Computer Science and Application 计算机科学与应用, 2020, 10(9), 1673-1682
Published Online September 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.109177

Hans汉斯

基于ALBERT和同义词词林的主观题自动评分方法

张展鑫，陈平华，刘佳荣

广东工业大学计算机学院，广东 广州
Email: 448201615@qq.com

收稿日期：2020年9月8日；录用日期：2020年9月18日；发布日期：2020年9月25日

摘　要

针对具有参考答案的主观题机器自动评分，既要考虑得分点契合度又要考虑文本整体相似度等问题，提出一种结合ALBERT和同义词词林的主观题自动评分方法。先利用ALBERT的Fine-tuning方法计算参考答案和考生答题之间的文本语义相似度；然后经关键词提取操作，利用同义词词林计算两份文本间面向得分点的关键词相似度；最后结合语义相似度和关键词相似度计算综合得分。真实数据集上的对比实验表明，本文的方法在评分准确率方面有明显提高。

关键词

主观题，自动评分，相似度，ALBERT，同义词词林

Automatic Scoring Method for Subjective Questions Based on ALBERT and Cilin

Zhanxin Zhang, Pinghua Chen, Jiarong Liu

School of Computer Science, Guangdong University of Technology, Guangzhou Guangdong
Email: 448201615@qq.com

Received: Sep. 8th, 2020; accepted: Sep. 18th, 2020; published: Sep. 25th, 2020

Abstract

Aiming at the machine automatic scoring of subjective questions with reference answers, both the fit of the score points and the overall similarity of the text must be considered. An automatic scoring method for subjective questions that combines ALBERT and Cilin is proposed. First use ALBERT's

文章引用: 张展鑫, 陈平华, 刘佳荣. 基于 ALBERT 和同义词词林的主观题自动评分方法[J]. 计算机科学与应用, 2020, 10(9): 1673-1682. DOI: 10.12677/csa.2020.109177

Fine-tuning method to calculate the textual semantic similarity between the reference answer and the test taker's answer. Then use the keyword extraction operation to calculate the score-oriented keyword similarity between the two texts using Cilin. Finally, the comprehensive score is calculated by combining semantic similarity and keyword similarity. Comparative experiments on real data sets show that the method in this paper has a significant improvement in scoring accuracy.

Keywords

Subjective Question, Automatic Scoring, Similarity, ALBERT, Cilin

1. 引言

近年来，得益于互联网的飞速发展和网络上海量的数据，人工智能被广泛应用于人类的生活中，如智能客服、语音助手、机器自动翻译等。简答题、叙述题等主观题机器自动评分研究始于上世纪 60 年代。传统的研究主要采用统计学方法，如：Ellis Page [1]提出的 PEG (Project Essay Grade)方法，评分结合考虑文章平均长度、分号数量和词语稀缺度等因素；高思丹等人[2]提出的评分方法先对关键词进行匹配，然后计算语句相似度，最后计算分数；倪应华[3]提出的评分方法除了对关键词进行匹配，还引入了模糊数学中的单向贴近度，计算考生答题和参考答案之间的贴近程度。传统的统计学方法，仅考虑答案文本长度、关键词数量等浅层因素，没有考虑到语义等深层复杂因素对评分的影响。

随着计算机性能的提高和自然语言处理技术的发展，基于统计的主观题机器自动评分方法渐渐被基于信息检索和自然语言处理技术的主观题机器自动评分方法取代，并且准确率有了较大地提高。如：David Callear 等人[4]提出了 ATM (Automated Text Marker)方法，该方法引入了近义词词典，初步考虑到语义因素；王逸凡等人[5]先用扩展的命名实体识别方法提取关键词，然后采用同义词词林进行相似度计算，最后得出分数；罗海蛟等人[6]将结合专家知识的 LDA 模型引入自动评分中；周洲等人[7]使用 TF-IDF 和 LSI 方法进行自动评分；张翠翠等人[8]采用双向遍历空间模型算法，依据关键字和答案贴合度计算主观题分数。

