基于双线性函数注意力 Bi-LSTM 模型的机器阅读理解

刘飞龙 郝文宁 陈 刚 靳大尉 宋佳星
(中国人民解放军理工大学 南京 210007)

摘 要 近年来,随着深度学习(Deep Learning)在机器阅读理解(Machine Reading Comprehension)领域的广泛应用,机器阅读理解迅速发展。针对机器阅读理解中的语义理解和推理,提出一种双线性函数注意力(Attention)双向长短记忆网络(Bi directional-Long Short-Term Memory)模型,较好地完成了在机器阅读理解中抽取文章、问题、问题候选答案的语义并给出了正确答案的任务。将其应用到四六级(CET-4,CET-6)听力文本上测试,测试结果显示,以单词为单位的按序输入比以句子为单位的按序输入准确率高 2%左右;此外,在基本的模型之上加入多层注意力转移的推理结构后准确率提升了 8%左右。

关键词 深度学习,机器阅读理解,注意力,Bi-LSTM
中图法分类号 TP181 **文献标识码** A

Attention of Bilinear Function Based Bi-LSTM Model for Machine Reading Comprehension

LIU Fei-long HAO Wen-ning CHEN Gang JIN Da-wei SONG Jia-xing
(PLA University of Science and Technology, Nanjing 210007, China)

Abstract With the wild usage of deep learning in machine reading comprehension in the past few years, machine reading comprehension has developed rapidly. In order to improve machine reading comprehension's semantic comprehension and inference abilities, an attention of bilinear function based Bi-LSTM model was proposed, which has good performance in extracting semantics of questions, candidates and articles, and producing the correct answers. We tested the model on CET-4 and CET-6 listening text materials. The results show that the accuracy rate of word-level input is about 2% higher than sentence-level input. Besides, the accuracy rate can increase about 8% after adding infe-rence structure with multi-layer attention.

Keywords Deep learning, Machine reading comprehension, Attention, Bi-LSTM

1 引言

阅读理解的一般形式是接收一篇给定的文章,测试者阅读并理解文章的内在意义,对文章的题目给出自己的理解和回答,完成这一过程需要抽象化和概念化语言文字。随着机器智能的发展,用机器完成阅读理解任务逐渐成为研究的热点。从文字处理的角度看,机器阅读理解[1-2]是自然语言处理(NLP)中的核心任务之一。本质上,机器阅读理解和人阅读理解面临的问题是类似的,不过相对于人类而言,机器对抽象性的非规则推理显得无所适从。为了降低任务难度,目前很多研究的机器阅读理解都将世界知识(类似于人类成长过程中所积累的经验)排除在外,采用人工构造的相对简单的数据集,回答一些比较简单的问题。

随着深度学习在自然语言处理领域的广泛使用[3-4],基于深度学习的机器阅读理解迅速发展。机器阅读理解语义的方式也从传统的依据语法、句法等专业性较强的领域知识转向端对端的深度神经网络方面。从近几年的研究来看[5-8],当前基于深度学习的机器阅读理解比较常见的任务形式是人工合成问答、实体补全[9]和备择答案预测。

人工合成问答是人工事先构造好由若干简单事实形成的文章以及相对应的问题,由机器阅读理解文章内容并进行一定的推理,从而得出正确答案,正确答案一般是文章中的某个实体词。

实体补全是在机器阅读并理解一篇文章内容后,对机器提出相关问题,而问题往往是抽掉实体词的句子,机器回答问题的过程就是预测问题句子中被抽掉的实体词,一般要求这个被抽掉的实体词是在文章中出现过的。

备择答案预测是机器依据文章、文章的相应问题及候选答案,经过理解和推理,在候选答案中预测出正确答案。答案的内容不是文章中的句子,这是该任务和上述任务的不同之处。对于典型的任务比如听力测试,其数据的存储形式为 $\langle D,Q,H,A\rangle$,其中 D 为待理解的文章,Q 为根据文章提出的问题,H 为候选答案的集合,A 为答案。图 1 展示了一份备择答案预测的示例。

刘飞龙(1994-),男,硕士生,主要研究方向为军用数据与知识工程,E-mail:genesis_fl@139.com;郝文宁(1971-),男,教授,主要研究方向为数据工程、海量高维数据规约;陈 刚(1974-),男,教授,主要研究方向为军用数据工程、信息安全;靳大尉(1979-),男,副教授,主要研究方向为军事运筹、数据工程、信息安全;宋佳星(1993-),男,硕士生,主要研究方向为兵器科学与技术。

Document：M：Hi，Janet，I hear you've just returned from a tour of Australia. Did you get a chance to visit the Sydney Opera House ？

W：Of course I did. It would be a shame for anyone visiting Australia not to see this unique creation in architecture. Its magnificent beauty is simply beyond description.

