计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300154

面向机器阅读理解的候选句抽取算法

郭 鑫[1] 张 庚[1] 陈 千[1,2] 王素格[1,2]
1 山西大学计算机与信息技术学院 太原 030006
2 计算智能与中文信息处理教育部重点实验室 太原 030006
(guoxinjsj@163.com)

摘 要 使机器理解人类自然语言是人工智能在认知领域的终极目标,机器阅读理解是自然语言处理技术中继语音识别、语义理解之后的一大挑战,要求计算机具有一定的背景常识,全面理解给定文本材料,并根据材料内容对相应的问题作答。随着深度学习的快速发展,阅读理解成为当前人工智能的热点研究方向,涉及机器学习、信息检索、语义计算等核心技术,在聊天机器人、问答系统、智能化教育等多个领域具有广泛的应用前景。文中聚焦微阅读模式,根据问题或选项从给定文本材料中抽取包含答案的候选句,缩小推理范围,为进一步实现机器阅读理解提供技术支持。传统基于特征的方法耗费大量人力,文中将答案候选句抽取看成一种语义相关度计算问题,提出了一种答案候选句排序方法,即 Att-BiGRU/BiLSTM 模型。首先,利用双向长短期记忆和门控循环单元来编码句子中表达的语义信息;其次,设计 Atten 结构,结合相异性和相似性对语义相关度进行建模;最后,采用 Adam 算法来学习模型的参数。在 SemEval-SICK 数据集上的实验结果显示,该模型在测试集上的 pearson 指标超过了基线方法 BiGRU 将近 0.67,在 MSE 指标上超过 BiGRU 方法 16.83%,收敛速度更快,表明双向和 Atten 结构能大大提高候选句抽取的精度。

关键词:长短期记忆模型;门控循环单元;候选句抽取;语义相关度计算
中图法分类号 TP391

Candidate Sentences Extraction for Machine Reading Comprehension

GUO Xin[1],ZHANG Geng[1],CHEN Qian[1,2] and WANG Su-ge[1,2]
1 School of Computer & Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing,Ministry of Education,Taiyuan 030006,China

Abstract The ultimate goal of artificial intelligence is to let machine understand human natural language in cognitive field. Machine reading comprehension raises great challenge in natural language processing which requires computer to have certain common knowledge,comprehensively understand text material,and correctly answer the corresponding questions according to that text material. With the rapid development of deep learning,machine reading comprehension becomes the current hotspot research direction in artificial intelligence,involving core technologies such as machine learning,information retrieval,semantic computing and has been widely used in chat robots,question answering systems and intelligent education. This paper focuses on micro-reading mode,and answer candidate sentences containing answers are extracted from given text,which provide technology support for machine reading comprehension. Traditional feature-based methods consumes lots of manpower. This paper regards candidate sentences extracting as a semantic relevance calculation problem,and proposes an Att-BiGRU/LSTM model. First,LSTM and GRU are used to encode the semantic expressed in a sentence. Then,the dissimilarity and similarity are captured with an Atten structure for semantic correlation. Last, adam optimizer is used to learn the model parameters. Experiment results show that Att-BiGRU model exceeds the baseline method of nearly 0.67 in terms of pearson,16.8% in terms of MSE on SemEval-SICK test dataset,which proves that the combination of the bidirectional and Atten structure can greatly improve the accuracy of the candidate sentences extraction,as well as the convergence rate.

Keywords Long short term memory,Gated recurrent unit,Candidate sentences extracting,Semantic correlation calculation

---

到稿日期:2019-03-28 返修日期:2019-07-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:山西省应用基础研究计划项目(201701D221101,201901D111032);国家自然科学基金项目(61502288,61403238,61673248);山西省重点研发计划项目(201803D421024)
This work was supported by the Natural Science Foundation of Shanxi Province(201701D221101,201901D111032),National Natural Science Foundation of China(61502288,61403238,61673248) and Key R&D Program of Shanxi Province(201803D421024).
通信作者:陈千(chenqian857@163.com)

