文章编号：1003-0077(2018)11-0112-05

一种基于数据重构和富特征的神经网络机器阅读理解模型

尹伊淳，张 铭

（北京大学 信息科学技术学院，北京 100871）

摘 要：该文描述了 ZWYC 团队在“2018 机器阅读理解技术竞赛”上提出的机器理解模型。所提出模型将机器阅读理解问题建模成连续文本片段抽取问题，提出基于富语义特征的神经交互网络模型。为了充分使用答案标注信息，模型首先对数据进行细致的重构，让人工标注的多个答案信息都能融合到数据中。通过特征工程，对每个词构建富语义表征。同时提出一种简单有效的问题和文档交互的方式，得到问题感知的文档表征。基于多个文档串接的全局表征，模型进行答案文本预测。在最终测试集上，该模型获得了目前先进的结果，在 105 支队伍中排名第 2。

关键词：机器阅读理解；数据重构；神经网络

中图分类号：TP391 **文献标识码**：A

A Neural Machine Reading Comprehension Model Based on Relabeling and Rich Features

YIN Yichun, ZHANG Ming

(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

Abstract: This paper describes the model proposed in “2018 NLP Challenge on Machine Reading Comprehension” by ZWYC team. Treated the machine reading comprehension as extracting the text span from the documents, this paper proposes a feature-rich neural interaction network. In order to effectively use the information of golden answers, our model first reconstructs the data in detail so that all golden answer information can be integrated. Then a feature-rich semantic representation is built for each word. Moreover, a simple but effective network is designed for question-aware representation for each document by captuing the interaction between questions and documents. The proposed model predicts answer text based on global representations of multiple candidate documents, leading to the runner-up position among 105 teams.

Keywords: machine reading comprehension; data reconstruction; neural network

0 引言

“2018 机器阅读理解技术竞赛”旨在让机器根据问题和搜索引擎返回的相应证据文档，抽取或生成合理的答案文本，推动搜索引擎场景下智能问答技术的发展。具体而言，竞赛要求构建机器阅读理解系统，自动对给定问题 q 及搜索引擎返回来的候选文档集合 $D=\{d_1,d_2,\cdots,d_n\}$，其中 $n=5$，输出对应的答案文本 a。评测基于人工标注问题答案，采用 ROUGE-L 和 BLEU4 作为评价指标，其中 ROUGE-L 是主阅读评价指标。

本文将机器理解问题建模成文本抽取问题，认为答案文本 a 来自候选文档中某个连续的文本片段，基于问题和数据的特点构造神经网络模型。为了充分利用人工标注答案信息，本文首先对数据进行重构，使同一个问题的多个候选文档具有各自不同的答案文本；接着提出了基于语义富特征的神经交互网络，得到问题感知的文档表征；最后采用基于不同词向量的模型集成方法，进一步提升模型效果。在测试集上，本文提出的模型得到 ROUGE-L60.99 和 BLUE-4 55.93 的结果，在所有 105 支参赛队伍

收稿日期：2018-06-25 **定稿日期**：2018-07-28

基金项目：国家自然科学基金(61472006,61772039)；国家自然科学基金(91646202)；北京市科技计划新一代人工智能技术培育项目(Z181100008918005)

中排名第 2 位。

1　数据重构

给定的数据集基于真实标注答案集(一个问题往往含有多个答案),在文档中寻找相对应的匹配文本段,进行伪答案的标注。这种简单的标注方法在答案覆盖度上存在以下不足:(1)以段落为单位标注和预测,忽略了候选文档其他段落的信息;(2)所有 5 篇候选文档中只把匹配度最高的一条真实标注答案作为基准信息,忽略了其他真实标注答案的信息。

为了缓解原始标注对真实答案信息利用不足的问题,本文对给定数据集进行了重构。具体来说,本文从文档级出发,基于每个真实标注答案对文档进行匹配。每个文档选择匹配分数最大的作为伪答案片段。这样真实的标注答案信息会出现在多个候选文档中,其分别匹配不同的真实答案。本文基于 F1 的词匹配指标,对提供的训练集进行重构,同时过滤掉匹配分数小于 0.65 的答案片段。为了提高数据重构的速度,使用多进程并行处理方式。