Q：What do we learn from this conversation ？

A) Janet loves the beautiful landscape of Australia very much.

B) Janet is very much interested in architecture.

C) Janet admires the Sydney Opera House very much.

D) Janet thinks it's a shame for anyone not to visit Australia.

A：C

图 1　机器阅读理解中备择答案预测的样例

人工合成问答由于缺乏训练数据，相关的研究还不成熟。从任务形式上看，实体补全和备择答案预测基本相同，不同之处在于实体补全任务中的候选集是文章中的实体词集合，而备择答案预测任务的候选集是给定的句子集合，这些句子通常不是文章中出现的句子，相比而言，备择答案预测任务更困难，但是更具有代表性，因此本文选择备择答案预测作为研究内容。在实验数据集方面，本文选择 1991 年至今的英语四六级听力测试部分；在模型的构建方面，注意力值的计算采用双线性函数，文字的初步语义理解采用能够学习长期依赖信息的双向长短记忆网络，推理则采取多层注意力转移的方式。实验表明，该模型能较好地理解语义及推理，并且相对于将听力内容以句子为单位的按序输入模型，以单词为单位的按序输入的准确率提高了 2%左右；此外，在基本的模型之上加入多层注意力转移的推理结构能更好地提升实验结果，准确率提升 8%左右。

2　相关工作

2.1　LSTM

长短时间记忆网络（Long Short Term Memory networks，LSTM）[10]，是一种特殊的 RNN，它能够解决深度学习中 RNN 梯度消失的问题。LSTM 的模型由专门的记忆存储单元组成，通过精心设计的遗忘门、输入门和输出门来控制各个记忆存储单元的状态，通过门的控制保证了随着隐藏层在新的时间状态下不断叠加输入序列，前面的信息能够继续向后传播而不消失。

LSTM 模型包括以下部分：输入序列 $X=\{x_1,x_2,\cdots,x_n\}$、步长及对应的输入、输入门 i_t、遗忘门 f_t 和输出门 o_t。记忆单元 C_t（见图 2）控制调整长期记忆单元记忆和遗忘哪些信息。在单位时间 t 的 LSTM 记忆单元 C_t 的值为经过输入门 i_t 对中间过渡内容 $\widetilde{C}_t$ 调整后的值和遗忘门 f_t 对上一个单位时间 $t-1$ 记忆内容 C_{t-1} 调整后的值之和。

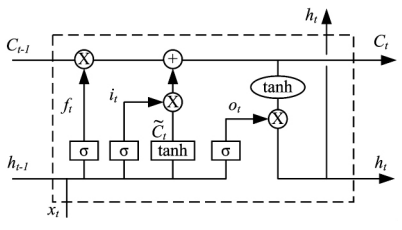

图 2　LSTM 记忆单元结构图

结合的公式如下：

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t$$

其中，

$$\widetilde{C}_t=\tanh(W_c x_t+U_c h_{t-1})$$

$$i_t=\sigma(W_i x_t+U_i h_{t-1})$$

$$f_t=\sigma(W_f x_t+U_f h_{t-1})$$

其中：输入门 i_t 控制新内容的输入，遗忘门 f_t 控制旧内容的遗忘；σ 为对应元素相乘的函数，如逻辑 S 形函数或双曲正切函数。一旦记忆单元 C 更新，当前隐藏层 h_t 则根据当前输出门 o_t 得到的结果计算得出：

$$o_t=\sigma(W_o x_t+U_o h_{t-1})$$

$$h_t=o_t*\tanh(c_t)$$

LSTM 不同的单元通过自适应地控制各自的控制门和记忆细胞来忘记、记忆和展示记忆内容。各个门的开闭不需要同步，同一时刻各个单元门的开闭自行控制。基于 RNN 的多重 LSTM 单元可以同时捕捉在网络中移动的数据。

对于一个输入序列 $S:t_1,t_2,\cdots,t_n$，其正向序列为 $\overrightarrow{S}:t_1,t_2,\cdots,t_n$，反向序列为 $\overleftarrow{S}:t_n,t_{n-1},\cdots,t_1$。对于序列 S 中元素 t_i，其上文信息即序列 S 中序列顺序在 t_i 之前的元素 $\{t_1,t_2,\cdots,t_{i-1}\}$ 包含的信息，其下文信息为序列 S 中序列顺序在 t_i 之后的元素 $\{t_{i+1},t_{i+2},\cdots,t_n\}$ 包含的信息。

