DOI:10.13878/j.cnki.jnuist.2019.06.008

曾明睿[1]　袁梦奇[1]　邵曦[1]　鲍秉坤[1]　徐常胜[1,2]

文本特征提取的研究进展

摘要

文本理解是人工智能的一个重要分支,其技术推动了人与计算机之间在自然语言上的有效交互.为了让计算机准确地理解和感知文本数据,文本特征提取是最为基础和关键的步骤之一.基于此,本文介绍文本特征提取研究的发展历史,以及近年来主流特征提取的方法,并对未来的研究方向进行展望.首先,介绍语义最底层的词级表示;接着,总结在词级表示基础上衍生出的句级表示上的研究进展;随后,介绍比词级表示和句级表示更高层的篇分析;最后,通过文本特征提取的一个典型应用——问答系统的介绍,阐述文本特征提取的最新方法和技术在问答系统上的应用,并对未来的研究方向做了展望.

关键词

自然语言处理;文本特征提取;问答系统

中图分类号 TP391.1

文献标志码 A

收稿日期 2019-10-15

资助项目 国家自然科学基金(61572503,61872424,6193000388,61872199);南京邮电大学高层次人才启动基金(NY218001);模式识别国家重点实验室开放课题(201900015)

作者简介

曾明睿,男,硕士生,主要研究方向为多媒体计算.894606932@qq.com

鲍秉坤(通信作者),女,博士,教授,博士生导师,主要研究方向为多媒体计算和计算机视觉.bingkunbao@njupt.edu.cn

1 南京邮电大学 通信与信息工程学院,南京,210044

2 中国科学院自动化研究所 模式识别国家重点实验室,北京,100190

0 引言

互联网技术的高速发展,以及硬件产品的不断更新换代,使得网络上的数据呈现出“爆炸式”的增长态势.特别是作为信息主要载体的文本数据,一方面数量迅速增长,另一方面其表现形式和结构也变得复杂多样,为文本理解带来了巨大的挑战.文本理解的核心是将文本数据通过数学运算转换为计算机可以感知和分析的信号,并根据任务的不同,对其进行自动处理以反馈结果.在文本理解中,最基础和最关键的步骤之一就是文本特征提取.文本特征提取是为文本数据集寻找一个具有判别力的特征空间,并将所有的文本数据映射到这一空间上,以抽取有代表性的、鲁棒的特征表示向量.

互联网上涌现的海量文本数据,既带来了丰富的语料资源,同时也使文本感知、分析和处理面临了巨大的挑战.首先,每个用户都可以产生和传播数据,而其中文本的占比又最大,这导致了文本语料规模的迅速增长,因此“大数据”是面临的第一个挑战;其次,在大数据的背后隐藏了大量重复且无意义的数据,这些数据良莠不齐,价值密度低,因此“大噪声”是面临的第二个挑战;最后,数据存在于各种各样的平台中,其类型包括了结构化数据、半结构化数据和非结构化数据等,因此“结构复杂”是面临的第三个挑战.

近年来,许多学者针对新环境下文本数据的这三个挑战,在文本特征提取上提出了大量有效的方法和技术.本文将对这些研究成果进行归纳和总结,为该方向的研究人员快速了解文本特征提取提供参考.依据语义单元的大小,本文首先介绍词上的特征提取方法和技术,包括利用上下文信息和外部知识引入;随后介绍比词级更高一层的句级特征表示,主要基于词级表示的方法,通过引入词和词之间的关联,对句子进行更高层语义的理解;再次,对语篇表示的研究成果进行总结,主要关注语篇关系挖掘的方法和技术.最后,介绍文本特征提取在问答系统上的典型应用,将结合双向 Transformer 的编码表示、注意力模型和卷积神经网络的方法展开阐述.新时代背景下所面临的大数据、大噪声和结构复杂三个挑战,也是词级表示、句级表示、语篇表示和问答系统需要解决的难点,因此本文在文献总结的过程中将侧重这三个方面详细阐述相关的应对方法和解决方案.