上述研究方法虽然能提取到文本的整体信息，但是没有考虑到文本内部深层语义和语序等细粒度特征。2018 年，Google 公司的 AI 团队发布了 BERT (Bidirectional Encoder Representation from Transformers)模型[9]。BERT 内部使用 transformer 编码器[10]搭建了双向的神经网络结构，以无标签的方式训练内部参数，提取含有上下文信息的词语和字符特征。实验结果表明，BERT 模型在机器阅读理解顶级水平测试 SQuAD1.1 中的表现已经超越人类，并且还在 11 种不同自然语言处理测试中创出最佳成绩。针对 BERT 的训练参数过多造成的训练周期长等问题，Google 公司又提出了基于 BERT 改进的 ALBERT 模型[11]，该模型在 BERT 的基础架构上使用因式分解和跨层参数共享的方式减少了需要训练的参数，极大地提高了训练效率。

将 BERT 模型直接应用于类似开放式作文等主观题机器自动评分自然是一个不错的选择，然而将 BERT 模型直接应用于简答题和叙述题等主观题，机器自动评分却难以取得最佳效果。因为，与开放的主观题不同，简答题和叙述题具有自身特点：首先，简答题和叙述题往往有参考答案，在人工评分时，

评卷者考虑考生答题的语言逻辑和语句通顺等因素之前，首先考虑的还是考生答题是否符合得分要点(关键步或关键词)；其次，汉语存在的一词多义、同义词等现象。基于此，本文模拟人工评分过程，将 BERT 和同义词词林同时应用于主观题自动评分中，具体步骤是：首先在领域公开数据集上应用 BERT 模型的 Fine-tuning 方法进行文本相似度训练；然后将主观题参考答案和学生答题输入到训练好的模型进行整体评价，获得参考答案和考生答题在文本级别的相似度；接着对参考答案和考生答题进行分词、提取关键词等操作后，通过同义词词林进行深层语义评价，计算参考答案和考生答题的在基于关键词的得分点上的相似度；最后综合上述两种相似度计算得到主观题最终评分。

2. ALBERT 模型

ALBERT 是 Google 公司 AI 团队发布的预训练语言模型[11]，采用双向的 Transformer 编码器结构[10]，如图 1 所示。其中：I_i 表示单个词或字的向量输入，Trm 即 Transformer，T_i 表示最终隐藏层输出，通过编码器里的注意力矩阵和注意力加权后，每个 T_i 都具有整句话上下文的语义信息。