单向的 LSTM 网络根据接收序列 S 的不同顺序可以训练不同的网络。对于序列 S 而言，单向 LSTM 如果接收正向序列 $\overrightarrow{S}$，则训练得到正向 LSTM，其网络中每个单元仅仅参考序列 S 中的上文信息，并传递给其后的单元；同样，接收反向序列 $\overleftarrow{S}$ 训练得到的则是反向 LSTM 网络，其网络中每个单元仅仅参考序列 S 中的下文信息（或者说是反向序列 $\overleftarrow{S}$ 的上文信息），并传递给其后的单元。概括来说，单向 LSTM 接收某个序列输入的信息，其网络中单个单元仅仅参考了该序列的上文信息，而对下文信息没有涉及，为了弥补该缺点，充分利用上下文信息，构建双向的 LSTM（简称 Bi-LSTM）网络显得很有必要。Bi-LSTM 网络以单向 LSTM 网络为基础，其基本思想是对每一个训练序列按照向前和向后分别训练两个 LSTM 网络（正向 LSTM 网络和反向 LSTM 网络），并且两个网络连接着同一个输出层。Bi-LSTM 的网络输出层能够完整地捕获输入序列中每一个点关于过去和未来的上下文信息。

2.2　Attention

注意力[11]的转移是人的大脑在接收或者处理外部信息时通过感官巧妙而合理地改变对外部信息的关注点，选择性地忽视与自身不太相关的内容，并对自身需要的信息做放大处理。通过改变关注点，人脑对所集中注意力部位的信息的接收灵敏度和信息处理速度都大大增强，能有效地过滤不相关信息，突出密切相关的信息。

自然语言处理任务中注意力的作用主要体现在对文本语义的抽取，假设一篇文档 D 中单个单词向量为 p_i，而某任务对该文档中单个单词 p_i 的关注度为 α_i，则该文档的表示如下：

$$V_D=\sum_{i=1}^{\|D\|}\alpha_i*p_i$$

其中，$\|D\|$ 为文档按序输入的单词数；α_i 根据任务不同取值不同，通过改变 α_i 的值可以改变任务对文档 D 中的关注点。假设任务由某个语义向量 V_Q 表示，则 α_i 可以表示为 p_i 和

V_Q 的函数 $F(p_i, V_Q)$，计算方式有以下几种常见函数。

向量的点积[12]：$F(p_i, V_Q) = p_i \cdot V_Q$。

前向神经网络[13]：$F(p_i, V_Q) = \tanh(W_D p_i, W_Q V_Q)$。其中，$W_D$，$W_Q$ 分别是对应文档、问题神经网络的权值矩阵。

双线性函数[14]：$F(p_i, V_Q) = p_i W V_Q$。其中，p_i 和 V_Q 的维度分别为 $\| p_i \|$ 和 $\| V_Q \|$，W 为 $\| p_i \| \times \| V_Q \|$ 维的矩阵。

通过研究发现，双线性函数能较好地提升神经网络的性能[15]，并且在尚未加入推理网络的情况下较其他函数能取得较好的结果，因此本文采取基于双线性函数的注意力值计算方式作为语义抽取的模型组成结构。

3 模型构建

本节针对备择答案预测任务构建双线性函数注意力 Bi-LSTM 模型，重点介绍模型的整体架构、文章和问题以及候选答案的语义向量表示、注意力计算、答案预测。

3.1 模型的整体架构

模型的整体架构如图 3 所示（子模型网络的展开部分将在接下来的小节内详细介绍），其中 D 表示文章，Q 表示该文章对应的问题，A, B, C, D 表示问题的候选答案。Net_Q 网络（网络结构见图 6）的输入是单个句子，输出是单个句子的语义向量表示，功能是抽取候选答案或者问题或者文章中单个句子的语义向量表示。Net_DQ（网络结构根据 D 的不同输入方式分两种不同形式，分别见图 4 和图 5）网络的初始输入是 Q 的语义向量表示 V_Q 和 D，输出是 D 和 Q 的综合语义向量表示 V_{DQ}^i（i 为叠加推理次数），在其后的多层推理过程中，Net_DQ 的输入为 D 和 V_{DQ}^i，输出是 V_{DQ}^{i+1}，n 次推理后得到 V_{DQ}^n，功能是抽取文章 D 和问题 Q 的综合语义向量表示。Net_A（网络结构见图 7）网络的输入是 D 和 Q 的综合语义向量表示 V_{DQ}^n 和 4 个候选项的语义向量表示 $\{V_A, V_B, V_C, V_D\}$，输出是 4 个选项 $\{A, B, C, D\}$ 中的一个，功能是答案预测。Net_DQ 中的 Net_Dp 或者 Net_Ds 的功能是抽取文章 D 的语义向量表示，区别在于 D 的输入形式不同，Net_DQ 中 Attention 的功能是计算问题关于文章的注意力值。

