文章编号：2096-1472(2017)-10-14-02

机器阅读理解软件中答案相关句的抽取算法研究

刘海静
(太原工业学院计算机工程系，山西 太原 030008)

摘　要：编写机器阅读理解软件中，一个基本步骤就是对于给定问题先在文档中找到和答案相关的语句。目前该领域大部分算法都使用递归神经网络，但由于很难序列并行化，这类算法在长文档上运行很慢。受人类在首次浏览文章时识别与问题相关的段落或语句，并仔细阅读这些内容得到答案的启发，本文采用一个粗糙但快速的模型用于答案相关句的选择。实验在WIKIREADING LONG 数据集上取得了较好的结果。

关键词：机器阅读理解软件；神经网络；答案相关句
中图分类号：TP391.1　　**文献标识码**：A

A Study of the Question-Aware Passage Extraction in Machine Reading Comprehension Software

LIU Haijing
(*Department of Computer Engineering*,*Taiyuan Institute of Technology*,*Taiyuan* 030008,*China*)

Abstract: In the programming of machine reading comprehension software,one basic step to extract the relevant sentences from the whole passage according to the given question.Recently,Recurrent Neural Network (RNN) is applied in most approaches,but it runs excessively slow when dealing with long documents because it is difficult to parallelize sequences.Inspired by the process that people first skim the document,identify relevant parts,and carefully read these parts to produce an answer,the paper proposes a rough but efficient model to extract the relevant sentences.Experiment results demonstrate the good performance on WIKIREADING LONG datasets.

Keywords:machine reading comprehension software;RNN;relevant sentences

1 引言(Introduction)

阅读理解是自然语言理解中的一个关键问题。近年来，随着大规模阅读理解语料的可用性，从无结构化文档中进行问答研究已经成为热门领域[1−3]。

与此同时，深度学习也已成为自然语言处理领域的热点，几乎所有的研究热点都已成为深度学习的天下。阅读理解中对于问题相关句的抽取任务也不例外，有许多专家学者利用深度学习来进行问题相关句抽取任务的研究[4−6]。

目前，在无结构化文档上进行问答研究的最好方法是基于递归神经网络，该模型将文档和问题进行编码来获得答案。尽管这样的模型能够获得所有相关信息，但是由于模型需要在可能成千上万的词语上序列化运行，并且计算不能平行化而变得很慢。事实上，这样的模型通常是把文档截短，并只考虑有限的一部分词语。受人类第一次浏览文档时识别问题相关部分，认真阅读这部分内容从而产生答案的研究[7−9]，本文提出了一种问题相关句选择模型。该模型能快速从文档中选择一些与回答问题相关的句子。

本文认为，答案并不需要显式地逐字出现在输入中(比如，即使没有被显式说明，一部电影的类型也可以轻易确定)。此外，答案经常会在文档虚假的上下文中出现很多次(比如，年份2012可能在文档中出现很多次，但真正与问题相关的可能只有一次)。因此，本文将句子选择看成是由仅使用增强学习的答案生成模型产生的答案并行训练的潜在变量。据我们所知，在分类研究领域有将句子选择看成是一个潜在变量，但是问答领域尚属首次。

本文展示了一个对于长文档阅读理解的模块化框架和学习步骤。它能够捕获文档结构的有限形式如句子边界，并能够处理长文档。WIKIREADING LONG数据集上的实验结果显示本文模型是有效的。

2 问题设置(Problem setting)

给定一个问题—文档—答案三元组的训练集$\{x^{(i)}, d^{(i)}, y^{(i)}\}_{i=1}^{N}$，我们最终目的就是一个模型能够对问题—文档对$(x, d)$生成一个答案$\mathrm{y}$。文档$\mathrm{d}$是一系列句子$s_1, s_2, \ldots\ldots, s_{|d|}$的集合。结合语言学研究和人类的直观感知，我们认为答案是可以从一个与问题相关的潜在的句子子集中生成的，这就是本文的研究内容。下面以儿童故事《小猫钓

www.rjgcz.com

鱼》为例说明研究内容，其中答案出现在s_5子句中。

文档d：

s_1：小猫坐在河边跟妈妈学习钓鱼。s_2：蝴蝶飞过，小猫放下鱼竿就去抓蝴蝶；s_3：蜻蜓飞来，小猫转过身就去追蜻蜓。s_4：一天的时间很快就过去了。s_5：傍晚，小猫低着头，提着空空的篮子跟着妈妈回家了。

问题x：小猫跟着妈妈去钓鱼，他钓到鱼了吗?