2　模型

模型的任务可以形式化定义为,对于给定问题 $q=\{q_1,q_2,\cdots,q_{|q|}\}$ 及返回来的候选文档集合 $\{d_1,d_2,\cdots,d_n\}$,其中 d_i 由词集合 $\{p_1^i,\ p_2^i,\ \cdots,p_{|d_i|}^i\}$ 组成;模型的目标是从所有的候选文档集合 D 抽取出一个文档片段 $\{p_{\text{start}}^t,\ p_{\text{start}+1}^t,\ \cdots,p_{\text{end}}^t\}$ 作为预测答案,其中 t 指示答案文本所在文档,[start, end]分别为答案的起至位置。

本节介绍富特征的神经交互网络(图 1),整个网络分为:(1)词富语义表征层,使用特征工程建模词向量表征;(2)序列语义编码层,使用句子序列信息对词进行语义编码;(3)问题与文档交互层,旨在得到有效的问题感知文档表征;(4)答案文本预测模块,在所有候选文档上抽取答案文本 a;(5)模型集成模块。

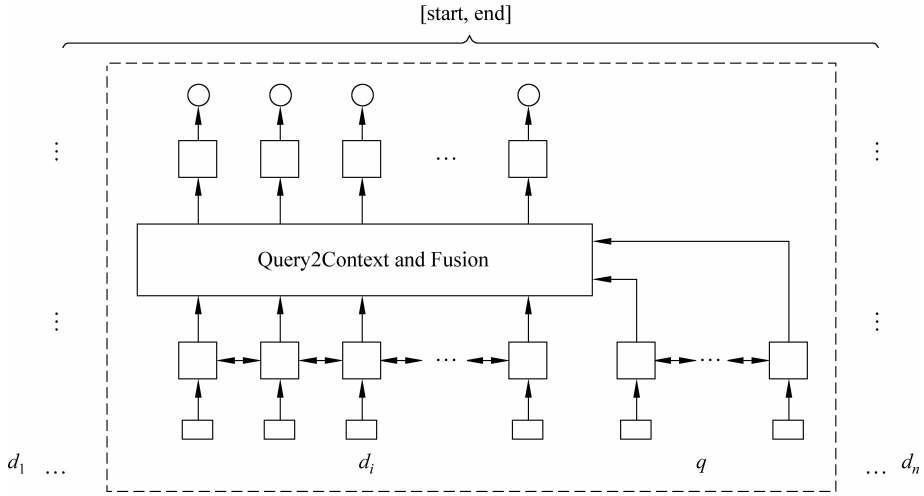


图 1　模型框架图

2.1　词富语义表征

为了得到丰富的词语义表征,本文使用了以下 8 种特征。

词向量特征:本文使用了从大规模 122GB 中文语料中使用 Word2Vec 预训练得到的 64 维词向量①。

文档排序特征:排名越靠前的文档越可能出现答案文本,因此本文使用了 $\boldsymbol{M}_{\text{rank}}\in\mathbb{R}^{5\times5}$ 待学习排序矩阵,每个位置上都对应一个 5 维的向量。

问题类别特征:考虑到问题类别对答案文本抽取影响较大,这里使用了 $\boldsymbol{M}_{\text{ques}}\in\mathbb{R}^{6\times5}$ 的待学习排序矩阵,分别对应“SEARCH _ DESCRIPTION”“SEARCH _ ENTITY”“SEARCH _ YES _ NO”“ZHIDAO_DESCRIPTION”“ZHIDAO_ENTITY”和“ZHIDAO_YES_NO”6 种问题类型,每种问题类型使用 5 维向量表示。

词性特征:本文同时使用了词性特征,统计全部语料,选取了 30 个不同词性作为特征,使用 $\boldsymbol{M}_{\text{pos}}\in\mathbb{R}^{30\times5}$ 进行表征。

① https://pan.baidu.com/s/1o7MWrnc

精确匹配特征：指示当前词是否同时存在于文档和问题中。

上下文匹配特征：指示在此前词的上下文中窗口词精确匹配的比例，例如考虑当前词上下文中的4个词，如果其中3个是精确匹配的，那么匹配比例即为0.75。

是否由数字组成：指示当前词是否由数字组成。

是否由字母组成：指示此前词是否由字母组成。

本文对问题和文档中的词采用了不同的特征表征形式，其中问题词使用了词向量特征、问题类型特征和词性特征；文档词使用了词向量特征、词性特征、排序特征、精确匹配、上下文匹配和是否为数字、字母特征。