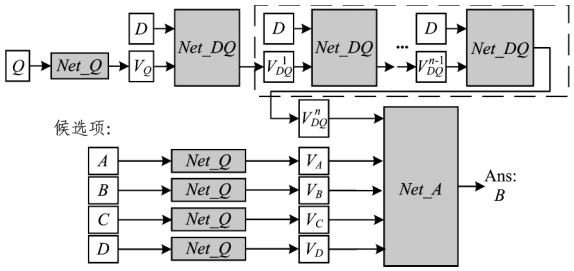

图 3 模型的整体架构

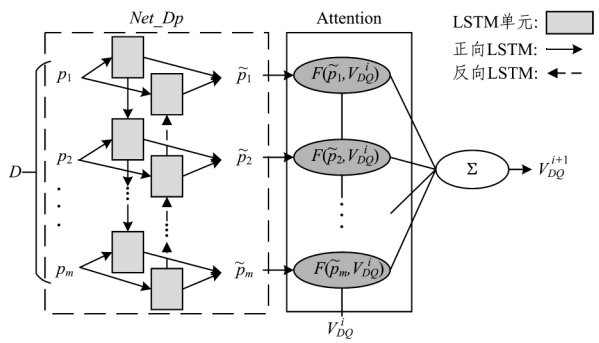

图 4 模型整体架构中 Net_DQ 的展开图（D 以单词形式输入）

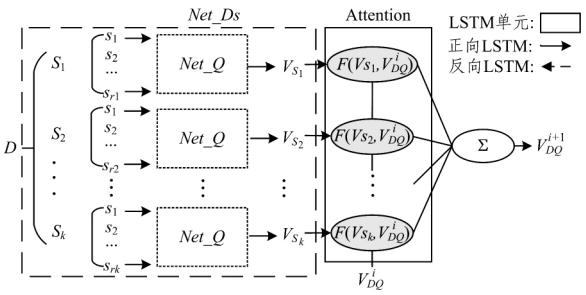

图 5 模型整体架构中 Net_DQ 的展开图（D 以句子形式输入）

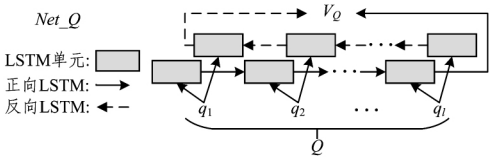

图 6 单个句子语义向量表示抽取模型模型整体架构中 Net_Q 的展开

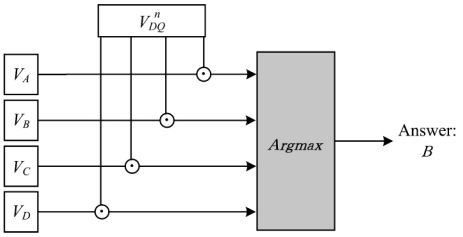

图 7 整体模型架构中 Net_A 的展开（答案预测部分）

模型运行的主要步骤如下：首先通过计算问题关于文章内容的注意力（网络结构如图 4、图 5 中的 Attention 部分），得出文章和问题的综合语义向量表示 V_{DQ}^n（n 为叠加推理次数）；其次，对候选答案进行语义抽取，得到候选答案的语义向量表示 $\{V_A, V_B, V_C, V_D\}$；最后，采用余弦相似性计算 V_{DQ}^n 与候选答案 $\{V_A, V_B, V_C, V_D\}$ 之间的相似度，选择与 V_{DQ}^n 最相近的一个候选答案作为结果返回。

3.2 文章、问题及候选项的语义向量表示

3.2.1 问句和候选项的语义向量表示

从结构上看，问句和候选项都是单个句子的形式，因此采用相同的网络结构 Net_Q 抽取语义，在此处以问题为例介绍 Net_Q 模型，该模型对于句子形式的语义信息抽取表现较好。在此之前，所有句子中的单词都映射到对应的 d 维词向量，即 $w_i \in R^{d \times |v|}$（v 表示所有互不相同的单词数目），则问题中的单词词向量可以表示为 $Q: q_1, q_2, \cdots, q_l \in R^d$。其次，关于问题中单词的表示，采取 Bi-LSTM 来表征每个单词及其上下文的语义信息；关于问题的表示，采取 Bi-LSTM 中正反向的尾节点的隐藏层状态进行拼接来表征整个句子的语义信息。对于正向 LSTM 来说，其尾部单词（句子最后一个单词）LSTM 隐藏层节点融合了整个句子正向的语义信息；而反向 LSTM 的尾部单词（句子的第一个单词）则逆向融合了整个句子反向的语义信息，将这两个时刻 LSTM 节点的隐藏层状态语义向量拼接起来则可以较好地融合表征问题的整体语义。正向隐藏层的状态表示为 $\vec{h}_i$（其中 i 为问题中单词的位置序列，$\vec{h}_i$ 的维度表示为 $\| \vec{h} \|$），逆向隐藏层状态表示为 $\overleftarrow{h}_i$，则问题的语义向量表示公式如下：

$$V_Q = [\vec{h}_{\|Q\|}, \overleftarrow{h}_1]$$

其中，$\| Q \|$ 表示问题中最后一个单词，[ ] 表示拼接两个向

量，则 $V_Q \in R^h$，$h=2\|\tilde{h}\|$，通过同样的处理可以得到关于候选答案的语义向量表示$\{V_A, V_B, V_C, V_D\} \in R^h$。