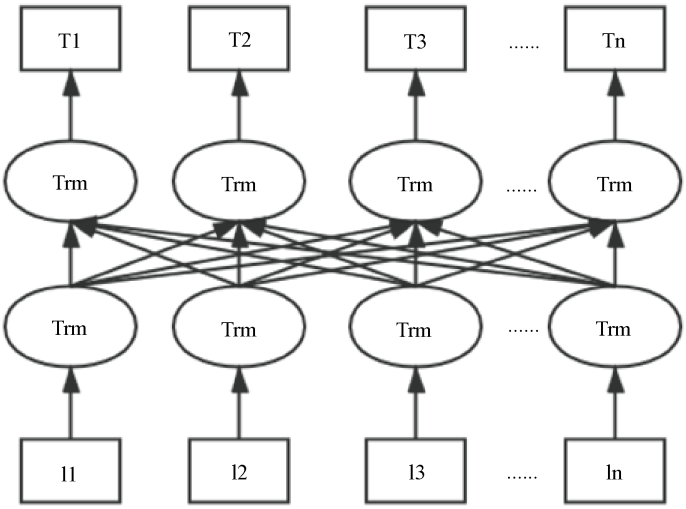

Figure 1. ALBERT model structure
图 1. ALBERT 模型结构图

Transformer 是 Encoder-Deocder 结构的 seq2seq 模型，ALBERT 模型使用了 Transformer 的 Encoder 层提取输入序列的特征信息，Encoder 层由 Self-Attention 层、Feed Forward 层和两个 Add & Norm 层组成，如图 2 所示。输入向量序列时，Transformer 会把输入词的位置信息向量化后与原词向量相加，让输入词带有位置信息。之后编码器对输入的序列经过 Self-Attention 处理，通过计算词间关系矩阵并以加权计算的方式更新输入序列的词向量。在 Self-Attention 层和 Feed Forward 层后面都连着一个 Add & Norm 层，该层的操作就是将上一层的输出和 Add & Norm 层输出直接相加，然后再对相加结果进行归一化。Feed Forward 层是一个简单结构的全连接前馈网络。

其中，在 Encoder 里，最重要的是 Self-Attention 层，如式(1)所示，其中 Q、K、V 是输入词向量矩阵，d_k 是 K 矩阵的维度。

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

在处理输入文本时，Self-Attention 层可以计算出句子内部词与词之间的关联度，并利用词间的关联

度进行权重计算调整原始输入，形成带有上下文信息的词向量。新的词向量由于融合了词性关系，相比原始输入更能体现词在全局序列中的关键度。

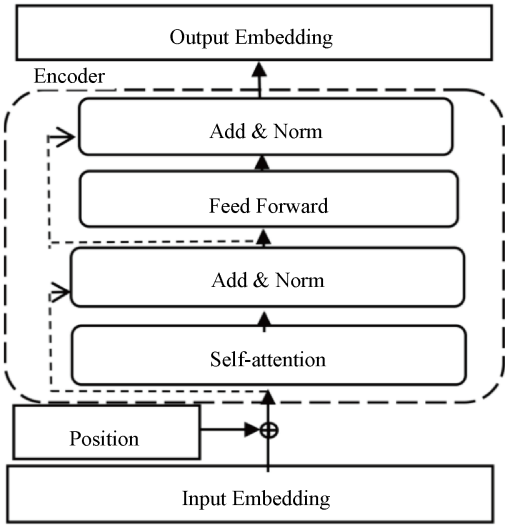

Figure 2. Transformer Encoder structure
图 2. Transformer Encoder 结构图

在实际应用中，Transformer 使用的是 Self-Attention 的扩展版 Multi-Head Attention。Multi-Head Attention 通过多个不同的线性变换对 Q，K，V 进行投影，最后将不同的 attention 结果拼接起来，提高了关注输入句子的不同位置的能力，如公式(2) (3)。

$$\text{MultiHead}\left(Q,K,V\right)=\text{Concat}\left(\text{head}_1,\cdots,\text{head}_h\right)W^o \tag{2}$$

$$\text{head}_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{3}$$

ALBERT 模型使用 Transformer 的 Encoder 层学习每个词的上下文信息，获得了更优秀的词向量。除此之外，ALBERT 模型还用两个无监督预训练任务强化对词向量的语义表示能力，分别是 Masked Language Model (简称 MLM)和 Sentence-Order Prediction (简称 SOP)。

MLM 任务：随机选取输入文本中 15%的词做三种形式的替换，然后去预测被替换掉的词。形式一是用[Mask]标记替换掉 80%的被替换词；形式二是用随机的一个词去代替本来要覆盖的词，随机替换的词占被替换词的 10%；形式三是对余下 10%的词不进行任何替换。这样设计是为了避免模型在微调时会过度依赖训练数据中不存在的[Mask]标记，同时也让模型在计算词向量的时候更依赖上下文信息。

SOP 任务：通过提取同一文本中的两个句子 A 和 B，预测句子中 A 和 B 的语序。取正序的 A 和 B 为正例，改变 A 和 B 的顺序作为反例。因为同一文本中描述的主题相同，所以在同一个文本中选取句子对可以减少文本主题的差异对训练效果的影响。