2.2 序列语义编码

本文使用了GRU(Gated Recurrent Unit)模型在词序列中建模文档中词语义表征。GRU基于前一个词隐层表征和当前词表征 $\boldsymbol{w}_k^i$，得到当前词隐层表征 $\overrightarrow{\boldsymbol{p}_k^i}=\mathrm{GRU}(\overrightarrow{\boldsymbol{p}_{k-1}^i},\boldsymbol{w}_k^i)$。同时为了丰富语义表征，本文也使用另一个GRU模型对文档中的词进行倒序建模，得到 $\overleftarrow{\boldsymbol{p}_k^i}$。将前向隐层和反向隐层进行串接得到最终的隐层表征 $\boldsymbol{p}_k^i=[\overrightarrow{\boldsymbol{p}_k^i},\overleftarrow{\boldsymbol{p}_k^i}]$。按照相同的方法，本文也使用双向GRU模型得到问题中词语义表征 $\boldsymbol{q}_j$。

2.3 问题感知的文档表征

为了得到问题感知的文档表征，本文使用了问题到文档的注意力模型(Query2ContextAttention)，对问题表征进行注意力操作，让每个文档词能得到反馈回的注意力求和向量，接着基于双向GRU模型将词和求和向量进行文档语义和问题语义的融合(fusion)。具体而言，基于每个文档词 $\boldsymbol{p}_k^i$，我们对所有的问题词表征做如下注意力操作 $\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_{|q|}\}$：

$$\alpha_{kj}=\frac{\exp(\boldsymbol{p}_k^i\cdot\boldsymbol{q}_j)}{\sum_{j'}\exp(\boldsymbol{p}_k^i\cdot\boldsymbol{q}_{j'})} \tag{1}$$

$$\boldsymbol{c}_k^i=\sum_j\alpha_{kj}\boldsymbol{q}_j \tag{2}$$

$$\boldsymbol{p}_k^i\leftarrow[\boldsymbol{p}_k^i;\boldsymbol{c}_k^i] \tag{3}$$

其中，符号“←”表示更新操作。本文使用了直接点积计算注意力权重而非带参数化的权重计算方式，在实验中也取得了较好的效果。为了将文档和问题信息进行融合，这里使用了双向GRU模型，以新的 $\boldsymbol{p}_k^i$ 作为输入，得到的隐层表征 $\boldsymbol{h}_k^i$ 作为最终的问题感知的文档词表征。对同一个问题的不同候选文档做相同的处理，分别得到问题感知的文档表征。

2.4 答案文本预测

为了让模型做出准确的全局判断，本文串接所有候选文档的词表征 $\{\boldsymbol{h}_1^1,\boldsymbol{h}_2^1,\cdots,\boldsymbol{h}_k^i,\boldsymbol{h}_{k+1}^i,\cdots,\boldsymbol{h}_{|d_n|-1}^n,\boldsymbol{h}_{|d_n|}^n\}$，基于此使用注意力机制分别进行start和end位置概率的计算。具体操作如式(4)、式(5)所示。

$$s_k^i=\boldsymbol{v}\tanh(\boldsymbol{W}^p\boldsymbol{h}_k^i+\boldsymbol{W}^q\boldsymbol{q}^s) \tag{4}$$

$$\beta_k^i=\frac{\exp(s_k^i)}{\sum_{i'}\sum_{k'}\exp(s_{k'}^{i'})} \tag{5}$$

其中，$\boldsymbol{v}$、$\boldsymbol{W}^p$、$\boldsymbol{W}^q$ 是待学习的参数；$\boldsymbol{q}^s$ 是问题的表征，本文使用注意力机制得到式(6)、式(7)。

$$\gamma_k=\frac{\exp(\boldsymbol{q}_k\cdot\boldsymbol{u})}{\sum_{k'}\exp(\boldsymbol{q}_{k'}\cdot\boldsymbol{u})} \tag{6}$$

$$\boldsymbol{q}^s=\sum_k\gamma_k\boldsymbol{q}_k \tag{7}$$

其中，$\boldsymbol{u}$ 是待学习参数。为了训练模型，本文最小化金标准start和end位置的**negative log probabilities**。