3.2.2　文章的语义向量表示

理论上，上述表示问题语义信息的模型也可以用来表征文章的语义信息，但是通常不采取这种方法，主要原因是文章内容往往比较长，而上述模型不能较好地表征太长的内容，所以类似上述模型的方法会丢失大量的信息。因此对长文章的语义向量表示一般保留正反向所有隐藏层的状态，而不是像上述模型那样只取正反向隐藏层尾节点的状态。具体的表征方法根据文章输入形式的不同分两种形式：基于单词序列的语义向量表示（如图 4 中长虚线 *Net_Dp*）和基于句子序列的语义向量表示（如图 5 中长虚线 *Net_Ds*）。

基于单词序列的文章语义向量表示将一篇文章看成有序的单词流序列，在这个有序序列上使用 Bi-LSTM 来对文章进行建模表达，每个单词对应 Bi-LSTM 序列中的一个单位时间 t 的输入，则该网络正向和反向隐藏层状态是分别融合了本身词义以及其上下文语义的语言编码。因此每个单词的语义可以由正向 LSTM 隐藏层状态和反向 LSTM 隐藏层状态拼接来表示。一篇文章 D 中的单词词向量可以表示为 $D: p_1, p_2, \cdots, p_m \in R^d$，用 $\vec{h}, \overleftarrow{h}$ 分别表示正向和反向隐藏层状态，则单个单词的输出表示如下：

$$\tilde{p}_i = [\vec{h}_i, \overleftarrow{h}_i] \in R^h$$

其中：

$$\vec{h}_i = LSTM(\vec{h}_{i-1}, p_i), i=1,2,\cdots,m$$

$$\overleftarrow{h}_i = LSTM(\overleftarrow{h}_{i+1}, p_i), i=m,m-1,\cdots,1$$

文章 D 的语义向量表示为 $V_D = \tilde{p}_1, \tilde{p}_2, \cdots, \tilde{p}_m \in R^h$。

基于句子序列的文章语义向量表示类似于基于单词序列的语义向量表示，基于句子序列的文章语义向量表示将一篇文章以句子为单位按序分开，即 $D: \{S_1, S_2, \cdots, S_k\}$（其中 S_i 为句子），对单个句子 $S_i: s_1, s_2, \cdots, s_{ri} \in R^d$ 的处理采取上述处理问题和候选项的模型 *Net_Q*，得到 S_i 的语义向量表示：

$$V_{S_i} = [\vec{h}_{\|S_i\|}, \overleftarrow{h}_1]$$

其中，$\|S_i\|$ 表示句子中的尾单词，[ ] 表示拼接两个向量，文章 D 的语义向量表示为 $V_D = V_{S_1}, V_{S_2}, \cdots, V_{S_k} \in R^h$。

3.3　基于双线性函数的注意力值计算

为了确定问题关于文章各个部分的注意力，需要集成问题语义向量表示和文章的语义向量表示。通过双线性函数计算问题和文章中各个单词或句子之间的关系，在有监督训练的过程中通过更新矩阵 W 学习凸显出其中与问题密切相关的内容，弱化与问题不相关内容，最后输出用于选择候选答案的综合语义向量表示 V_{DQ}^n。在 *Net_DQ* 中注意力值的计算方法是使用双线性函数计算问题语义向量表示和文章的语义向量表示之间的权值 α_i。通过整合文章各个部分的权值以及该部分的语义向量表示 *Net_DQ* 输出文章关于问题的 h 维语义向量表示：

$$V_{DQ} = \sum_{i=1}^{\|D\|} \alpha_i D_i \in R^h \ (D_i \text{ 表示 } \tilde{p}_i \text{ 或者 } V_{S_i})$$

其中，$\alpha_i = \underset{i}{\text{softmax}} V_Q^{\mathrm{T}} W D_i \ (W \in R^{h \times h})$。