综上所述，通过 MLM 任务和 SOP 任务的结合，ALBERT 模型可以更精准地描绘出文本的语义信息。

3. 同义词词林

同义词词林[12]是梅家驹等人在 1983 年所完成的同义词词典。由于时间长远，不适合当前年代的使用，于是哈尔滨工业大学信息检索实验室对该词林进行了扩展，完成了词林扩展版[13]。扩展版共收录了 77343 条词语，比前一版本增加了 2 万多词语，并且增加了两层结构，如表 1 所示。

Table 1. Comparison before and after expansion of Cilin
表 1. 词林扩展前后对比

特征	总数	大类	中类	小类	层数	编码长度
扩展前	53895	12	94	1428	3	4
扩展后	77343	12	97	1400	5	8

扩展版通过建立一个树状层次结构将所有词语组织包含其中，如图 3 所示。词林整个树状结构从上到下分成五层，分别是大类层、中类层、小类层、词群和原子词群。同一个原子词群里的词语语义相同或十分接近或关联性很强。

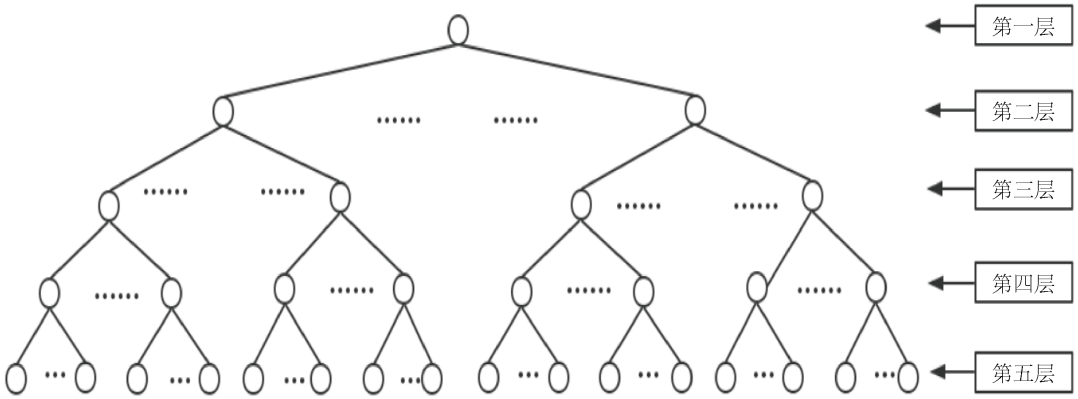

Figure 3. The tree-like hierarchical structure of Cilin
图 3. 同义词词林的树状层次结构

词林对每个词语赋予唯一的 8 位编码，其中，高 7 位为树状层次结构各层次编码最低位为标记位，编码规则如表 2 所示。标记位主要是用于区分原子词群层次下的词语之间的语义关系，分常规同义词、相关词和只有词语本身(即没有同义词也没有相关词)三种情况，分别用“=”、“#”和“@”三个字符表示。

Table 2. Cilin coding rule table
表 2. 词林编码规则表

编码位	符号举例	符号性质	级别
1	A	大类	第一级
2	b	中类	第二级
3	2	小类	第三级
4	6	小类	第三级
5	C	词群	第四级
6	0	原子词群	第五级
7	2	原子词群	第五级
8	=/#/@		

如编码“He03B14 = 出让转让让”表示词语“出让”、“转让”和“让”同属于 H 大类、e 中类、

03 小类、B 词群、14 原子词群下三个词语，他们是同义词。

4. 基于 ALBERT 和同义词词林的主观题自动评分

计算评分的具体流程如图 4 所示。把主观题参考答案和学生答题输入到训练后的 ALBERT 模型，获得参考答案和学生答题在文本级别的相似度；同时使用同义词词林对从参考答案和学生答题提到的关键词序列进行相似度计算，得出两者基于关键词的得分相似度；最后综合两种相似度得到学生答题的综合预测评分，并针对题目总分以预测评分生成预测分数。

