文章编号：1003-0077(2018)11-0135-08

D-Reader：一种以全文预测的阅读理解模型

赖郁婷[1]，曾倪颖[1]，林柏诚[2]，萧瑞辰[2]，邵志杰[1]

（1. 台达电子股份有限公司 台达研究院，台湾 台北；
2. 台达电子股份有限公司 知识管理部，台湾 台北）

摘　要：该文针对 2018 机器阅读理解技术竞赛提出一个基于双向注意流(BiDAF)BiDAF 的阅读理解模型，实作于 DuReader 中文问答数据集。该文观察到基线系统采用与问题最相近的段落，作为预测的筛选条件，而改以完整段落来预测答案，结果证实优于原方法。并利用 fastText 训练词向量以强化上下文信息，最后通过集成学习优化结果，提升效能与稳定性。此外，针对 DuReader 的是非类题型，该文集成两个分类模型，分别基于注意力机制(attention)与相似性机制(similarity)来预测答案类别。该模型最终在“2018 机器阅读理解技术竞赛”的评比中得到了 ROUGE-L 56.57 与 BLEU-4 48.03。

关键词：机器阅读理解；DuReader；双向注意流；集成学习

中图分类号：TP391　　**文献标识码**：A

D-Reader：A Reading Comprehension Model by Full-text Prediction

LAI Yuting[1], TSENG Yiying[1], LIN Pocheng[2], HSIAO Vincent[2], SHAO Chihchieh[1]

(1. Delta Research Center, Delta Electronics, Taipei, Taiwan, China;
2. Delta Management System, Delta Electronics, Taipei, Taiwan, China)

Abstract: This paper proposed a reading comprehension model based on Bi-Directional Attention Flow (BiDAF) network. It predicts the answers using complete paragraphs and the results outperformed baseline system. The fastText is applied to train word embedding to include contextual information. The ensemble learning is adopted to improve performance and stability. Specifically, for the Yes/No questions, this paper ensembles two classification models based on attention and similarity mechanism, respectively. The model reaches a ROUGE-L score of 56.57 and a BLEU-4 score of 48.03 in the MRC 2018.

Keywords: machine reading comprehension; DuReader; BiDAF; ensemble learning

0　引言

机器阅读理解是近年来自然语言处理的重点研究项目之一，我们相信当机器具备高水平的阅读理解能力时，将能大幅提升数据及知识检索的效率。近年来，多个机器阅读理解数据集的发布使得机器阅读理解的研究大幅增加，常见的任务形式包含填空题、选择题与简答题。其中，简答题最为接近实际的应用情境，相关的英文数据集有 SQuAD[1]、MS MARCO[2]，中文数据集则有 DRCD[3] 和 DuReader[4]。

本文描述为了 2018 年举办的机器阅读理解技术竞赛所建构的模型，该竞赛采用 DuReader 数据集，其题型为简答题，每个问题提供最多五个文章段落，及人工整理的答案。本文基于经典模型 BiDAF[5] 进行数据分析与系统改良，提交机器阅读理解模型 D-Reader。我们的方法加入了预训练的词向量，并组合多次训练的模型成为一个集成模型。也针对训练数据做了预处理及筛选，以确保训练数据质量，并对预测结果进行标点符号正规化与是非题分类处理，以提高答案分数。

收稿日期：2018-06-14　**定稿日期**：2018-09-05

本文结构如下，第 1 节介绍数据与预处理方法，第 2 节介绍本文使用的模型及实现细节，第 3 节介绍实验环境及实验结果，第 4 节为分析与讨论，第 5 节总结本文内容与发现。

1 数据

1.1 数据集描述

本文的实验数据采用的是 2018 年举办的机器阅读理解技术竞赛中公开的 DuReader 数据集，此数据集包含 30 万个问题，每个问题对应 5 个候选文档及人工整理的答案，所有的问题与内文都来自于真实的数据——百度搜索引擎数据和百度知道问答社群。

百度数据集可以从两个方面来分类：问题类型与观点。第一类是问题类型，DuReader 将问题类型分成 Entity（实体）、Description（描述）和 YesNo（是非）。实体类问题，其答案都是单一确定的回答或是一连串的字词，例如："三国演义的作者有谁?"；对于描述类问题，其答案长度较长，是多个句子的总结，是一种典型的 how/why 的问题，例如："如何在计算机安装 Linux 系统?"；对于是非类问题，其问题比较简单，通常回答是或否，比如："怀孕可以吃姜黄吗?"

