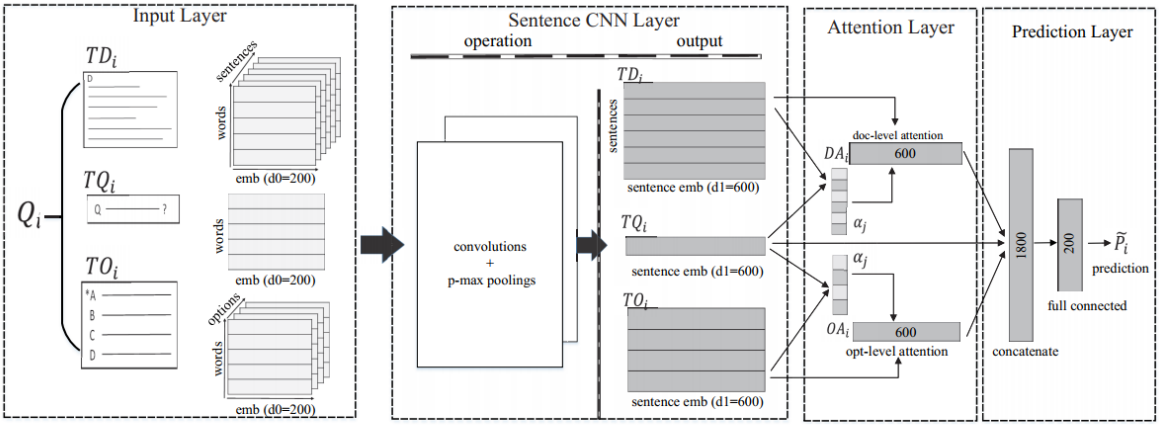

Figure 3: TACNN framework. The numbers in TACNN are the dimensions of corresponding feature vectors.

Input Layer. 输入为所有的文本信息，输出为结构化的数据。其中TD由M个句子构成，TQ和TO中的每个选项由单句构成。每个句子又由N个单词构成，而每个单词则是一个预处理好的d_0维的向量。

因此输出TD是一个 $M \times N \times d_0$ 的张量，而TQ和TO中的每个选项则是一个 $N \times d_0$ 的矩阵。

Sentence CNN Layer. 类似于图像处理中的方法，这里选取大小为k的滑动窗口，对于Input Layer中输出的每个含有N个单词的句子，进行卷积操作，卷积时直接将k个词向量连接，$h_i^c = \sigma(G \cdot [w_{i-k+1} \oplus \cdots \oplus w_i] + b)$ 。卷积后的得到$N-k+1$个d维的向量。再采用池化操作选取每p个中最大的一个。交替采用卷积核池化操作，最终将每个句子映射到一个d_1维的向量空间。此时，TD是一个$M \times d_1$的矩阵，而TQ与TO中的选项是d_1维的向量。

Attention Layer. 采用注意力机制选取和TQ相关的TD和TO，这里相关性的大小采用向量余弦值来衡量。输出值DA是M个向量的加权平均，OA是几个选项的加权平均，权重为其与TQ的向量余弦值。输出的DA和OA是d_1维的向量。

Prediction Layer. 首先将DA，OA和TQ三个d_1维向量采用连接操作聚合，再通过一个全连接层得到难度预测值。

为了训练TACNN模型，还需要定义合适的损失函数。

一个最基本的 Test-independent loss function 可以定义为一个最小二乘loss加上一个l_2正则化项。其问题在于没有考虑到不同考试的应试者不同所带来的误差。

而 Test-dependent pairwise loss function 则是选取同一次test中的不同的两个题目难度的差值作为比较的对象。这种pairwise的方法保证了每一对选取的题目都具有相同的应试者，从而消除了应试者不同所带来的误差。具体的，则是采用 $(P_i^t - P_j^t) - (M(Q_i) - M(Q_j))$ 替代了test-independent中的 $P_i - M(Q_i)$ 。

4 实验

（1）模型与训练设置

Dataset. 数据集是由讯飞提供的约三百万个考试记录。预处理中，过滤掉了那些没有考试记录的题目，因为无法得到其难度值。

Word Embedding. 使用word2vec 的工具对词项进行embedding，每个词项为一个 $d_0 = 200$ 维的向量表示。

TACNN Setting. Input Layer中，根据对于数据分布的观察，选取TD中的句子个数M = 25，每个句子中单词个数N = 40，卷积池化层共有四层，其中每一层卷积层的 k = 3，四层池化层的 p = (3, 3, 2, 1)，每一层输出的特征向量的维度分别为 (200, 400, 600, 600) 。

Training Setting. 网络中所有的参数的初始化均采用在 $(-\sqrt{6/(nin+nout)}, \sqrt{6/(nin+nout)})$ 中随机取值。其中nin和nout是输入输出的feature size。训练中mini batch = 32，dropout的概率为0.2。

（2）基线

因为之前并没有相关问题的研究，所以实验中分别选取了三个TACNN模型的变体进行测试，分别为CNN，ACNN，TCNN，其中A代表采用了注意力机制，T代表采用了 test-dependent 的损失函数进行训练。另外，HABCNN网络结构与TACNN较为相似，同样将其进行调整后作为另一个baseline。

（3）评估方法

分别采用了 Root Mean Squared Error (RMSE)，Degree of Agreement (DOA)，Pearson Correlation Coefficient (PCC)，passing ratio (PR) 四个测试指标对实验结果进行评估。

（4）实验结果

TACNN，CNN，ACNN，TCNN，HABCNN 这几个模型对比来看，TACNN在各项指标的评估中都处于领先，ACNN和TCNN的结果均优于CNN，说明了注意力机制和 test-dependent 的损失函数均对于模型有价值，对比来看，TCNN的结果优于ACNN，说明了 test-dependent 的损失函数具有非常重要的作用。而HABCNN并没有取得比较好的效果，也说明了该模型并不适合QDP的问题。

通过与专家的难度评估对比，也可以看出TACNN模型的表现要优于专家的表现。

另外，注意力机制也对模型的可解释性与可视化提供了非常重要的渠道。注意力机制可以选取出文章中对于当前题目相关度较高的那些语句，帮助人们更好的理解模型的判断。

（5）展望

之后可以将TACNN模型从阅读扩展到听力，写作，口语的题目中，或是其他科目（如数学）中去。

学习心得与总结

参考文献

Zhenya Huang, Qi Liu*, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, Guoping Hu, Question Difficulty Prediction for READING Problems in Standard Tests, *The 31st AAAI Conference on Artificial Intelligence **(AAAI'2017)***: 1352-1359, San Francisco, California USA, February 4-9, 2017.