1　引言

机器阅读理解是当前人工智能的一个重要研究方向，涉及自然语言处理、机器学习等多种领域核心技术，在自动问答系统、聊天机器人中具有广泛的应用[1-2]。阅读理解是一种综合能力的考察，根据背景文本材料，针对选项判断其是否符合材料文意。一般来说，选项的正误大多能根据背景材料中的少量几个句子来推断，本文将这些关键句称为候选句，候选句抽取有助于对选项直接相关的句子进行准确定位，为下一步阅读理解任务提供必要的支撑。目前已有工作专门针对散文阅读理解开展基于抽象词语关联的答案获取方法[3]的研究，针对散文选择题开展选项和背景材料的一致性蕴含问题[4]，这些工作的前提都需要进行基于候选片段的抽取。文献[4]采用了较为简单的词语匹配抽取算法，由于级联产生的错误传递导致最终正确率不高。另外，多数阅读理解系统从问题出发将候选句抽取同答案生成融合为一个问题来提出联合模型，但该方法缺乏可解释性，且最终答题的精确度并未得到显著提高[5]。本文聚焦如何从背景材料中自动抽取候选片段，作者认为该问题是机器阅读理解的首要任务，能极大程度地缩小推理范围，为阅读理解提供前期支撑。

作为传达人类意见、情感等语义信息的一种载体，自然语言具有高度弹性、模糊性和可扩展性，阅读理解任务不仅要求理解自然语言传达的场景和过程，更要求理解自然语言中的情感和观点。传统句子建模基于语言的文法、句法模型，根据文法构建句子对应的分析树，从而理解句子的层次化结构。由于不同语言的句法通常并不一致，一个句子的语义并非是句子各个部分语义的简单组合，因此句法驱动的句子模型构建往往有很大的局限性。目前，机器阅读系统中的答案候选句抽取模块多采用基于词重叠或基于向量空间模型的句子相似度计算方法，由于缺乏对语义相关度的把握，很难将蕴含答案的候选句子定位出来。基于概率主题模型的方法虽然避免了词重叠、一词多义的现象，但背景材料通篇都是关于特定主题的句子，适用于多文本处理而不适用于表达句意的相关度，因此抽取精度仍然较低。此外，基于特征工程的传统机器学习方法将该问题看成一个分类问题，虽然精度有一定提升，但随着背景材料领域的转换，需要重新耗费大量人力来设计合理的特征以适应新的领域材料，缺乏一定的可扩展性。

基于深度神经网络构建的句子模型，不依赖于特定语言的句法，其输入往往为句子中的每一个词的词向量，神经网络根据其学习的特征函数从词向量中自动提取特征。将提取的特征作为原始句子的高层次的抽象，可以用于句子语义分析或者句子分类任务。本文将答案候选句抽取转换为一种语义相关度计算问题，提出了一种基于长短期记忆模型/门控循环单元（Long Short Term Memory/Gated Recurrent Unit，LSTM/GRU）的端对端深度神经网络的答案候选句排序方法。该方法一方面聚焦语义相关度建模而不是相似度，另一方面采用端对端方式省去了大量繁琐的特征设计工作。本文提出了一种 Att-BiGRU 模型，不仅保留了句子间的相似关系，同时考虑了句子间的相异程度，从而更准确地刻画了句子的相关度语义关系。对于该模型性能的验证，本文没有选择斯坦福大学著名的 SQuAD 数据集[6]，因为该数据集没有具体候选句标注信息，本文选择 SEMEVAL 英文标注数据集，通过实验证明了采用 Atten 结构和双向结构相较于未采用这两种结构的神经网络模型在 MSE 和 pearson 指数上均具优势。

2　相关工作

答案候选句抽取算法大体可以分为 3 类：基于信息检索的方法和基于概率主题模型的方法和基于深度学习的词嵌入方法。基于信息检索的方法将候选句抽取任务看成一个检索问题，早期语义相似度计算方法主要基于空间向量模型。利用 TF-IDF 刻画文档语句相似度，是一种无监督的学习方式，这种方法简单地采用词匹配和重叠的方式，难以捕获句子之间微妙的相关信息。基于概率主题模型方法[7]的基本思路是通过每一篇文档的主题分布和每一个主题中词的分布来确定句子的相关度，一般采用经典的概率主题模型（Latent Dirichlet Allocation，LDA）或潜在语义分析方法（Latent Semantic Analysis，LSA）[8]等，将高维的文档降维到潜在语义低维空间来计算语义相关度，这类方法主要从篇章级别对文档进行建模，而对于微阅读模式的句子级建模表现较差。基于深度学习的词嵌入方法主要基于 skip-gram 神经网络的词向量来构建句子向量[9]，其中，ALEX 等[10]提出了一种用神经网络构建二元语言模型的方法，其思路与后续的语言模型差别较小，计算方法是将一个句子通过模型训练之后的词向量表示为一个二维矩阵，通过计算矩阵的余弦相似度来刻画词义相似度，但该方法仅刻画了语句相似度，而相关度包含相似程度和相异程度两个方面，因此不足以刻画语句相关度。