多层推理的结构（如图 3 中长虚线部分）即通过多次叠加注意力层转移注意力，得到文章关于问题的高层语义表征：

$$V_{DQ}^j = \sum_{i=1}^{\|D\|} \alpha_i D_i \in R^h, j=1,2,\cdots,n$$

其中，$\alpha_i = \underset{i}{\text{softmax}} V_{DQ}^{j-1\mathrm{T}} W D_i$。

3.4　预测答案

对于上述输出的文章关于问题的最终语义向量表示 V_{DQ}^n 和候选答案集合$\{V_A, V_B, V_C, V_D\}$，答案的选择过程即通过计算候选答案和 V_{DQ}^n 之间的相似度，选择其中最相关的选项作为结果输出：

$$ans = \underset{i \in \{A,B,C,D\}}{\arg\max} (V_i^{\mathrm{T}} \cdot V_{DQ}^n)$$

4　实验

4.1　数据集

实验数据来源于从 1991 年实行至今的四六级（CET-4，CET-6）考试的听力部分，经过整理得到样本共 908 个（其中训练样本 728 个，测试样本 100 个，模型评估样本 80 个），每个样本由 4 部分组成，每部分占一行，每个候选答案各占一行，组成形式为〈D, Q, H, A〉，其中 D 为待理解的文章，Q 为根据文章提出的问题，H 为候选答案的集合（共 4 个），A 为答案。对各个部分单独统计，可以得到 D 的平均长度为148.6，Q 的平均长度为 9.7，单个候选答案的平均长度为8.21。

4.2　模型初始化

词向量的维度 $d=100$，对于词向量的构建，选择已经预训练的 100 维谷歌词向量文件[16-17]来映射；对于未在词向量文件中出现的单词，通过随机函数将其映射为向量元素范围在[−1，1]的 100 维随机向量。注意力计算中双线性函数的权值 W 的元素随机初始化为在[−0.01，0.01]内的矩阵，Bi-LSTM 网络中的参数则随机初始化为符合 $N(0, 0.01)$ 分布的数值。

4.3　模型参数的设置

Bi-LSTM 网络中的正向和反向隐藏层的节点数均设置为 128 个，为了避免过拟合，所有的 Bi-LSTM 网络共享一套权值；训练过程中误差传递采用随机梯度下降（SGD）[6]的方法，学习率设置为 0.1；对于误差的更新，则采用批处理的形式，每次由 30 个样本一起更新，即 batch 设置为 30 个，dropout 设置为 0.2，推理层数设置为 4，轮回次数设置为 100，选择在评估数据集上性能最好的一组参数作为训练的模型参数输出。

4.4　对比实验

对比实验分两类，第一类只考虑候选答案的字符长度，第二类将初始化的词向量进行简单的向量相加得到问题、答案的整合向量，通过计算向量之间的相似度来选择。

基于候选答案长度的模型：选择候选答案长度最长的方案、选择候选答案长度最短的方案。

基于整合向量的模型：根据初始化的词向量，问题可以表示为 $Q: q_1, q_2 \cdots q_l \in R^d$，进行简单的向量相加得到问题向量：

$$V_Q' = \sum_{i=1}^{\|Q\|} q_i$$

同样可以得到候选答案的向量表示：

$$\{V_A', V_B', V_C', V_D'\}$$

基于整合向量的模型有两种选择方案：

1）候选答案之间最不相同的选择方案。首先计算候选答案两两之间的余弦相似度，之后对每个选项与其他 3 个选项之间的相似度求平均，得到该选项与其他选项之间的平均相

似度，最后从 4 个平均相似度之间选择最小的作为结果输出，该结果即为与其他选项最不同的选项。

2)选择与问题最相似的答案。计算候选答案与问句之间的相似度，选择与问句相似度最大的作为结果输出。

4.5 实验结果

使用测试集上的正确率(测试集上输出答案正确的样本数/总测试样本数)作为评判的依据，实验结果如表 1 所列。

表 1 实验结果

模型		平均准确率/%
基于候选答案长度	长度最长	21.66
	长度最短	30.45
基于整合向量	候选答案之间最不相同	28.62
	候选答案与问题最相似	34.21
本文模型(不加入多层推理)	基于单词	45.51
	基于句子	44.24
本文模型(加入多层推理)	基于单词	54.65
	基于句子	52.71

基于候选答案长度的模型：在全部的 908 个样本数据集上，每次随机抽取 100 个样本作为测试集，共抽取 10 次，计算 10 次正确率的平均值作为该类模型的最终结果(如表 1 第一部分)。经过实验可以得到选择候选答案长度最长的方案的准确率为 21.66%，选择候选答案长度最短的方案的准确率为 30.45%。

基于整合向量的模型：与基于候选答案长度的模型类似，10 次随机抽取各 100 个样本作为测试集，最终输出 10 次正确率的平均值作为该类模型的结果(如表 1 第二部分)。经实实验可以得到选择候选答案之间最不相同的方案的准确率为 28.62%，而选择与问题最相似的答案方案的准确率为 34.21%。