第二类是观点，即回答是事实（Fact）还是观点（Opinion），通过两个划分方法，DuReader 的问题类型总共可以分成六类，如表 1 所示。

表 1 DuReader 问题类型

Type	Fact	Opinion
Entity	哪一年举办北京奥运?	你喜欢三国演义哪个角色?
Description	怎么买去上海的火车票?	百度的员工福利如何?
Yes-No	孕妇可以吃姜黄吗?	王者荣耀好玩吗?

1.2 数据预处理

数据预处理采用 DuReader 数据集提供的分词。另外，在数据集中，每一个问题匹配到多个参考答案。因此，我们对每一个参考答案皆取文档中最近似的区段，故一个问题会根据每一答案产生数笔训练数据。前述处理后，训练数据会比原先的 30 万笔还要多，预期扩充训练数据将会使准确率提升。参考答案与文章段落的相似度计算使用 F1-score，计算预测答案与参考答案的平均重叠次数。

由于 DuReader 数据集的答案为人工产生，可能无法在文档中找到准确对应的句子，故滤除与标准答案之 F1-score 低于 0.7 的数据，以保持训练数据质量。同时，由于时间与设备的限制，本文并未使用完整资料，而是使用 347 723 个问题作为训练数据。

2 方法

本文的系统架构如图 1 所示。首先训练词向量，并基于词向量训练 BiDAF 模型，组合 6 个单一模型为一个集成模型，最后进行后处理，并对是非题的答案进行分类。下面将介绍各步骤的细节。

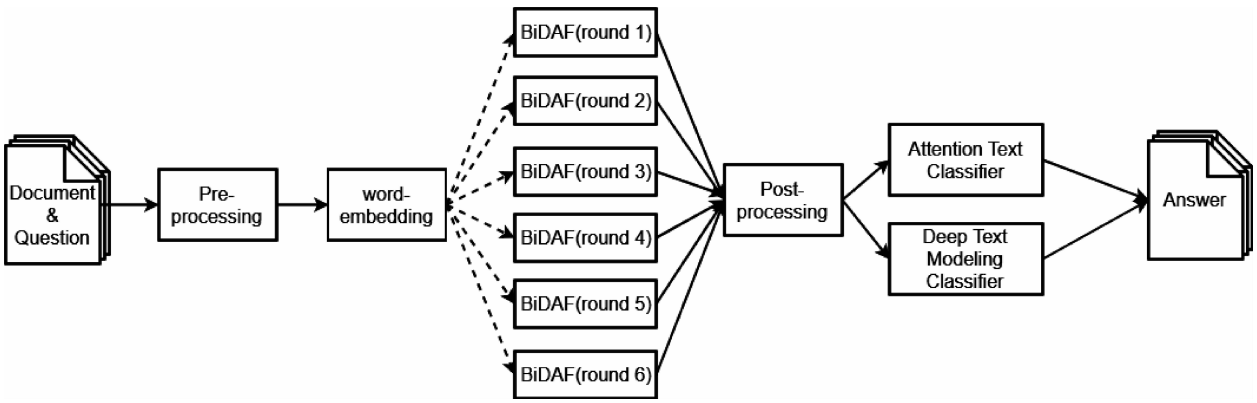

图 1 模型架构

2.1 词向量

本文使用 Joulin 等于 2016 年提出的 fastText[6] 模型进行词向量训练。此模型基于 Word2Vec[7]，通过上下文的信息来训练词汇的语意表示，并同时考虑子词的信息，将词汇中 N-gram 的子词向量加总作为该词汇向量。此作法有别于过往，能获取未登录词之词向量。此外借助子词的信息，也能有效提升低频词的词向量质量。fastText 的训练也相当快速，是当前的主流方法。

我们选用其中的 Continuous Bag-of-Word (CBOW) 算法，其以中心词汇的前后文词汇来预测中心词汇，在优化语言模型的同时更新词向量。

2.2 BiDAF

Bi-Directional Attention Flow (BiDAF)[5] 是由 Minjoon Seo 等发表于 2017 年的一个分层多阶段的训练网络。其引入不同级别的文章粒度，包含字符级及词级，对段落上下文进行模型的训练。并计算问题到内文与内文到问题之间两种关注(Attentions)，来获得 query-aware 的特征向量，最后使用双向 LSTM[8] 进行语义信息的聚合，得到答案的开始位置以及结束位置为预测结果。