近年来，随着深度学习在图像处理、语音领域的突破性进展，大量的神经网络开始应用于答案候选句抽取的问题。Johnson 等提出了一种金字塔卷积神经网络模型（Deep Pyramid Convolution Neural Networks，DPCNN）[11]用于文本相似度计算，该模型对一定窗口下的词向量维度进行卷积操作，三层金字塔的构造能够有效地减少卷积神经网络带来的高维计算量。Tai 等[12]提出了一种树型 LSTM 的网络架构，句子被标记并表示成语义树，再被递归放到树型 LSTM 神经网络的结构中。大量 LSTM 被用于对句子进行建模，有效解决了句子长短不一的问题。一种基于卷积神经网络（Convolution Neural Networks，CNN）和 LSTM 混合神经网络被提出[13]，用于社区自动问答系统中，其特点是将输入的问题、选项以及答案候选句分散到 LSTM 和 CNN 层中，并用注意力机制来刻画语义相似度。

以上方法要么采用卷积神经网络 CNN，虽然其具有特征学习和抽取功能，但无法适用于句子长短不一的场景；要么采用 LSTM 模型进行句子相似度计算，但该方法缺乏相关度的建模。本文综合多个模型的结构特点，结合 Atten 结构和双向结构对语句相关度进行联合建模，探索不同的递归神经网络模块在英文语料上对答案候选句抽取任务的适应性和准确率的影响。对 3 种基本的句子模型进行比较，实验结果表明，在实验所使用的 3 组英文数据集中，本文提出的 Att-BiGRU 模型可以有效地用于提取数据特征，在句意相关度上取得了比较满意的结果。

3 基于 Att-BiGRU 的答案候选句抽取方法

本节首先根据样例给出答案候选句的基本定义，然后将其表示成一个语义相关度计算问题，最后给出模型的结构和训练方法。

3.1 问题描述

在用直觉进行阅读理解答题时，为了回答特定问题，需要在背景材料中定位答案候选句。表 1 列出了 H 这个选项在材料原文《白鹿原上奏响一支老腔》中直接相关的 3 句话，要判断 H 的正误，首先需要从背景材料中将 P_1，P_2，P_3 3 个能支持构成 H 是否成立的证据抽取出来，显然这比从整个原文中得出结论要容易得多。

表 1 选项的候选句样本示例

Table 1 Sample record of options and related candidate segments

P_1：后来，有作家朋友看过老腔的演出，不无遗憾地对我说过这样的话，你的小说《白鹿原》是写关中大地的，要是有一笔老腔的画面就好了。
P_2：我却想到，不单是一笔或几笔画面，而是整个叙述的文字里如果有老腔的气韵弥漫…
P_3：直到后来小说《白鹿原》改编成话剧，导演林兆华在其中加入了老腔的演唱，让我有了一种释然的感觉。
H：朋友为小说《白鹿原》没有写老腔的笔墨而感到遗憾，我对此深有同感。

根据以上示例，本人认为面向阅读理解的候选句自动抽取对于机器阅读理解的意义重大。本文给出了候选句抽取问题的形式化描述，如算法 1 所示。

算法 1 候选句抽取算法

输入：给定的文章 context，给定的候选句 option

输出：选择文章中相关度最高的句子

Given：context＝{s_1，s_2，s_3，…，s_n}，option＝h

Ouput：candidate＝{s_t}，

Require：η＝score(s_0，h)

for sentence in context

  if score(s_j，h)＞η then

    η＝score(s_j，h)，t＝j

  end if

end for

算法 1 中，$score(s_i, s_j)$用于度量两个句子的相关程度，因此将问题转化为计算两个句子的相关度。

传统深度神经网络仅对词语做词向量训练，然后通过计算两个句子的词向量夹角余弦值来计算语句的相似度，但是这并不能准确表达语句相关度语义。本文采用一种 Atten 结构的方法，即在隐藏层设计中对隐藏层做绝对值差和点积运算之后，再对运算后的值进行计算表达。前者的目的是计算两个句子之间的相异关系，后者是为了计算两个句子之间的相同关系，从而从正反两个角度的量化值刻画句子的相关度。