本文模型：在训练集(728 个样本)上训练 5 次，每次选择不同的数值初始化并迭代 100 个轮回，选择在评估数据集(80 个样本)上取得最好结果的 5 组模型参数输出，共得到 5 组模型参数，分别用在测试集(100 个样本)上，最终输出为 5 组模型在测试集上准确率的平均值(如表 1 第三部分)。经过实验可以得到在不加入多层注意力转移推理结构的情况下，以单词为单位的文章语义向量表示的准确率为 45.51%，以句子为单位的文章语义向量表示的准确率为 44.24%；在加入多层注意力转移推理结构的情况下，以单词为单位的文章语义向量表示的准确率为 54.65 %，以句子为单位的文章语义向量表示的准确率为 52.71%。

为了验证本文所提方法在统计上显著不同于其他方法，另外设置配对 t 检验的实验。选择本文模型中基于句子的多层推理模型作为待检验模型。实验数据方面，为了避免本文模型训练所采取的训练集对准确率造成的影响，选择评估数据集的 80 个样本和测试数据集的 100 个样本构成抽样总体库。实验过程：在抽样总体库中各随机抽取 50 个样本作为测试集，重复 10 次，每个测试集分别在基于候选长度的模型、基于整合向量的模型、基于句子的多层推理模型上测试，最终每个模型得到 10 个准确度值(为计算方便，准确度值取为准确率 $*100$)，将 5 组准确度值送入 spss 数据分析软件，得到准确度的平均值、标准差以及与待检验本文模型的配对 t 检验 p 值，结果如表 2 所列。从表 2 中可以看出，其他方法与本文所提方法的配对 t 检验 p 值均小于 0.05，因此可以验证本文所提方法在统计上显著不同于其他方法。

表 2 配对 t 检验结果

模型	平均值	标准差	配对 t 检验 p 值
候选答案长度最长	21.80	3.293	0.000
候选答案长度最短	31.90	2.998	0.000
候选答案之间最不相同	29.10	4.057	0.000
候选答案与问题最相似	33.90	3.900	0.000

4.6 数据分析

为了深入分析以句子为单位的文章向量表示的效果弱于以单词为单位的文章向量表示的原因，随机从所有样本中抽取 100 个样本对问题进行统计分析。

4.6.1 问句分类

通过仔细分析 100 个样本的问题，得出问题的样本大致分为如下几类(如果一个样本满足多个类别，则将其分类到下面最先出现的类别)：

1)答案匹配式问题：问题的答案在文章中出现，并且答案关键字是判定候选答案的主要决定因素，关键字直接匹配即可得到正确答案，不需要推理。

2)意义问题：问题要求在全面理解文章的前提下总结出结果，不能只是简单地进行关键字匹配。

3)位置问题：问题要求简单的推理，类似于机器阅读理解任务中的人工合成问答。

4)时间问题：问题的答案是准确的时间或者频率、日期等。

5)判断问题：问题要求从候选答案中选出符合文章内容的答案，即每一个选项都是一个陈述句，陈述一个事实，需要根据文章来判断正确的事实输出。

表 3 列出了 100 个样本中各类问题的数目分布，表 4 列出了不同类问题的样例。

表 3 100 个样本中各问题类别及数目

问题分类	数目/个
答案匹配式问题	21
意义问题	45
位置问题	20
时间问题	9
判断问题	5

表 4 不同类问题及样例

问题分类	样例
答案匹配式问题	Who did the man buy the books for? What is the man's favorite sport in summer?
意义问题	What do we learn from the conversation? What does the man mean?
位置问题	Where is the conversation most probably taking place? Where are the man and woman going?
时间问题	what time did Suzy leave home? How often will the woman's son have piano lesson from next week on?
判断问题	According to the man, which option is more likely to be correct? According to the speakers, which of the following statements is not appropriate?

4.6.2 实验分析

表 5 列出了将 100 个不同类别问题的样本输入预先训练好的模型后的实验结果，评估结果以准确率表示。

(下转第 122 页)