答案y：没有。篮子空空的。

3　数据(Data)

本文方法在WIKIREADING LONG数据集上进行了测试。

WIKIREADING LONG是一个问答数据集，由Wikipedia and Wikidata自动生成：给定一个关于某实体的维基页面和维基数据资源，比如职业或性别，目标就是基于该文档推断其价值。和其他最新发布的大规模数据集不同，WIKIREADING LONG并不标注答案界限，这使得句子选择更加困难。

4　句子选择模型(Sentence selection model)

句子选择模型定义了一个给定输入问题x和文档d、基于句子层面的$p(s|x,d)$分布。按照近年来在句子选择方面的研究，本文构建了一个前馈网络来定义句子上的分布。我们考虑了三种简单的句子表示：词袋模型、块结构词袋模型、卷积神经网络模型。这些模型能够有效处理长文档，但是不能完全捕获文本的序列性，即句子中词语的位置信息。

4.1　词袋模型

给定一个句子s，我们定义了s中词语的词袋表示。为了定义文档句子的分布，我们使用了一个标准注意模型，将问题的词袋表示级联到文档中每个句子的词袋表示上，然后输入给一个单层前向网络：

$$h_l = [BOW(x)\ ; BOW(s_l)]$$

$$v_l = v^T ReLU(Wh_l)\ ; p(s = s_l|x,d) = softmax(v_l)$$

其中，[;]表示问题和文档中句子的词袋表示级联，矩阵W、向量v和word embeddings是学习参数。

4.2　块结构词袋模型

为了获得更精细的粒度，我们将句子分割成一些长度固定的小块，并且对每个小块单独进行了打分。这种做法适用于用句子的子块回答问题，在不同小块上学习注意力词袋表示，并和词袋模型一样进行了评分。最终获得了一个块上分布，并通过边缘化相同句子组块来计算句子概率。假定$p(c = c_{l,j}|x,d)$是来自所有句子的块上分布，则有：$p(s = s_l|x,d) = \sum_{j=1}^{J} p(c = c_{l,j}|x,d)$，参数与词袋模型参数一致。

4.3　卷积神经网络模型CNN

尽管我们的句子选择模型有快速性要求，我们还是使用了一种卷积神经网络来更好地表达临近词的语义。相对而言，卷积神经网络模型CNN仍然是有效的，因为所有的过滤器可以平行计算。

借鉴了目前该领域的最新研究，我们将问题x和句子s中word embedding进行了级联，并使用特征抽取器(Filter)和宽度w运行一个卷积层。这产生了每个具有w长度的F特征，我们利用了MaxPooling Over Time来获得一个最终的表示。MaxPooling Over Time是NLP中CNN模型中最常见的一种下采样操作，意思是对于某个Filter抽取到若干特征值，只取其中得分最大的那个值作为Pooling层保留值，其他特征值全部抛弃。值最大代表只保留这些特征中最强的，而抛弃其他弱的此类特征。

在类似于阅读理解的NLP任务中使用Max Pooling的好处在于可以把变长的输入x整理成固定长度的输入。因为CNN最后往往会接全联接层，而其神经元个数是需要事先定好的。如果输入是不定长的，那么很难设计网络结构。CNN模型的输入X的长度是不确定的，而通过Pooling操作，每个Filter固定取一个值，那么有多少个Filter，Pooling层就有多少个神经元，这样就可以把全联接层神经元个数固定住。如图1所示。

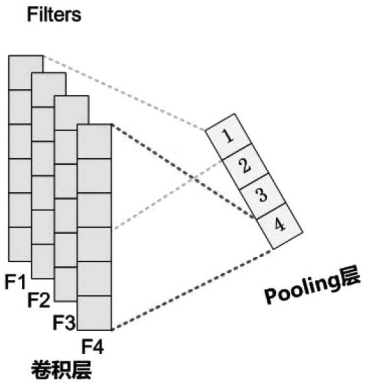

图1 Pooling层神经元个数等于Filters个数

Fig.1 Numbers of fiters equals numbers of neurons in pooling

最后，和词带模型一样，将h通过一个单层前向网络计算句子概率。

5　实验(Experiments)

文中使用Word2Vec对训练数据进行处理，每个单词就可以得到对应的word embedding，这是一种低维度向量形式的单词表示，能够表征单词的部分语义及语法含义。

在模型参数的学习方面，本文采用了一种管道模型，使用远距离监督来训练句子选择模型。

实验数据中70%为训练集，10%为开发集，20%为测试集。每篇文档中前35条句子作为启发式模型的输入，其中每条句子最大长度为35。将每篇文档的前5个词语加到句子序列的结尾，并将句子索引编号作为句子表示的一个重要向量。

www.rigcz.com

使用三种方法进行问题相关句子选择的结果，详见表1。

表1 三种方法用于问题相关句子选择

Tab.1 Three approaches for selection of query related sentences

数据集	模型	准确率
WIKIREADING LONG	词袋模型	67.2%
	块结构词袋模型	70.4%
	卷积神经网络模型	71.9%

6 结论(Conclusion)