3.2 基于深度学习神经网络答案候选句的抽取算法

LSTM 是由 Hochreiter 等于 1997 年提出的[14]，属于循环神经网络(Recurrent Neural Network，RNN)的一种变种，近几年 LSTM 被广泛应用于自然语言处理任务中，如机器翻译中的 Encoder-Decoder 模型[9]和问答系统(Question-Answering，QA)。LSTM 模型很好地处理了远距离依赖问题，它含有的记忆细胞可以长时间存储信息，它的 3 个门结构分别是输入门 i、遗忘门 f 和输出门 o，可用来控制信息的流动。在时刻 t，输入为 x_t，$h(t)$表示 t 时刻的输出。LSTM 网络的转换方程如式(1)所示：

$$\begin{aligned} i_t &= sigmoid(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}) \\ f_t &= sigmoid(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \\ g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hc}h_{(t-1)} + b_{hg}) \\ o_t &= sigmoid(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \\ c_t &= f_t \odot c_{(t-1)} + i_t \odot g_t \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{1}$$

其中，$\odot$表示向量之间逐元素相乘。

Cho 等[15]于 2014 年根据 LSTM 单元架构，提出了一种简单但不失效果的 RNN 神经网络 GRU。Chung 等在序列建模任务上的实验结果表明，GRU 的收敛速度比普通 RNN 要快，它既保持了 LSTM 的效果同时又具有简单的结构，所需参数明显少于 LSTM[16]。GRU 模型只有两个门，分别为更新门和重置门，更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，其值越大说明前一时刻的状态信息带入得越多。重置门用于控制忽略前一时刻状态信息的程度，其值越小说明忽略得越多。GRU 转换方程如式(2)所示：

$$\begin{aligned} r_t &= sigmoid(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \\ z_t &= sigmoid(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \\ n_t &= \tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{(t-1)} + b_{hn})) \\ h_t &= (1 - z_t) \odot n_t + z_t \odot h_{(t-1)} \end{aligned} \tag{2}$$

本文实验部分 LSTM 和 GRU 的对比结果进一步验证了该模型的高效性和准确性。

3.3 Att-BiLSTM/Att-BiGRU 模型

标准循环神经网络 LSTM 能有效处理时序序列数据，但它忽略了未来的上下文信息，因此一种含有双向结构的循环神经网络(Bidirectional LSTM/Bidirectional GRU，Bi-LSTM/Bi-GRU)[17]被提出。由于候选句抽取同时强调未来信息和过去信息，因此采用双向循环神经网络在实现序列学习的同时还能保留过去和未来的上下文信息。本文在该双向结构的基础上，设计并实现了一个层级迭代的 Att-BiLSTM/Att-BiGRU 模型。其具体网络结构如图 1 所示。

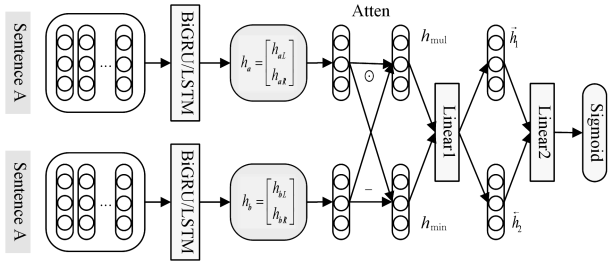


图 1 Att-BiLSTM/BiGRU 的网络结构

Fig. 1 Network architecture of Att-BiLSTM/GRU

首先使用 BiLSTM 模型为每个句子生成句子表示 $\boldsymbol{hL}$ 和 $\boldsymbol{hR}$，令：

$$\boldsymbol{h}_a = \begin{bmatrix} \boldsymbol{h}_{aL} \\ \boldsymbol{h}_{aR} \end{bmatrix}, \boldsymbol{h}_b = \begin{bmatrix} \boldsymbol{h}_{bL} \\ \boldsymbol{h}_{bR} \end{bmatrix} \tag{3}$$

为了对句子相关度的相似成分和相异成分进行联合建模，如式(4)所示，利用 $\boldsymbol{h}_{\text{mul}}$ 刻画两个对象的相似程度，利用 $\boldsymbol{h}_{\text{min}}$ 刻画两个对象的相异程度。由于 BiLSTM 存在双向结

构，因此在计算过程中将 $\boldsymbol{h}_{\mathrm{mul}}$ 和 $\boldsymbol{h}_{\mathrm{min}}$ 两个部分的双向结构连接起来，并通过线性变换和激活函数 tanh，最终计算得出 $\boldsymbol{h}_s$，如式(5)所示[16]。

$$\boldsymbol{h}_{\mathrm{mul}}=\boldsymbol{h}_a\odot\boldsymbol{h}_b,\boldsymbol{h}_{\mathrm{min}}=|\boldsymbol{h}_a-\boldsymbol{h}_b| \tag{4}$$

$$\boldsymbol{h}_s=\tanh(W_{sa}\boldsymbol{h}_{\mathrm{mul}}+W_{sb}\boldsymbol{h}_{\mathrm{min}}+b_s) \tag{5}$$