BiDAF 网络包含六层：

① **Character Embedding Layer**：使用 CNN 将问句和内文的每个字符映像到一个多维向量空间。

② **Word Embedding Layer**：将问句和内文的每个词映射到一个 300 维的向量空间，使用的是前面提到的预先训练好的 fastText 的 CBOW 词向量模型。双层 Highway Network 会形成两个一维矩阵，包含内文的矩阵 X 和问句的矩阵 Q。

③ **Contextual Embedding Layer**：将 X 和 Q 向量分别输入一个双向的长期短期记忆网络(LSTM)[12]，并连接双向 LSTM 的输出，捕捉 X 和 Q 各自的特征来优化向量。此层输出为两个二维的矩阵，内文的矩阵 H 和问句矩阵 U。

④ **Attention Flow Layer**：将向量 H 和向量 U 链接，做 context-to-query 以及 query-to-context 两个方向的关注 (Attention)，输出为内文中的每个单词的查询感知特征向量(query-aware vector)，以及前一层传过来的内文与问句向量。双向关注做法是，先计算矩阵的相似性，利用内文和问句的相似度矩阵 $S\in R^{T*J}$，相似度计算方法如式(1)所示。

$$S_{tj} = \alpha(H_{:t}, U_{:j}) \in R \tag{1}$$

其中，

$$\alpha(h,u) = w_{(S)}^{T}[h;u;h\odot u] \tag{2}$$

S_{tj} 表内文的第 t 个字和问句的第 j 个字的相似度，α 是一个可训练的 scalar function，$H_{:t}$ 为 H 的第 t 列向量，$U_{:j}$ 为 U 的第 j 列向量，w 是一个可训练的权重向量，$\odot$ 为逐元素的乘积，“;”表示在向量上做拼接，计算后得到双向的关注向量 S。

1) Context-to-query Attention (C2Q)：计算对每一个问句的词而言，哪些内文的词与它最相关。对前面得到的相似度矩阵 S 做 softmax 归一化，得到内文对问句的关注权重 a，然后计算问句的向量加权和得到 $\widetilde{U}$。

$$\widetilde{U}_{:t} = \sum_{j} a_{tj} U_{:j} \tag{3}$$

$$a_t = softmax(S_{t:}) \in R^{J}$$

其中，a_t 表示第 t 个内文的词对问句的词的关注权重。

2) Query-to-context Attention (Q2C)：计算对每一个内文的词而言，哪些问句词与它最相关，当作为此问句的关键回答。取得相似性矩阵每列的最大值，并做 softmax 得到关注权重 b，即：

$$b = softmax(max_{col}(S)) \in R^{T} \tag{4}$$

归一化计算关注的内文向量。

$$\hbar = \sum_{t} b_t H_{:t} \in R^{2d} \tag{5}$$

其中，$\hbar$ 表示上下文中关于查询的最重要单词的加权总和，计算 T 次后得到 $\widetilde{H}\in R^{2d}$。最后将内文向量和关注向量组合产生 G，其中每个列向量可以被认为是每个内文词的查询感知特征向量(query-aware vector)，G 的定义为：

$$G_{:t} = \beta(H_{:t}, \widetilde{U}_{:t}, \widetilde{H}_{:t}) \in \mathrm{R}^{\mathrm{d_G}} \tag{6}$$

其中，$G_{:t}$ 表第 t 个列向量，对应于第 t 个内文的词，β 为一个任意可训练的向量函数，$\mathrm{d_G}$ 是 β 的输出维度。β 采用的方法是如上面 α 所述的拼接方式。

$$\beta(h,\bar{u},\hbar) = [h;\bar{u};h\odot\bar{u};h\odot\hbar] \in R^{8\mathrm{d}*\mathrm{T}} \tag{7}$$

⑤ **Modeling Layer**：建模层的输入为 G，对 G 做编码，经过双向 LSTM 后得到 $M\in R^{2d\times T}$，M 的每个列向量包含关于整个内文段落和问句的词的交互信息。

⑥ **Output Layer**：使用上一层的 M 做分类得到内文每个位置为起始位置的机率 p^1，然后将 M 输入双向 LSTM 得到 M^2，再将 M^2 分类得到结束位置的机率 p^2。

$$p_1 = softmax(W_{(p^1)}^{T}[G;M]) \tag{8}$$

$$p_2 = softmax(W_{(p^2)}^{T}[G;M^2]) \tag{9}$$

训练：其中 W 是一个可训练的权重向量，定义训练损失函数为真实答案的开始和结束的负对数概率总和，并对所有例子取平均值。

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N}\log(p_{y_i^1}^1) + \log(p_{y_i^2}^2) \quad (9)$$