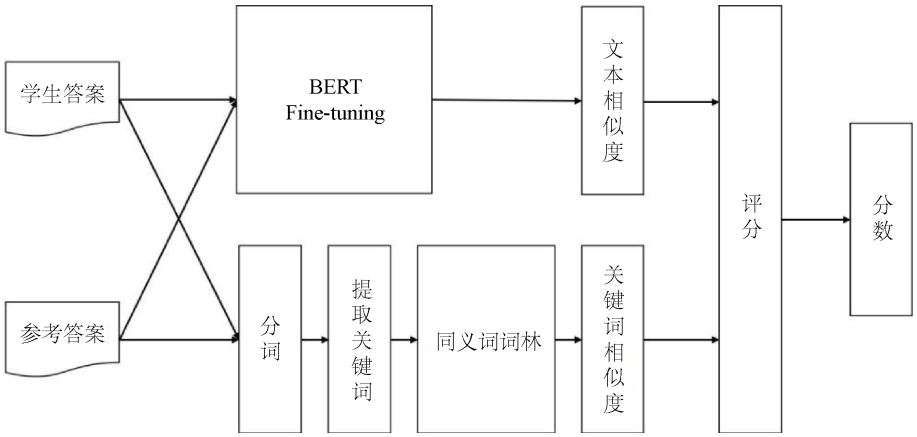

Figure 4. Automatic scoring flowchart for subjective questions
图 4.主观题自动评分流程图

4.1. 文本相似度计算

利用 ALBERT 的 Fine-tuning 方法可以简单高效地实现自然语言序列之间的语义相似度计算。Fine-tuning 方法把语义相似度任务归类为二分类任务，通过计算相似概率与不相似概率得出分类结果，具体流程如图 5 所示。首先将学生答题和参考答案输入到 ALBERT 模型中进行编码，然后将编码后的结果输入到式(4)的线性函数和式(5)的 softmax 函数进行 Fine-tuning，最后得出两个文本的相似度。使用的损失函数是交叉熵损失函数，如式(6)所示。

$$Z = WX + b \tag{4}$$

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{5}$$

$$L = -\left[y \log \hat{y} + (1-y) \log (1-\hat{y}) \right] \tag{6}$$

4.2. 关键词相似度计算

关键词相似度的计算，就是对学生答题文本和参考答案文本进行关键词提取，然后对提到的关键词序列进行词语相似度的计算。

目前词语相似度的计算主要有两种方法，第一种是使用大规模的语料库，并通过一定的概率模型去计算；第二种是基于一些世界知识来计算，比如通过具有完备性的语义词典中的结构特点去计算[14]。