本文第 1 章到第 3 章将依次详细阐述在对于词级表示、句级表示和语篇关系三层语义做特征提取时所采用的技术,并对每层语义级

再次细分做介绍.第 4 章是对文本特征提取方法进行结合和实际在问答系统的应用.最后,展望了文本特征提取的未来研究方向并对全文进行总结.

1 词级表示

词作为文本中最基础的单位,是构成句子和语篇的最小元素.对词的特征提取通常称为词级表示,但在文本中,不管是英文单词还是中文词汇的数量都是非常庞大的,仅仅对这些词进行顺序编码,不仅人力花费高昂,还难以揭示词与词之间的语义关系,因此对词级进行语义距离可度量的向量化表示是非常必要的.具体来说,在给定某一语义度量准则下,将每个词或词组投影为高维的向量,这些向量形成的空间称为词级的向量空间,以此将非结构化的文本转化为了可处理的结构化形式.然而这种工作是属于预训练的范畴的,当我们把词级表示应用到实际问题的时候,无须从零开始训练一个新的模型,这为后面的训练大大节省了时间.目前关于词的预训练方法,可以分为两条思路:利用上下文相关信息和外部知识关系的结合.

1.1 利用上下文相关信息

在自然语言中,很多单词有着多种含义,而其真实含义是根据所在的上下文语境来决定的.因此在设计词的特征提取模型时,需要引入上下文相关信息,以消除一词多义的影响.根据模型种类的不同,基于上下文信息的词级表示方法可以分为基于 LSTM 模型和基于 Transformer 模型两类.

基于 LSTM 模型这类方法,是针对于传统方法(如 word2vec 等)忽略词的上下文关系,无法建模词的一词多义的缺陷所提出的.具体实现是通过将整句的单词,输入进 LSTM 神经网络中,通过 LSTM 建模目标词和句子里其他单词的上下文的语义关联,来获得融合其他单词信息的词级表征.根据融合单词与目标词的位置不同,这类词级表示的方法可以分为两类:前向融合[1](图 1)和双向融合[2](图 2).前向融合只考虑目标词之前的词对其产生的语义影响,如图 1 所示,对“into”进行词级表示,将“into”之前的单词“problems”、“turning”等依次输入至 LSTM 模型中,根据单词与目标词的远近,进行有选择的记忆存储和遗忘,并将记忆信息融合至“into”的词级表示中.很显然,不仅“into”之前的单词对其有语义影响,其之后的单词“banking”、”crises”、“as”等也会有影响,因此学者又在前向融合的基础上,考虑目标词之后的词,提出双向融合的方法(图 2).具体而言,建模由两个 LSTM 构成的 Bi-LSTM 模型,分别从前往后和从后往前对单词进行输入,以融合目标词前后的所有单词的语义.基于 Bi-LSTM 模型,Melamud 等[2]改进了基于 word2vec 的 CBOW 图,提出 context2vec.其中,基于 word2vec 的 CBOW 图计算窗口内所有词嵌入的平均值(图 3),而 context2vec 是基于 Bi-LSTM 融合目标词的上下文(图 4).为了建模更为复杂的上下文语义关系,Peters 等[3]提出了 ELMo(Embeddings from Language Models)模型,这是一种深度语境化词表示方法,由两层 Bi-LSTM 组成的语言模型内部状态函数生成的词向量,通过 Bi-LSTM 连接的语言模型将每个输入字上方堆叠的向量的线性组合以组合表示多样的文字特征,表示更加丰富的语义.

相比于基于 LSTM 模型的方法,Transformer 模型不仅不需要通过循环来并行处理句中的单词,还能结合上下文的信息,因此在处理长语句时,效率较高.Radford 等[4]最先基于 Transformer 提出了 Open AI GPT,该模型克服了 LSTM 的短期预测能力,通过捕捉长距离的语言结构,来学习一个通用表示. 2018

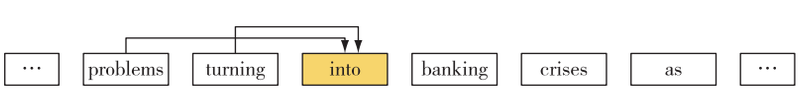

图 1 前向融合[1]

Fig. 1 Forward convergence[1]