其中，θ 是可训练的权重的集合，N 是数据集的实例数量，y_i^1 和 y_i^2 分别是第 i 个实例的真实的开始和结束位置，p_k 表示向量 p 的第 k 个值。

测试：选择最大的 $p_k^1 p_l^2$，$(k \leqslant l)$，作为答案。

本文以 Wei He[4] 等人实现的 BiDAF 作为基线系统。在该程序代码中，在训练与测试阶段，对每篇文章挑选出最具代表性的一个段落，以改良效能。其挑选的方法为，在训练阶段，比较答案与段落的 recall。而在测试阶段，因答案不可取得，则是与问题进行比较。然而，我们发现在测试阶段若以段落与问题的 recall 来筛选，将会导致许多正确答案所在的段落落选，反而选择复诵问题但无内容的段落。另外也发现，有部分文章被切割为数个段落，若只取一个段落，将取到不完整的文章。因此，为了提升召回率，我们将文章段落以句号串接起来，以整篇来预测答案。

2.3 集成模型

本节将 6 个 BiDAF 单一模型的开始与结束位置的机率取平均值，再计算机率乘积最大的区间作为答案，如果候选答案区间为空或为单一句号，则视为无效答案，将跳过并取下一个答案直至找到有效答案。

2.4 后处理

由于在 2.2 节中以句号串接段落文章，在此处将清除多余的标点符号和移除换行符号“\n”和“\r”，并于句尾补上句号，使答案句更加完整。

2.5 是非题答案分类

因 MRC 2018 主办方规范的评价指标，增加了对正确识别是非题答案类别的得分奖励。故我们对 BiDAF 模型预测过后的是非题结果进行分类。是非题答案共有四个类别：Yes、No、Depends 和 No_Opinions。

本文的分类模型基于 LSTM 设计了两种不同的模型架构：Attention Text Classifier 和 Deep Text Modeling Classifier，前者采用注意力机制，后者采用相似度机制。因两种模型侧重的特征不同，该分类模型采用集成的方式组合两者，以提升模型的泛化能力。

两个分类模型都采用相同的预处理动作，先对问题与答案进行分词，使用清华大学推出的中文词法分析工具包 THULAC[9]，其在简体中文分词中具有准确率高及效能佳的特点。

2.5.1 Attention Text Classifier

此模型分成五个部分：Embedding Layer、Bi-LSTML Layer、Attention Layer、Merge Layer 与 Softmax Layer。模型的架构如图 2 所示。

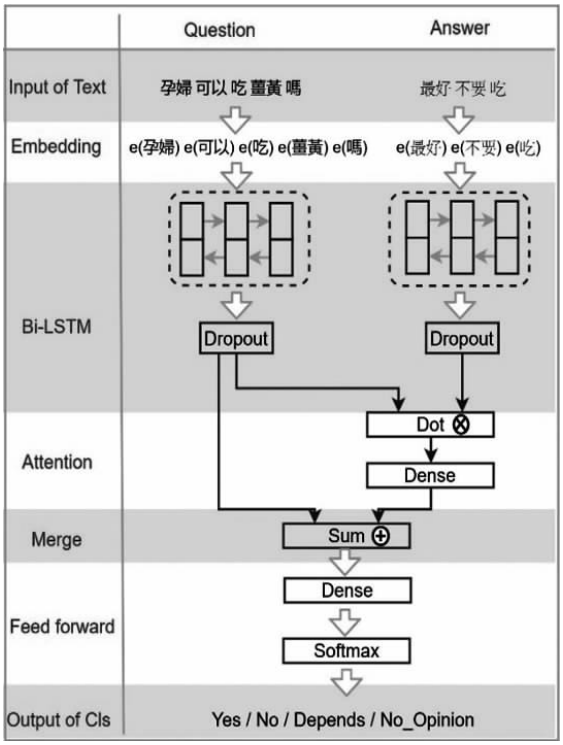

图 2 Attention Text Classifier 模型

① **Embedding Layer**：使用 fastText，以 DuReader 的数据集训练一个 300 维词向量模型。

② **BiLSTM Layer**：利用 LSTM 模型累加的线性形式，处理序列数据的信息，避免梯度消失的问题也能学习长周期的信息。将 Question 和 Answer 分开表示，透过双向 LSTM 结合上下文信息，分别学习 Question 和 Answer 的表示向量，分别将两个表示向量传递给后面。