本文用的是第一类的方法，基于同义词词林扩展版，并根据彭琦等人[15]提出的基于信息内容词语相似度计算公式计算词语相似度，如式(7)所示。

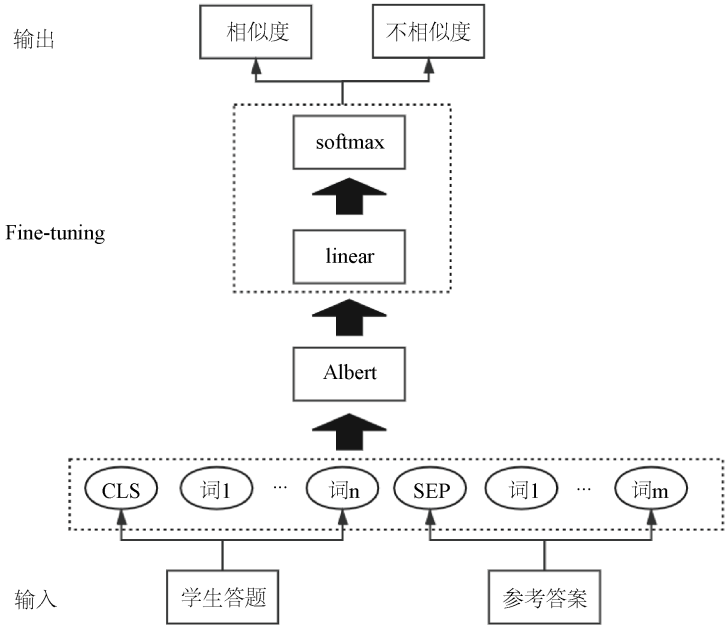

Figure 5. Block diagram of text similarity calculation
图 5. 文本相似度计算框图

$$sim(C_1,C_2)=\frac{MaxDIFF-dis(C_1,C_2)}{MaxDIFF-MinDIFF} \tag{7}$$

其中 *MaxDIFF* 和 *MinDIFF* 分别表示词林的最大差异性与最小差异性，当两个词语是两个不同大类下的叶子节点，最近公共父节点为根节点时，具有最大差异性，当两个词语在同一叶子节点上时，具有最小差异性，根据式(7)的差异性计算公式，可得出最大差异性为 2，最小差异性为 0。

$$dis(C_1,C_2)=IC(C_1)+IC(C_2)-2IC(LCS(C_1,C_2)) \tag{8}$$

$$IC(C)=1-\frac{\log(hypo(C)+1)}{\log(maxnodes)} \tag{9}$$

在式(8)中，$LCS(C_1,C_2)$表示两个词语的最近公共父节点，$IC(C)$表示词语 C 的信息内容含量，由式(9)计算得出。式(9)中的 $hypo(C)$表示词语所在节点的子节点个数，*maxnodes* 表示词林里的全部节点的数量，取值 90114。

具体方法如算法 1 所示。

算法 1 关键词相似度计算

input: 参考答案关键词序列 A_n、学生答案关键词序列 B_m
1. initialize i← 0, j ← 0, ConKeyNum← 0, KeySim
2. while i++ < n **do**
3. **while** j++< m **do**
4. **if** sim(A_i, B_j) >= 0.9 **then**
5. ConKeyNum++
6. end if
7. end while
8. end while
9. KeySim←ConKeyNum/n
10. output: 关键词相似度 KeySim

4.3. 综合评分

教师在批改试题时，通常是先观察学生答案的整体作答情况，然后再查看学生答案中是否具有各个得分点的关键词，最后给出分数。通过参考教师批改试题的评分过程，本文提出了综合评分公式，如式(10)所示，其中 *TextSim* 是利用 ALBERT 得到的文本相似度，取值范围为 0~1，*KeySim* 是基于同义词词林计算得出的关键词相似度，取值范围也是 0~1，*MaxScore* 为题目的分值。其中 α、β 为超参数，由实验筛选出最佳取值。

$$Score=\begin{cases}\left[\alpha\times TextSim+(1-\alpha)\times KeySim\right]\times MaxScore, & TextSim\geq\beta\\ KeySim\times MaxScore, & TextSim<\beta\end{cases} \tag{10}$$

5. 实验

5.1. Fine-Tuning 实验

本文使用的中文预训练模型为 albert_tiny_zh[1]，使用来自哈尔滨工业大学的 LCQMC [16]数据集，该数据集一共有 26 万多个带有是否相似标注的句子对，包括有训练集约 24 万个，验证集约 9 千个，测试集约 1 万 2 千多个。本文对该模型训练了 5 个 epochs，最终在测试集上的准确率为 85%。

5.2. 自动评分实验

5.2.1. 实验设置

本文选取了某中学高中阶段语文月考的一道阅读试题(满分为 6 分)作为测试数据，随机抽取了 100 份考生答卷，每份答卷均已由认可教师进行过人工评分。具体试题和参考答案如下：

试题：请简要概括我国大数据交易面临的困境有哪些？

参考答案：数据的开放和共享跟不上市场的发展，无法满足需求。各类数据主体缺乏共享理念。大数据交易缺乏统一标准，各大数据交易平台的交易规则也存在差异和缺陷，适用范围小。大数据专业人才匮乏。