本文称上述网络结构为 Atten 结构。该结构有利于表示相关度的正反方面。此外，$\boldsymbol{h}_s$ 可以进一步表示为以下形式：

$$\boldsymbol{h}_s=\begin{bmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{bmatrix} \tag{6}$$

为了能够将两个语句的相似度进行恰当的表示，本文将计算之后的双向向量变为 $\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$，通过 relu 激活函数，得到结果 $\boldsymbol{h}_t$。为了能够将运算之后的值锁定到[0,1]区间内，采用 sigmoid 变换函数进行输出，如式(7)、式(8)所示：

$$\boldsymbol{h}_t=relu(W_{ta}\boldsymbol{h}_1+W_{tb}\boldsymbol{h}_2+b_t) \tag{7}$$

$$y=sigmoid(W_o\boldsymbol{h}_t+b_o) \tag{8}$$

这样，本文就计算出了句子相关度的关系值 $\boldsymbol{y}$。

3.4　隐藏层 Sum 化

Att-BiLSTM/Att-BiGRU 模型仍存在一些不足之处，这是因为双向 LSTM/GRU 模型仅对模型双向输出做了一个简单的连接操作，并没有很好地表达向量之间的量化操作关系。隐藏层向量之间的关系是通过一定的比例来确定相似度关系值。因此，基于该问题，本文设计了一个全连接的多层感知机网络(Multi-Layer Perceptron，MLP)，并且通过求和比例的方式来计算句意之间的相似度关系值，目的是能够清晰反映出每一个双向隐层单元差值与乘积值之间的比例关系，如式(9)所示：

$$\begin{aligned}&\boldsymbol{h}_1=h_{aL}\odot h_{aR}\\&\boldsymbol{h}_2=|h_{aL}-h_{aR}|\\&\boldsymbol{h}_3=h_{bL}\odot h_{bR}\\&\boldsymbol{h}_4=|h_{bL}-h_{bR}|\end{aligned} \tag{9}$$

输出向量值隐藏层 $\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3,\boldsymbol{h}_4$ 体现了两个句子之间双向输出间的差异关系，它们拼接成 $\boldsymbol{h}_k$。最后通过 relu 激活函数计算出每一层的特征向量 $\boldsymbol{h}_k'$：

$$\boldsymbol{h}_k'=relu(W_{hk}\boldsymbol{h}_k+b_{hk}) \tag{10}$$

特征向量表达了语句之间的向量比例关系，因此对式(10)的每个隐藏层做求和运算 $p_k=\sum_i h_{ki}'$，并重新排列向量 $\boldsymbol{T}=(p_1,p_2,p_3,p_4)$。同样，输出值经过激活函数进行处理，$\boldsymbol{y}=sigmoid(W_T T+b_T)$。

3.5　模型训练

针对 3.4 节提出的模型，本文采用 Adam (Adaptive Moment Estimation)梯度下降算法[18]来进行参数学习。相比于随机梯度下降算法(Stochastic Gradient Descent，SGD)，Adam 能够加速 SGD 在正确方向的下降并抑制震荡。在训练过程中，损失函数选取交叉熵损失函数来计算目标值与预测值之间的损失，对于批输入的问题句使用矩阵式的交叉熵损失函数。

$$loss(o,t)=-\frac{1}{n}\sum_{k=0}^{n}(t_k\log o_k+(1-t_k)\log(1-o_k)) \tag{11}$$

训练前需要对模型进行初始化，包括模型参数的定义、优化器定义(包含梯度下降算法优化器)以及损失函数定义，根据训练语料中的基本格式，本文采用单句训练的方法，训练步骤如下。1)清除梯度优化器的梯度值：这是每次循环训练的第一步，需要清除上一次计算过程中的梯度值，即将梯度值置位为零。2)隐藏层参数初始化：通过随机初始化赋予网络权重参数较小的数值。3)预测值句子对的数值：通过神经网络层的后向传播过程，计算两个句子的相似度，输出层使用了 Sigmoid 函数，将相似度的表示方法转化为[0,1]区间内的数值，方便训练具体的数据集。4)用损失函数计算损失值：使用交叉熵损失函数，得到误差值。5)损失值反向传播：使用误差的反向传播，将误差值反向传播到每一个神经网络层中，进行梯度的更新操作。6)利用优化器计算下一步的梯度值：用来更新神经网络中的权重矩阵值，使梯度值更新至每一个神经网络权重矩阵。