③ **Attention Layer**：透过注意力(Attention)机制，增强关联性较强的词权重并降低关联性较低的词权重，将 Question 和 Answer 的表示向量，采用点积(Dot)方式进行计算，产生答案对于问题的注意力(Attention)的表示向量。

④ **Merge Layer**：保留问题的信息并加入特定词汇的权重，把 Attention 及 Question 的表示向量

进行加总运算，将结果传递给后面。

⑤ **Feed forward Layer**：使用 Softmax 回归模型，针对 Merge Layer 传递过来的信息进行学习，计算待分类数据归属于各个类别的机率，Softmax 回归模型是 Logistic 回归模型的一种形式，拥有良好的数学特性。

模型的参数设定为：采用 Adam 算法进行优化、词向量维度为 300，batch size 设定为 256，LSTM units 设定为 128，Hidden Layer Number 设定为 1，dropout rate 设定为 0.3。

此模型在开发集的正确率可达 72.71%。

2.5.2　Deep Text Modeling Classifier

此模型参考 Basant Agarwal[10] 等人提出的分类模型修改而成，分为五个部分，依序为 Embedding layer，CNN layer，RNN layer，Interaction layer，Feed forward layer，模型的架构如图 3 所示。

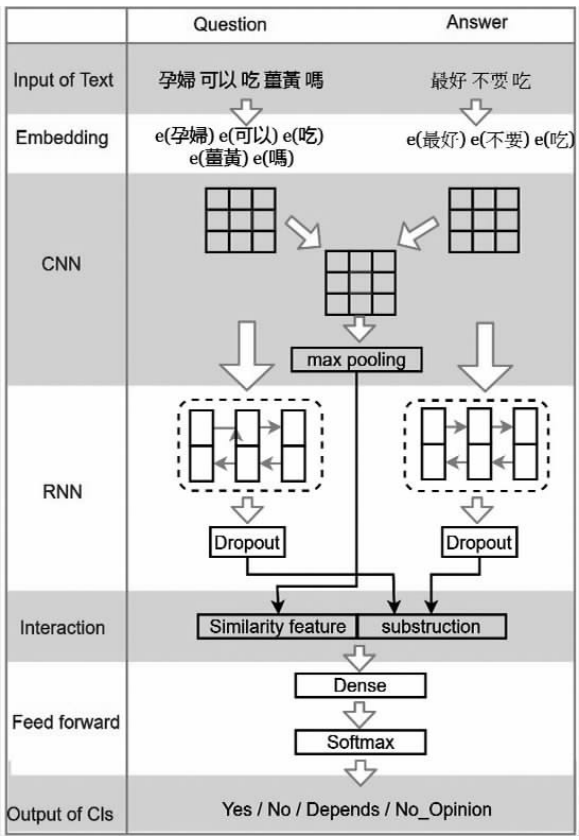

图 3　Deep Text Modeling Classifier 模型

① **Embedding layer**：使用 Word2Vec，以 DuReader 的数据集训练一个 300 维词向量模型。

② **CNN layer**：CNN 通过不同大小的 filter 撷取重要的特征，对于抽取局部特征有优异的表现。使用不同长度的 filter 同步进行卷积，视为不同长度的 N 元组语意信息，最后通过 Max pooling 将卷积后的重要信息撷取出来。

③ **RNN layer**：上述 CNN 的输出除了将重要信息撷取出来外，同时也保留文字的顺序关系，因此将之作为 LSTM 的输入，将序列数据基于文字顺序迭加以获取文字语意信息。两个输入字符串分别经过 CNN，RNN layer 后相减，此方法可视为句对间语意的差异信息。

④ **Interaction layer**：将句对的词向量作内积，内积可作为向量在另一向量的投影，因此通过词向量的内积了解句对间词汇的相似程度，视之为相似矩阵 Similarity Matrix，最后将结果经过 CNN 的特征撷取来表示句对间重要的语意信息。

⑤ **Feed forward layer**：将前述方法所得之结果串联后通过 Feed forward layer，通过 Softmax 来仿真各标签的机率，采用 Cross Entropy 作为损失函数，最后通过 Adam 进行参数更新。

3　实验

3.1　系统运行的环境及硬件条件

本节实验以 Wei He[4] 等人实现的 BiDAF 为基线，进一步组合我们提出的改良方法。实验中 BiDAF 模型的超参数设置如下：batch_size 设为 64，dropout_keep_prob 设为 1，embed_size 设为 300，epochs 次数为 2，hidden_size 为 150，learning_rate 为 0.001，max_a_len 设定为 250，max_p_len 设定为 500，max_p_num 设定为 5，max_q_len 设定为 60。