利用同义词词林扩展版进行关键词相似度计算时，首先要对参考答案和考生答题进行分词以及关键词提取的操作。在本实验里，选用 NLPIR[2] 分词工具进行分词，筛选文本的实词作为关键词。

对于式(10)中 α 和 β 参数的设定，经过多次实验结果表明，当 α 为 0.5，β 为 0.7 时，自动评分和人工评分的分数最接近。

5.2.2. 实验结果分析

本实验选用准确率、QWKappa、平均误差分作为评价指标，选取基于同义词词林的方法、基于双向遍历空间的方法以及基于 LDA 的方法进行实验对比分析。评分结果如表 3、图 6 所示。

Table 3. Experimental results
表 3. 实验结果

方法	准确率	QWKappa	平均误差分
基于 BERT 与词林的方法	91.0%	0.84	0.54
基于词林的方法	87.7%	0.71	0.74
基于双向遍历空间的方法	84.3%	0.65	0.94
基于 LDA 的方法	83.3%	0.54	1.00

[1]https://github.com/brightmart/albert_zh.
[2]https://ictclas.nlpir.org.

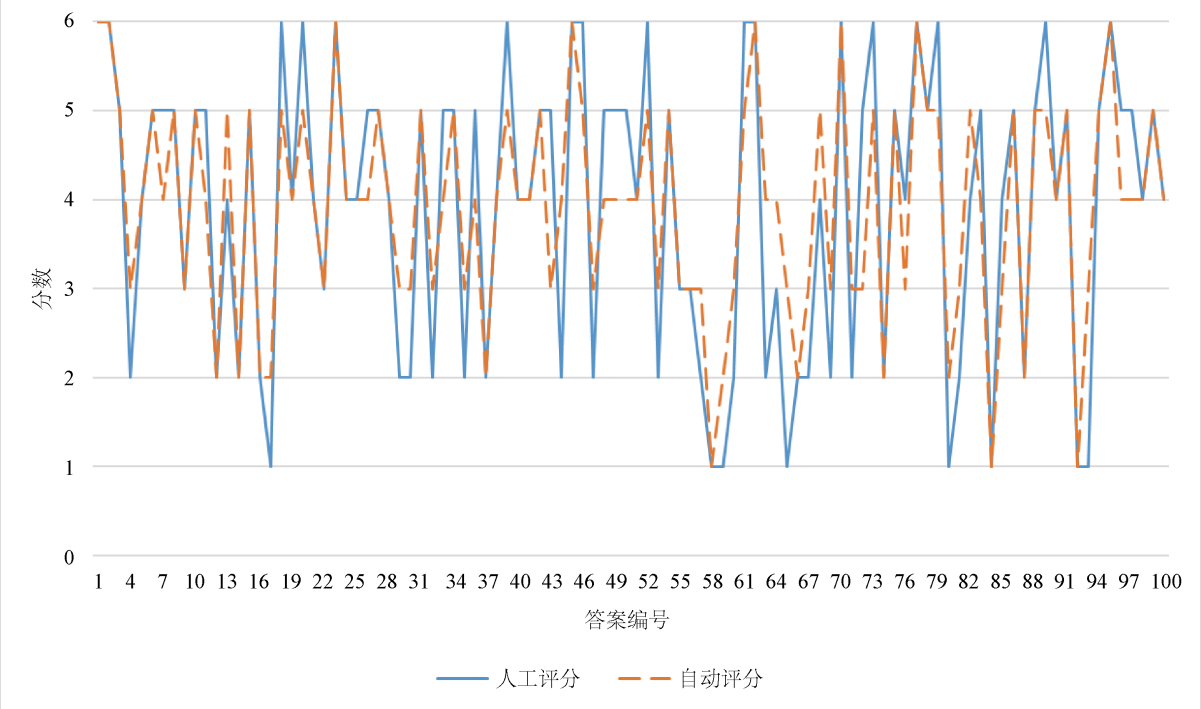

Figure 6. Comparison of manual scoring and automatic scoring
图 6. 人工评分与自动评分的对比

由表 3 可以看出，本文提出的方法效果明显：在三个评价指标上均优于三个 baseline 方法。与表现最好的基于词林的方法相比，本文提出的方法在准确率、QWKappa、平均误差三项评价指标上分别提升了 3.3 个百分点、提升了 0.13、降低 0.2。与人工评分结果相比，本文方法自动评分分数已经非常贴近人工评分分数。总体来看，本文提出的自动评分方法已经取得良好的效果。