4　实验设计与结果分析

4.1　数据集及评测标准

为了验证 Atten 结构和双向结构的有效性，本节设计并实现了 Att-BiGRU 和 Att-BiLSTM 模型，具体代码已发布到 github 网站[1)]，接着在标准的英文标注文本数据集上设计了算法收敛、模型性能等实验来评估本文方法。采用 SemEval-SICK [2)] 语义文本相似性视频描述数据集作为评测数据集。对于给定的一对句子，语义相关性任务是预测两个句子在意义上相关度的评分。该数据集包括 train，test，trail 3 个部分，其规模分别为 4500，500，4927，合计 9927 个句对。每个句对都被注释意义上的相关性和两者之间的蕴涵关系。数据集的每条记录为一个句对，包括 ID 号、句子对、相关度评分以及情感蕴涵标签。数据集的样本示例如表 2 所列，其中 C 代表矛盾，N 代表无关，E 代表蕴含。

表 2　评测语料数据集样本展示

Table 2　Sample record of Datasets

pair_ID	*sentence_A*	*sentence_B*	*score*	*judgment*
4	The young boys are playing outdoors and the man is smiling nearby	There is no boy playing outdoors and there is no man smiling	3.6	*C*
105	Four children are doing backbends in the gym	Four girls are doing backbends and playing outdoors	3.8	*N*
253	A hiker is on top of the mountain and is doing a joyful dance	A hiker is on top of the mountain and is dancing	4.7	*E*

参照文献[12]，本文拟采用均方差指数(Mean Square Error，MSE)来评估所提模型和人工标注的差距，采用 pearson 指数来度量模型的结果同人工标注数据的线性相关关系，这两个指标分别从正反两个方面刻画了模型的优劣，*MSE* 值越小，*pearson* 指数越大，模型的表现就越好，具体指标计算公式如下：

$$pearson=\frac{x_1\cdot x_2}{\sqrt{\sum_{k=1}^{n}(x_{1k}-\bar{x})^2}\cdot\sqrt{\sum_{k=1}^{n}(x_{2k}-\bar{x})^2}} \tag{12}$$

1) https://github.com/MobTgZhang/AttRNN

2) http://alt.qcri.org/semeval2014/task1/

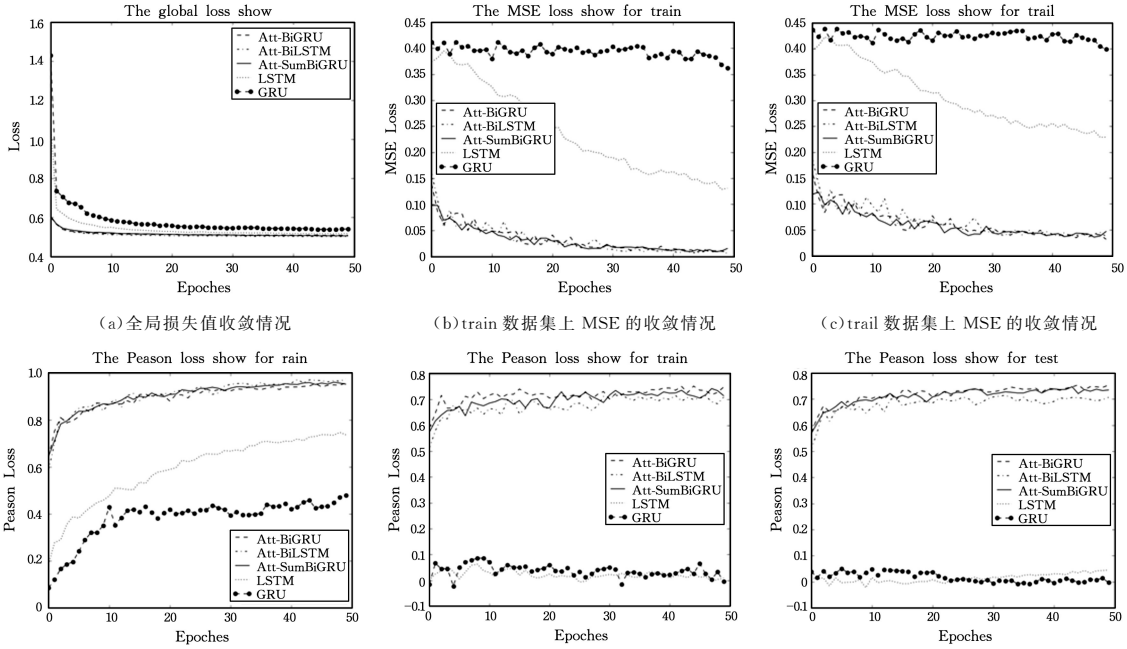
$$MSE=\frac{1}{N}\sum_{k=1}^{N}(x_k-y_k)^2 \tag{13}$$