实验环境为：CPU Intel Core E5-2698；GPU NVIDIA DGX-1 搭载 Tesla V100；显存 128GB；操作系统为 64 位元 Ubuntu 16.04 LTS。

3.2　实验结果

实验以 MRC 2018 主办方规范的 ROUGE-L[11] 及 BLUE-4[12] 作为评价指标，并以 ROUGE-L 为主要参考指标。主办方适当增加了正确识别是非题的答案类型及匹配实体的得分奖励，以弥补传统 ROUGE-L 和 BLEU-4 指标对是非题和实体类型问题评价不敏感的问题。

3.2.1　后处理实验

本节比较答案后处理，包含是非题答案分类及标点正规化的效果，具体说明及参数设置见节 2.4 与 2.5，实验结果如表 2 所示。加入了 Yes/No 标

签(cls)后使 ROUGE-L 提高了 0.42%,BLEU-4 提升0.35%。标点正规化(norm)则使 ROUGE-L 再提高了 0.16%,BLEU-4 的提升则更为显著,有0.56%,这可能是源于 BLEU-4 在短答案上的波动较大的特性。在本节后续的实验中,都会加上 cls+norm 的后处理。

表 2 后处理实验数据

	ROUGE-L/%	BLEU-4/%
Baseline	43.93	38.83
Baseline+cls	44.35	39.18
Baseline+cls+norm	44.51	39.74

3.2.2 词向量比较

基于上一节最佳的后处理设置,我们进一步实验不同的词向量。训练语料使用 DuReader 的全部分词文本。算法为 CBOW,模型窗口设定为 3,维度设定为 300,学习速率设定为 0.5,训练 5 轮。N-gram 的最大值设定为 2,最长子词则设定为 4 字符,最小为 1 字符。损失函数选用 hs。

结果如表 3 所示,可以看出使用 fastText 算法预训练词向量较随机词向量效果有显著提升,ROUGE-L 提高了 3.93%,BLEU-4 提高了 2.57%。

表 3 词向量实验数据

	ROUGE-L/%	BLEU-4/%
Random Vector	44.51	39.74
fastText CBOW Vector	48.44	42.31

3.2.3 预测方式比较

基线系统为了提升效能,以先挑选文章中最具相关的段落来加速预测。而我们使用以句号串接的完整数据来预测,实验结果见表 4,在 ROUGE-L 上提高了 7.23%,BLEU-4 提高了 6.75%,进步非常显著,可见基线系统以问题与段落相似度来代表其相关度的假设并不合理。这也是本文模型分数大幅提升的关键,虽然会使预测阶段的运算时间拉长,但通过平行处理,运算时间约两小时,并不至于太慢。

表 4 预测方法实验数据

	ROUGE-L/%	BLEU-4/%
基线系统段落筛选方法	48.44	42.31
不筛选段落	55.67	49.06

3.2.4 集成模型

表 5 是不同集成模型的实验结果,此实验中的每个单一模型所使用的参数设置皆相同。从实验数据可看出,集成模型确实能优化结果,并且表现比单一模型稳定,由 7 个模型的集成在 ROUGE-L 上较单一模型提高了 0.96%,可推测越多的集成模型表现越佳。此实验中的模型以相同权重集成,我们认为线性加权应可再提升结果。

表 5 也针对不同前处理的训练数据进行实验,"标准数据"指的是原始训练数据集,经筛选后有 242 132 个问题。"扩充数据"则是在 1.2 节重新预处理过所得之数据,共有 347 723 个问题。

因上传次数限制,没有足够数据进行同基准的比较,但可看出扩充数据在 ROUGE-L 上有显著提升效果,但其 BLEU-4 下降。我们推测,扩充数据也可能混淆训练方向,会有同一个题目却对应到不同的答案的矛盾情形。又因时间关系,扩充的数据并不完整,这也可能是导致分数下降的原因之一。

表 5 集成模型实验数据

	ROUGE-L/%	BLEU-4/%
标准数据(Single)	55.77	48.73
标准数据[Ensemble(3 runs)]	56.31	50.82
标准数据[Ensemble(5 runs)]	56.73	50.68
标准数据[Ensemble(7 runs)]	56.73	50.70
扩充数据[Ensemble(2 runs)]	56.71	49.47
扩充数据[Ensemble(6 runs)]	**57.19**	48.88