6. 总结与展望

针对有参考答案的主观题自动评分既要考虑关键得分点契合度，又要考虑文本整体相似度等问题，提出基于 ALBERT 和同义词词林的主观题自动评分方法。方法使用 ALBERT 预训练模型和同义词词林计算出参考答案与考生答题之间的文本语义相似度以及针对每个得分点的关键词相似度，最后结合两种相似度计算总体分数。实验表明，相比传统的方法，本文的方法在一定程度上提高了评分的性能。

该方法存在一定的弊端，它强调的是两个答案之间的相似度、贴合度，如果是没有标准答案的试题，该方法将束手无策。在下一步的研究工作中，将引入知识图谱等技术，让自动评分的适用性更广。

基金项目

广东省研究生教育创新计划项目(2017JGXM-ZD14)。广东省科技计划项目（2020B1010010010、2019B101001021）。

参考文献

[1] Page, E.B. (1994) Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education*, **62**, 127-142. https://doi.org/10.1080/00220973.1994.9943835

[2] 高思丹, 袁春风. 语句相似度计算在主观题自动批改技术中的初步应用[J]. 计算机工程与应用, 2004, 40(14): 132-135.

[3] 倪应华. 基于 XML 自动阅卷算法的设计与实现[J]. 仪器仪表学报, 2005, 26(z1): 704-705.

[4] Callear, D.H., Jerrams-Smith, J. and Soh, V. (2001) CAA of Short Non-MCQ Answers. *Proceedings of the* 5*th International Computer Assisted Assessmentconference*, Loughborough, July 2001, 1-14.

[5] 王逸凡, 李国平. 基于语义相似度及命名实体识别的主观题自动评分方法[J]. 电子测量技术, 2019, 42(2): 84-87.

[6] 罗海蛟, 柯晓华. 基于改进的 LDA 模型的中文主观题自动评分研究[J]. 计算机科学, 2017, 44(z2): 102-105+128.

[7] 周洲, 侯开虎, 姚洪发, 张慧. 基于 TF-IDF 及 LSI 模型的主观题自动评分系统研究[J]. 软件, 2019, 40(2): 158-163.

[8] 张翠翠, 周国祥, 俞磊, 石雷, 王青青. 基于 MVC 的试卷生成及主观题判卷算法研究[J]. 系统仿真学报, 2020, 32(1): 105-112.

[9] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[10] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, December 2017, 5998-6008.

[11] Lan, Z.Z., Chen, M.D., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R (2019) Albert: A Lite Bert for Self-Supervised Learning of Language Representations. arXiv: 1909.11942.

[12] 梅家驹. 同义词词林[M]. 上海: 上海辞书出版社, 1983.

[13] 哈工大信息检索研究中心. 同义词词林扩展版[DB/OL]. https://www.ltp-cloud.com/download, 2005.

[14] 郭小华, 彭琦, 邓涵, 朱新华. 基于边权重的 WordNet 词语相似度计算[J]. 计算机工程与应用, 2018, 54(1): 172-178.

[15] 彭琦, 朱新华, 陈意山, 孙柳, 李飞. 基于信息内容的词林词语相似度计算[J]. 计算机应用研究, 2018, 35(2): 400-404.

[16] Liu, X, Chen, Q., Deng, C., *et al*. (2018) LCQMC: A Large-Scale Chinese Question Matching Corpus. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, August 2018,1952-1962.