4.2 实验设计

在预处理过程中，本文使用词向量对语句进行初始化建模，其具体的处理步骤如下：1）训练模型之前建立包含所有可能出现的词构成的词典，统计每个句子对的词语并生成词典；2）生成句子向量，在训练模型前，对词嵌入层进行初始化，词典的词向量大小为 $vocab\_size * embedding\_dim$，其中 $vocab\_size$ 是词典的大小，$embedding\_dim$ 是嵌入词向量的维度大小。对句子对进行分词处理后，生成序列化的词语，再经词典转化后得到句中对应词的索引序列，这样在训练过程中就能够对应每个词生成的初始随机词向量。在比较 Att-BiGRU 模型和 LSTM/GRU 模型时，通过改变隐藏层的维度来控制 LSTM 参数以及 Att-BiGRU 模型隐藏层的数量。选取 BiGRU 作为系统的 baseline，并将所提模型以及只含有 LSTM/GRU 和双向网络结构的神经网络进行对比。为公平起见，所有神经网络模型的输入词向量统一设置为 300，若模型中含有隐藏层，则隐藏层的输入维度为 200，输出维度为 100，此外 sum 化层的维度设置为 100。接下来，从全局损失值和 MSE 的收敛速度方面来对比各个模型训练的优劣，并将各个模型在 train 和 trail 数据集上的预测结果进行了 pearson 和 MSE 两项指标的比较，最后展示了效果较差和较好的若干示例，并分析了模型的优缺点。

4.3 实验结果分析

实验对每个模型训练 50 个批次，在每一个批次训练完成之后，保存模型参数等信息，然后对 3 个数据集上收敛的情况进行测试。每一个批次的测试结果如图 2 所示。



(a)全局损失值收敛情况 (b)train 数据集上 MSE 的收敛情况 (c)trail 数据集上 MSE 的收敛情况

(d)train 数据集上 pearson 的收敛情况 (e)trail 数据集上 pearson 的收敛情况 (f)test 数据集上 pearson 的收敛情况

图 2 模型收敛性能比较

Fig. 2 Convergence on global loss, MSE and pearson index comparison with various models

图 2(a) 给出了全局损失值的收敛情况对比，图 2(b)和图 2(c) 给出了 MSE 在 train 和 trail 数据集上的收敛情况对比；图 2(d)－图 2(f)给出了 3 个训练集在 pearson 指数上的收敛测试结果。从图 2 可以看出，在训练过程中，相比 baseline 模型，Att-BiLSTM/GRU 以及 Att-sumBiGRU 在 3 种数据集上，在 pearson 指数、MSE 以及全局误差上的收敛速度以及最终值都要明显优于 GRU 和 LSTM 模型。此外，Att-BiGRU 模型在测试集上具有优良的线性相关的性质，对训练的样本有非常良好的拟合效果。MSE 均值误差表示了目标值与实际测试值之间的距离值，虽然训练过程仅针对 train 数据集，但在 test/trail 数据集上同样具有不错的拟合效果。

表 3 列出了 Att-BiGRU，Att-BiLSTM，Att-SumBiGRU，Att-GRU，word2vec，LSTM，GRU，BiLSTM，BiGRU 等多个神经网络模型配置在 test，trail2 个子集上的 MSE 和 pearson 指标预测结果。综合比较表中的数据可以看出，Att-BiGRU 模型能更准确地刻画语义相关度的建模特征，排名第二的 Att-SumBiGRU 模型同样表现不错。

表 3 各种模型在测试集和 trail 集合上的 pearson 和 MSE 指标结果

Table 3 MSE and pearson index on Test and Trail dataset using various models

模型	pearson		MSE	
	Test	Trail	Test	Trail
Att-BiGRU	**0.7544**	**0.7521**	**0.0313**	**0.0314**
Att-BiLSTM	0.7142	0.7160	0.0388	0.0366
Att-SumBiGRU	0.7472	0.7423	0.0337	0.0356
Att-GRU	0.5392	0.5280	0.0539	0.0555
Word2Vec	0.2902	0.2278	0.0740	0.0807
LSTM	0.0451	0.0623	0.2123	0.2289
GRU	0.0506	0.0862	0.3854	0.3991
BiLSTM	0.0622	0.0715	0.2727	0.2620
BiGRU	0.0819	—	0.1996	0.1866