4 分析与讨论

由于该数据集答案为人工产生,以区段的方式难以表示结果,因为答案可能摘要自原段落的不同位置。在实体类问题中尤为明显(表 6)。我们认为也许可以尝试生成式模型,或是将多个答案进行合并。

表 6 摘要式答案之范例

问题	2017 有什么好看的小说
段落	1.《将夜》作者:猫腻(起点白金作家) 简介:与天斗,其乐无穷。故事概要:主角宁缺带…… 2.《择天记》作者:猫腻 too 相对将夜,这是一本正在写的新书,值得一看。…… 3.《冒牌大英雄》作者:七十二编。这本书……
答案	1.《将夜》2.《择天记》3.《冒牌大英雄》4.《无限恐怖》5.《恐怖搞校》6.《大国医》7.《龙魔导》。

预测答案时，我们使用以句号串接的方式预测完整数据。这个做法虽能一定程度上提高命中率及改善段落被错误切割的情形，但在段落很短时，很大机率会取到跨段落的答案。表 7 中，框起来的句号为串接处，问题 1 串接了数个段落的答案，使其更完整，但问题 2 却因为串接了两个独立答案，计分受到长度惩罚。

此外，可以观察到被误切为数段的多为列点或步骤描述，也许可以此为线索合并段落，将有类似格式的内文在预处理时串接起来，以修正数据。至于预测时则可改用特殊符号为分隔符，降低预测到跨文章结果的机率。

表 7　串接段落对答案的影响之范例

问题 1	怎么清除 dnf 缓存
预测答案 1	1、打开电脑，进入桌面，鼠标双击“这台电脑”并进入。2、选择要清理的磁盘，如 C 盘，鼠标右击 C 盘，点击“属性”。3、进入“属性”窗口后，点击右下角的“磁盘清理”。4、点击“磁盘清理”后，会弹出一个窗口扫描此磁盘下的各种垃圾文件和缓存，如果磁盘文件较多，扫描的时间会有点长一些。
问题 2	遣倦
预测答案 2	是“缱绻”，形容情投意合，难舍难分；缠绵。缱绻。应该这样写。形容情投意合，难舍难分，缠绵。

最后，DuReader 数据集中三种问题类型的特性迥异，我们认为应对三种类型分别训练。最初虽曾尝试训练个别模型，但性能变差，推测原因是数据量下降，故未继续研究这个方向。但也许可以尝试多任务学习，通过共享信息来放大数据，降低拆开训练的不利影响。

5　结论

机器阅读理解是近年来自然语言处理的研究重点，随着更多中文阅读理解数据集的发布，中文阅读理解的技术将能更好地发展。本文为 DuReader 数据集设计一基于 BiDAF 的阅读理解系统。除改良数据前处理及使用 fastText 预训练词向量，亦发现基线系统为简化运算而以问题与段落之相似度筛选文本的假设并不合理，故改用全文预测，获得大幅度的性能提高，为本文系统分数提高的主因。本文亦以集成学习降低单一模型的偏差，使模型效果更佳，也更稳定，有效地提升正确率。并使用两种分类模型，分别基于注意力与相似性，对是非题答案进行分类。

本文实现的方法在 MRC 2018 的评比中得到了 ROUGE-L 56.57％与 BLEU-4 48.03％的结果，说明该方法是可行且有效的。

本文研究也存在不足之处，首先，由于时间与设备限制，资料扩展并没有完整完成。对于预测时的段落串接，虽然大幅改善结果，但也造成一些多抓的情况，应可再尝试其他的处理方式。另外，未能针对各问题类型及不同资料来源进行实验，也是可惜之处，我们认为 DuReader 的三种问题类型各有不同特性，这可以是今后继续研究的方向。

参考文献

[1] Pranav Rajpurkar, et al. SQu AD: 100,000+ questions for machine comprehension of text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016.

[2] Tri Nguyen, et al. MS MARCO: A human generated machine reading comprehension dataset. CoRR, abs/1611.09268, 2016.

[3] C Shao, et al. 2018 DRCD: a Chinese Machine Reading Comprehension Dataset[J]. arXiv pre-print arXiv: 1806.00920, 2018.

[4] Wei He, DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications [J]. arXiv preprint arXiv: 1711.05073, 2018.

[5] M Seo, A et al. Bidirectional attention flow for machine comprehension[J]. arXiv pre-print arXiv: 1611.01603, 2016.

[6] Piotr Bojanowski, et al. Enriching word vectors with subword information[C]//Proceedings of TACL 5: 2017: 135-146.

[7] Tomas Mikolov, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the NIPS, 2013.