在得到start和end的概率之后，本文基于两个概率的积，在全局文本上搜索答案文本，选择积最大者作为最终答案文本，其中设定文本的最大长度为max_a_len。

2.5 模型集成

本文假设不同语料学习得到embedding具有不同的语义知识，采用基于embedding的方法[1]进行模型集成。具体而言，本文使用了：(1)从大规模122GB中文语料中使用Word2Vec[2]预训练得到的64维词向量 $\boldsymbol{E}^1$；(2)将训练数据集作为语料采用fasttext[3]预学习得到的128维词向量 $\boldsymbol{E}^2$。模型集成采用两个64维向量学习得到的模型和两个128维词向量学习得到的模型，最终的起止位置的概率值采用4个模型结果的平均值。

3 实验

3.1 实验设置

实验采用了比赛提供的数据集(此数据集是之前DuReader[4]公开数据集的超集)，总共30万个问题，其中训练数据27万；开发数据1万；测试数据2

万。直接使用数据集提供分词结果。

本文的模型构建采用 Tensorflow1.4.1、Python2.7.12，整体调优在开发集上进行，参数值设置见表 1；模型采用 Adam 优化模型参数。实验运行的硬件条件为：TITAN12GB 显存；内存为 132GB。

得到结果文本之后，去掉了 HTML 符号标记的冗余部分，作为最终的预测文本。

表 1　模型参数设置

参数	数值
Hidden_size	156
Learning_rate	0.001
Batch_size	32
Max_a_len 最大答案长度	300
Max_p_len 最大文档长度	1000
Epoch 训练轮数	5

3.2　实验结果

基于不同词向量的模型结果如表 2 所示，可以发现我们的模型显著高于 baseline，模型集成对效果也有一定的提升。最终，我们的模型在测试集上的排名第 2。

表 2　模型结果

模型	ROUGE-L	BLUE-4
开发集		
基于 $\boldsymbol{E}^1$ 模型	50.90	47.56
基于 $\boldsymbol{E}^2$ 模型	50.47	46.03
集成模型	**51.54**	**48.12**
测试集		
集成模型	60.99	55.93

为了评价富特征和数据重构的有效性，我们在开发集上做了相应的对比实验：(1)有无富特征模块的效果对比；(2)是否进行数据重构处理的效果对比；(3)富特征模块和数据重构两者一起对模型效果的影响。整个实验结果如表 3 所示。从结果中可知，富特征模块和数据重构对整个模型贡献很大，分别在 ROUGE-L 指标上有 5%和 7%的提升，其中数据重构比富特征对结果提升更大；同时加入两个模块，模型得到了接近 14%的提升，这说明了本文提出方法的有效性。

表 3　模块对比

模型	ROUGE-L	BLUE-4
总体结果	**50.90**	**47.56**
无富特征	45.52(−5.38)	43.96(−3.60)
无数据重构	43.78(−7.12)	40.62(−6.94)
无富特征和数据重构	37.19(−13.71)	36.62(−10.94)

注：实验结果基于开发集，所有模型使用 $\boldsymbol{E}^1$；无富特征表示只是用词向量。

3.3　错例分析

从模型在开发集上的表现，可以总结出以下几种错误类型。

(1) 真实答案为非连续片段；例如 ID＝186572，问题是["2017"，"有"，"什么"，"好看"，"的"，"小说"]，答案由多个文本片段组成，每个答案之间包含大量的无关介绍文本。目前本文框架还无法解决此类问题，未来可以将其转化成序列标注问题或者生成问题。

(2) 真实答案不在文档中，例如，ID＝181576。这种错误只能寄希望于补充更多的事实文本信息，进行答案补充处理。

(3) 文档标题信息融合不够导致的错误，例如，ID＝181579；问题是["csgo"，"读取"，"游戏"，"后"，"提示"，"已"，"停止"，"工作"]，所提出模型给出的答案是定位在题目为"csgo 一点开始就停止工作"的文档下。下一步本文将考虑如何将标题信息进行融入。

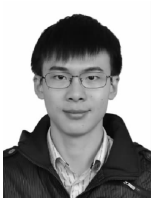
4　总结

本文面向机器阅读理解提出了一种简单且有效的神经交互网络。首先，为了充分使用标注的答案信息，本文对发布的数据进行了重构；在建模词表征的时候，使用了多种词语义特征；接着使用问题到答案的注意力操作以及双向 GRU 完成文档和问题信息的有效融合。在最终的测试集上，本文模型达到了现有的先进的效果，名列所有参赛队的第 2 名。