从表 3 可以看出，加入 Atten 结构的 4 个模型 Att-BiGRU，Att-BiLSTM，Att-SumBiGRU，Att-GRU 的 pearson 和 MSE 指标与未加入 Atten 结构的模型相差较大，这也证实了

Atten结构中的隐藏层相减操作以及相乘操作有助于解释两个句子之间的相关关系。表4列出了若干实例，可以看出，在一些矛盾和蕴含类型的句对上，Att-BiGRU模型效果非常好，第1—2行和第4行实例预测误差分别为0.3，0.139和0.057，但第3行实例较差，仅为1.354。经分析，本文认为系统对broccoli属于flower没有常识支持，且训练语料未出现该词，因此本文提出模型还无法处理未登录词。

表4　评测语料数据集样本展示

Table 4　Sample records of Datasets

ID	*sentence_A*	*sentence_B*	*score*	*predict*
4	The young boys are playing outdoors and the man is smiling nearby	There is no boy playing outdoors and there is no man smiling	3.6	3.304
124	Three women are dancing	A few women are dancing	4.6	4.461
1525	A woman is planting some flower	A woman is cutting broccoli	1.2	2.554
4661	A woman is cutting a potato	A woman is slicing a potato	4.8	4.857

结束语　本文将机器阅读理解中的答案候选句抽取问题看作是句子相关度计算，同传统相似度计算不同，该算法不仅要识别句子间相似成分，还要识别句子间相异成分。本文提出了一种Att-BiGRU神经网络模型，Atten结构考虑了句子间的相似和相异成分，双向GRU结构较好地解决了句子长度不一致问题。在SemEval的SICK数据集上的实验结果证明了双向结构和Atten结构的结合能极大提高算法的性能，有效地抽取候选句的局部语义，从而很好地判断句意的相关性。

由于未登录词问题造成模型判断错误，拟引入预训练词向量结合增强学习技术解决该问题；此外，语料过于单一，本文将在不同语种（中文、英文）、不同规模和不同类型的若干语料上进一步探讨该模型的适用性。

参 考 文 献

[1] YAN Z, TANG D, DUAN N, et al. Assertion-based QA with Question-Aware Open Information Extraction[J]. arXiv: 1801.07414, 2018.

[2] WANG S, YU M, CHANG S, et al. A Co-Matching Model for Multi-choice Reading Comprehension[J]. arXiv: 1806.04068 [cs], 2018.

[3] CHEN Q, CHEN X F, GUO X, et al. Multiple-to-One Chinese Textual Entailment for Reading Comprehension[J]. Journal of Chinese Information Processing, 2018, 32(4): 87-94.

[4] QIAO P, WANG S G, CHEN X, et al. Word Association Based Answer Acquisition for Reading Comprehension Questions from Prose[J]. Journal of Chinese Information Processing, 2018, 32(3): 135-142.

[5] WANG Z, LIU J, XIAO X, et al. Joint Training of Candidate Extraction and Answer Selection for Reading Comprehension[J]. arXiv: 1805.06145 [cs], 2018.

[6] RAJPURKAR P, JIA R, LIANG P. Know What You Don't Know: Unanswerable Questions for SQuAD[J]. arXiv: 1806.03822 [cs], 2018.

[7] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[10] ALEX R. Can Artificial Neural Networks Learn Language Models? [C]// Proceedings of the Sixth International Conference on Spoken Language Processing, ICSLP 2000/INTERSPEECH 2000. Beijing, China, 2000: 202-205.

[11] JOHNSON R, ZHANG T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]// 55th Annual Meetings of the Association for Computational Linguistics. Vancouver, Canada, 2017: 562-570.

[12] TAI K S, SOCHER R, MANNING C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[C]// 53rd Annual Meeting of the Association for Computational Linguistics, 2015, 5(1): 1556-1566.

[13] ZHOU X, HU B, CHEN Q, et al. Answer Sequence Learning with Neural Networks for Answer Selection in Community Question Answering[C]// Proceedings of the 53rd Annual Meeting of the Association for Computationl Linguistics. Beijing, China, 2015: 713-718.

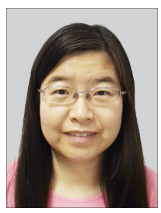
[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[M]// Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, 1997: 1735-1780.

[15] CHO K, VAN M B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734.

[16] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [C]// International Joint Conference on Neural Networks. 2015.

[17] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[M]// Piscataway, NJ: IEEE Press, 1997.

[18] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]// Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015: 1-15.



GUO Xin, Ph.D, lecturer. Her main research interests include feature learning and natural language processing.

CHEN Qian, associate professor. His main research interests include topic detection and natural language processing.