[8] S Hochreiter, J Schmidhuber. Long short-term memory[J]. Neural Computation,1997: 9(8): 1735-1780.

[9] 孙茂松，等. THULAC：一个高效的中文词法分析工具包[EB/OL]. 2016. http://thulac.tinunlp.org/.

[10] Basant Agarwal, et al. A Deep Network Model for Paraphrase Detection in Short Text Messages[J]. arXiv pre-print arXiv: 1712.02820, 2017.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of Text Sum-

marization Branches Out: Proceedings of the ACL-04 Workshop, 2014: 74-81.

[12] Kishore Papineni, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

赖郁婷(1992—),硕士,主要研究领域为自然语言处理、机器学习。
E-mail: yuting.lai@deltaww.com

曾偲颖(1988—),硕士,主要研究领域为自然语言处理、语音处理、机器学习。
E-mail: yiying.tz@deltaww.com

林柏诚(1990—),硕士,主要研究领域为自然语言处理、机器学习、深度学习。
E-mail: pocheng.lin@deltaww.com

全国知识图谱与语义计算大会(CCKS 2018)在天津隆重召开

2018年全国知识图谱与语义计算大会(CCKS 2018)于8月14日至17日在天津召开,本次会议由中国中文信息学会语言与知识计算专业委员会主办,南开大学软件学院承办,天津大学智能与计算学部协办。会议主题为"知识计算与语言理解"。大会吸引了来自海内外的千余名知名学者、工业界专家和知名企业代表参加,其中主会注册人数超860人。会议回顾了知识图谱与语义计算的进展情况,探讨了领域内的新发现、新技术和新应用,旨在让社会各界了解知识图谱与与语义计算的新方向和新趋势,以推动我国语言与知识计算领域的进一步发展。

CCKS2018会议分为讲习班和主会两个阶段。8月14日至15日,中国中文信息学会《前沿技术讲习班》(ATT)第十一期在南开大学泰达学院报告厅举行。前沿技术讲习班邀请了国内外优秀青年学者及工业界专家,内容涵盖了知识图谱的推理、构建,自然语言的推理、关系抽取及知识图谱应用等方面,分别从知识图谱的构建及在实际场景中的应用等角度介绍了知识图谱的最新进展和实战经验。讲习班开班仪式由浙江大学陈华钧教授和加拿大皇后大学(Queen's University)的助理教授朱晓丹主持。加州大学圣巴巴拉分校(University of California, Santa Barbara)的助理教授 William Wang 作了题为"Deep Knowledge Graph Reasoning"的报告,阿伯丁大学(University of Aberdeen)的 Jeff Z. Pan 教授作了题为"Exploiting and Reasoning With Open Knowledge Graph"的报告,皇后大学(Queen's University)的助理教授朱晓丹作了题为"Deep Learning for Natural Language Inference"的报告,卡塔尔计算研究所(Qatar Computing Research Institute)的 Preslav Nakov 博士作了题为"Semantic Relation Extraction from Text"的报告,阿里巴巴商品知识图谱负责人张伟博士作了题为"特定领域知识图谱的构建及应用案例"的报告;科大讯飞 AI 研究院语音交互研究主管刘权作了题为"语义计算与知识问答技术在实际场景中的应用"的报告,系统地介绍了在商业领域垂直知识图谱构建和服务的实践。

8月16日至17日,CCKS主会在天津滨海新区滨海一号酒店举行,中国中文信息学会理事长、中国工程院院士方滨兴、南开大学党委副书记李义丹、天津滨海新区副区长夏青林等嘉宾出席大会并致辞。主会包括特邀报告、知识图谱相关顶级会议 Review、优秀学术论文报告、优秀测评系统报告、Poster Spot Highlight、知识图谱工业界论坛及知识图谱 Panel 等环节。特邀报告环节邀请了海内外知名学者和工业界代表介绍了学科前沿信息及重要成果,中国科学院院士张钹作了题为"知识与人工智能"的特邀报告,介绍了知识在未来人工智能发展中的重要性;语义网(Semantic Web)创始人之一 James A. Hendler 作了题为" The Semantic Web: Vision, Reality and Revision"的报告;阿里巴巴集团副总裁墙辉在"阿里巴巴生态下的知识引擎" 中介绍了阿里通用知识图谱、商品知识图谱、客服知识图谱等的构建和应用;罗马大学(Sapienza University of Rome)Roberto Navigli 教授作了题为"What can you do with multilingual knowledge graphs? Experiences at Sapienza and Babelscape"的报告。