机器阅读理解是自然语言处理任务中的一个核心问题。数据语料集的复杂度、是否引入了世界知识、深度学习模型的改进等都对该问题有重要影响。目前，机器自然语言阅读理解很难对于给定的问题一次性给出答案。本文认为，首先从文档中选择与答案相关的句子是很重要的一个步骤。借鉴了目前最新的研究，本文分别从词袋模型、块结构词袋模型、卷积神经网络模型三种方法来进行考察。结果表明，神经网络方法对于该问题是最有效的。

参考文献(References)

[1] Collobert R,Weston J,Karlen M,et al.Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research(JMLR),2011,12:2493-2537.

[2] Masson M E.Conceptual Processing of Text during Skimming and Rapid Sequential Reading[J].Memory & Cognition,1983,11(3):262-274.

[3] Williams R J. Simple Statistical Gradient following Algorithms for Connectionist Reinforcement Learning[J].Machine Learning,1992,8(3-4):229-256.

[4] Bordes A,Chopra S,Weston J.Question answering with subgraph embeddings[C].EMNLP,2014:615-620.

[5] Tan M,Santos C N,Xiang B,et al.Improved Representation Learning for Question Answer Matching[C].Meeting of the Association for Computational Linguistics,2016:464-473.

[6] Hermann K M,Kocisky T,Grefenstette E,et al.Teaching Machines to Read and Comprehend[C].In Proc. of NIPS,2015, 19:1684-1692.

[7] 刘江鸣,徐金安,张玉洁.基于隐主题马尔科夫模型的多特征自动文摘[J].北京大学学报:自然科学版,2014,50(1):187-193.

[8] 谭红叶,赵红红,李茹.面向阅读理解复杂问题的句子融合[J].中文信息学报,2017,31(1):8-16.

[9] 张志昌,张宇,刘挺.开放域问答技术研究进展[J].电子学报.2009,37(5):1-6.

作者简介：

刘海静(1985-)，女，硕士,讲师．研究领域：自然语言处理．

(上接第23页)

物信息的处理变得更加安全、迅速、精准。另外，无线传感器网络技术通过利用传感器节点来感知大型耗能设备的动态参数信息，使实验室管理变得环保节能。

本系统在硬件上选用了ER302读写器和若干电子标签、TSL2561无线传感器等，进而实现数据信息的采集。在软件环境上，开发环境为Microsoft Visual Studio 2010，数据库为SQL Server 2008，开发平台为ASP.NET，整体架构为B/S结构。

在这次研究中，也存在不足或不完善的地方。比如，没有考虑到传统的条形码技术和RFID技术的融合。如何优化目前的RFID系统将是今后的主要研究方向。目前本管理系统只是完成了基本功能的实现，还有待于进一步完善。

参考文献(References)

[1] Anne Bremer.Diffusion of the "Internet of Things" on the world of skilled work and resulting consequences for the man-machine interaction[J].Empirical Research in Vocational Education and Training,2015,7(1):1-13.

[2] Xingyanli,Ninglu.Exploration and practice of self-service in university library—case study of Beijing University of Agriculture Library[J].Advanced Materials Research,2014,988:724-728.

[3] Ninglu,Xingyanli.The application of RFID Technology in the University Library[J].Applied Mechanics and Materials,2013,263-266:3279-3283.

[4] Ninglu,Liyunle,Xingyanli.The construction and innovation of information service system for academic libraries of agriculture[J]. Applied Mechanics and Materials,2013,263-266:3284-3287.

[5] 欧文.物联网技术及其在农业生产中的应用研究[D].昆明:昆明理工大学,2015:131.

[6] Ninglu,Xingyanli.Cloud computing applications in library information services.Applied Mechanics and Materials, 2013,373-376:1719-1723.

[7] 张兴,韩冬,曹光辉,等.基于PRESENT算法的RFID安全认证协议[J].通信学报,2015,36(S1):65-74.

[8] 陈志辉,王颖纯,刘燕权.基于物联网环境的图书馆RFID技术应用现状的研究[J].情报杂志,2015,34(05):196-201;189.

作者简介：

张倩怡(1994-)，女，硕士生.研究领域：农业信息化．

李小顺(1984-)，男，硕士，实验师.研究领域：实验室建设与管理.本文通讯作者.

胡　萍(1980-)，女，硕士，实验师.研究领域：实验室管理．

宁　璐(1972-)，男，硕士，研究馆员.研究领域：信息技术应用，网络与信息安全．

www.rjgcz.com