notation[J]. Computers&Mathematics with Applications, 2013, 66(10):1920-1934.
[9] SALEM Y B, NASRI S. Automatic recognition of woven fabrics based on texture and using SVM[J]. Signal, Image and Video Processing, 2010, 4(4):429-434.
[10] SHIN J, KIM H J, KIM Y. Adaptive support vector regression for UAV flight control[J], Neural Network, 2011, 24(1): 109-120.
[11] AUEPANWIRIYAKUL S, SUMONPHAN E, THEERA-UMPON N, et al. Automatic Nevirapine concentration interpretation system using support vector regression[J]. Computer Methods and Programs in Biomedicine, 2011, 101(3):271-281.
[12] LEVIS A A, PAPAGEORGIOU L G. Customer demand forecasting via support vector regression analysis[J]. Chemical Engineering Research and Design, 2003(83):1009-1018.
[13] GAOA C, BOMPARD E, NAPOLI R, et al. Price forecast in the competitive electricity market by support vector machine[J]. Physica A: Statistical Mechanics and its Applications, 2007, 382(1):98-113.
[14] VAN GESTEL T, SUYKENS J A K, BAESTAENS D E, et al. Financial time series prediction using least squares support vector machines within the evidence framework[J]. IEEE Transcations on Neural Networks, 2001, 12(4):809-821.
[15] PAI P F, LIN K P, LIN C S, et al. Time series forecasting by a seasonal support vector regression model[J]. Expert Systems with Applications, 2010, 37(6):4261-4265.
[16] LI X, HU B, DU R. Predicting the parts weight in plastic injection molding using least squares support vector regression[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Review, 2008, 38(6):827-833.
[17] KUHN H W, TUCKER A W. Nonlinear Programming[C]// Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1951:481-492.
[18] MERCER J. Function of positive and negative type and their connection with the theory of integral equations[J]. Philosophical Transactions of the Royal Society, 1909, 209:415-446.

(上接第 96 页)

表 5 不同问题类别下基于单词和基于句子的实验

问题分类	基于单词		基于句子	
	正确数/个	准确率/%	正确数/个	准确率%/
答案匹配式问题	17	81	16	76.2
意义问题	20	44.4	23	51.1
位置问题	11	55	8	40
时间问题	5	56	3	33.3
判断问题	2	40	3	60

通过实验分析可以得出结论，基于单词的文章语义向量表示相对于基于句子的文章语义向量表示在位置问题、时间问题和答案匹配式问题上占据优势，能取得更好的结果；而基于句子的文章语义向量表示在意义问题、判断问题这类需要完整把握文章语义并进行总结的任务上的效果较好。

结束语 本文提出了双线性函数注意力 Bi-LSTM 模型，并将其应用于机器阅读理解中的备择答案预测任务，使用多层注意力转移的推理结构后，在四六级听力数据集上的准确率达到了 54.65%，较好地完成了该任务；此外，对任务中的问题进行详细分析，对比了基于单词的文章语义向量表示和基于句子的文章语义向量表示，发现基于单词的文章语义向量表示能较好地把握文章中的关键字，在关键字选择推理型的任务中性能较好；而基于句子的文章语义向量表示能较好地整体把握文章的语义信息，较适用于需要通读整篇文章并作出总结的任务。

虽然本文的模型能较好地理解文章的语义，但是对于需要根据文章信息抽象总结出关于问题的答案的任务的表现仍有较大欠缺，今后的工作将考虑采用叠加记忆网络的方法[18]来提升模型语义理解的能力。

参 考 文 献

[1] BURGES C J C. Towards the machine comprehension of text: An essay: MSR-TR-2013-125[R]. 2013.
[2] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks[J]. arXiv preprint arXiv: 1506.02075, 2015.
[3] KUMAR A, IRSOY O, SU J, et al. Ask me anything: Dynamic memory networks for natural language processing[J]. arXiv preprint arXiv: 1506.07285, 2015.
[4] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(8):2493-2537.
[5] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks[C]// Advances in Neural Information Processing Systems. 2015:2440-2448.
[6] KALCHBRENNER N, GREFENSTENTTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv: 1404.2188, 2014.
[7] 尹宝才，王文通，王立春. 深度学习研究综述[J]. 北京工业大学学报，2015, 41(1):48-59.
[8] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv preprint arXiv: 1509.00685, 2015.
[9] CHEN D, BOLTON J, MANNING C D. A thorough examination of the cnn/daily mail reading comprehension task[J]. arXiv preprint arXiv: 1606.02858, 2016.
[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[11] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473, 2014.
[12] KADLEC R, SCHMOID M, BAJGAR O, et al. Text Understanding with the Attention Sum Reader Network[J]. arXiv preprint arXiv: 1603.01547, 2016.
[13] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C]// Advances in Neural Information Processing Systems. 2015:1693-1701.
[14] 北京大学数学系前代数小组. 高等代数(第四版)[M]. 北京：高等教育出版社，2013.
[15] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv: 1508.04025, 2015.
[16] http://nlp.stanford.edu/data/glove.6B.zip.
[17] PENNINGTON J, SOCHER R, Manning C D. Glove: Global Vectors for Word Representation[C]// EMNLP. 2014, 14:1532-1543.
[18] WESTON J, CHOPRA S, BORDES A. Memory networks[J]. arXiv preprint arXiv: 1410.3916, 2014.