参考文献

[1] Yichun Yin, Yangqiu Song, Ming Zhang. NNEmbs at semeval-2017 task 4: Neural twitter sentiment classification: a simple ensemble method with different embeddings[C]//Proceedings of the 11th International

Workshop on Semantic Evaluation (SemEval-2017), 2017: 621-625.
[2] Tomas Mikolov, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of NIPS 2013, 2013: 3111-3119.
[3] Bojanowski Piotr, et al. Enriching word vectors with subword information[J]. arXiv:1607.04606, 2016.
[4] Wei He, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications [J]. arXiv preprint arXiv:1711.05073, 2018.



尹伊淳(1990—),博士,主要研究领域为自然语言处理、深度学习。
E-mail: yichunyin@pku.edu.cn



张铭(1966—),通信作者,博士,教授,主要研究领域为机器学习、文本挖掘、健康大数据等。
E-mail: mzhang_cs@pku.edu.cn

第七届未来数据论坛暨第五届大搜索论坛在哈尔滨举办

2018 年 8 月 2 日—4 日,第七届未来数据论坛暨第五届大搜索论坛在黑龙江省哈尔滨市举行。本次大会由中国工程院、FFD and BigSearch Steering Committee 主办,哈尔滨工业大学、国防科技大学与中国中文信息学会网络空间大搜索专委会联合承办。来自中国、美国、澳大利亚、香港等国家和地区共 30 多位大数据及相关领域杰出科学家应邀出席大会,交流研讨大数据研究最新成果。中国中文信息学会理事长方滨兴院士、副理事长黄河燕教授、副理事长宗成庆研究员和副理事长兼秘书长孙乐研究员出席了会议。

未来数据论坛(FFD)是 2011 年由美国伊利诺伊大学芝加哥分校的工程学院计算机科学系 Philip S. Yu 教授、香港城市大学计算机科学系的李青教授、澳大利亚维多利亚大学工程与科学学院的张彦春教授,以及中国国防科技大学计算机学院的贾焰教授联合倡议并发起的一个由海内外大数据及相关领域有影响的科学家出席的小规模国际研讨会,旨在交流大数据研究的最新进展,特别是在实际应用领域的新突破。自 2016 年起,未来数据论坛(FFD)与"大搜索"(BigSearch)论坛合并举行,为各个数据相关领域的国内外知名专家提供了交流平台。

此次论坛上,与会专家交流了大数据研究的最新进展, Philip S. Yu 教授、李青教授、张彦春教授、周晓方教授、贾小华教授、陈熙霖教授、林学民教授、许进教授、李飞飞教授等 20 多位嘉宾分别作题为《Broad Learning via Fusion of Heterogeneous Information for Mobile User Identification》《Event Cube》《医学数据分析与睡眠疾病辅助检测/诊断》《Developing a General-Purpose Spatial Trajectory Data Management System》《Privacy-Preserving Service Matching Platforms on Public Clouds》《Towards Scene Understanding》《Towards Big Graph Processing: Applications, Challenges, and Advances》《探针计算机》《Large-scale Spatio-Temporal Data Learning and Analytics》等报告,通过交流大数据与搜索领域的新见解、新成果和新计划,致力于分享大数据研究的最新进展,展望大数据技术及新兴应用的发展方向和产学研用相结合的发展途径,全面深入解读了国内外大数据产业当前态势与未来发展。

此外,大数据和人工智能技术的发展带来了新的机会,网络空间大搜索需求应运而生。网络空间大搜索是指面向泛在网络空间中的人、物体、信息和服务,在正确理解用户意图基础上的,基于从网络空间大数据获取的知识,给出满足用户需求的智慧解答。复旦大学王晓阳教授组织了大搜索 Panel,围绕大搜索的"5S"特性进行了讨论:(1)泛网获取(Sourcing from the Cyber),从多通道获取信息,提炼知识;(2)知识融合(Synthesizing the Knowledge),构建由巨规模实体及关系的知识仓库,并进行综合;(3)意图感知(Sensing the Context),即理解用户的搜索需求;(4)智慧解答(Intelligent Solution),产生智慧解决方案;(5)安全可信(Security(Privacy and Trust)),能够保护用户隐私。