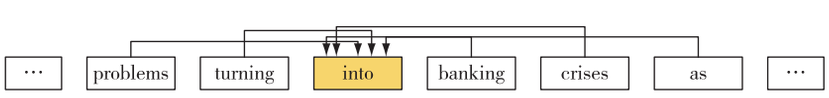

图 2 双向融合[2]

Fig. 2 Bi-direction convergence[2]

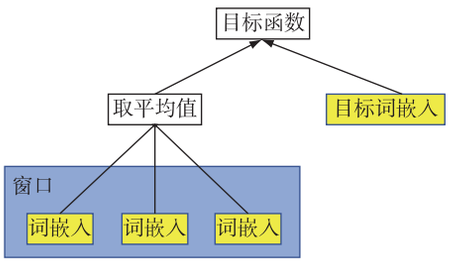

图 3　CBOW 模型[2]

Fig. 3　CBOW model[2]

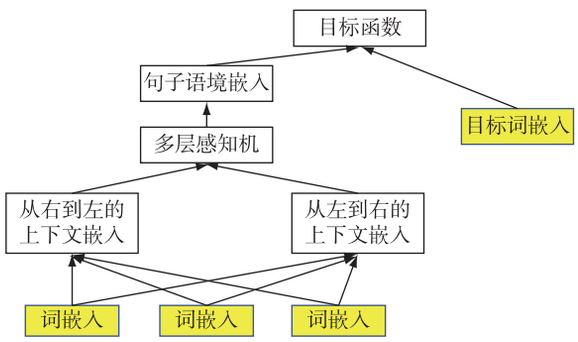

图 4　Context2vec 模型

Fig. 4　Context2vec model

年,Devlin 等[5]提出了基于深度双向 Transformer 的 BERT (Bidirectional Encoder Representation from Transformer)模型,与 Open AI GPT 单方向的预训练模型不同的是,BERT 提出了一种遮蔽语言模型(Mask Language Model)来训练出上下文的特征(图 5),它通过遮蔽一个单词,训练一个深度双向 Transformer 模型,从单词的左右两个方向来预测遮蔽单词.2019 年,Dai 等[6]通过引入相对位置编码和片段循环机制对 Transformer 模型进行改进,提出 Transformer-XL 模型,循环机制在每处理完一个片段之后都会将输出保留在隐藏层中以便后面循环,建立长期的依赖关系.而相对位置编码则是通过对隐藏状态的相对位置进行编码,克服了不同片段编码可能导致编码一样的问题.两种方法的融合解决了由于固定上下文的长度所带来的无法获取超出定义长度的依赖关系的问题.

1.2　外部知识的引入

传统的词级表示方法在情感分类、文本分类等任务上取得了令人满意的结果,但当处理稀疏词汇时,由于词汇出现的频率较低,无法对其抽取得到准确的语义,甚至容易受到噪声的干扰.因此,学者们提出通过加入维基百科等其他语料库,引入外部的知识,以获得更为准确的词级表示.

如何将外部语料库有效地引入到目标语料库中,生成融合外部知识的词嵌入,是目前这部分工作面临的挑战.2017 年,Cao 等[7]建模文本和知识库之间的关联,以解决多义词引起的歧义的问题.Sarma 等[8]分别在目标语料库上训练一个通用词嵌入和在外部语料库上训练一个外来词嵌入,然后对两组嵌入使用线性 CCA[9]或非线性 CCA[10],沿着最大相关的方向投射,再取平均值,最终得到引入外部知识的词级特征表示.Xu 等[11]将通用词嵌入和外来词嵌入的双重嵌入机制与 CNN 网络结合,让 CNN 网络决定两种嵌入中可用信息的比重,从而使文本特征提取更加高效、简单.相较于 BERT 的 Mask Language Model 无法对显式语义单元进行建模,百度的 Paddle 发布了知识增强的预训练模型 ERNIE[12](Enhanced Language Representation with Informative Entities),该模型通过将知识图谱在编码输入至预训练模型,从而有效地挖掘了图谱中实体间关系,最终增强了模型语义表示能力.例如在图 6 中,“哈尔滨”作为一个整体被抹去时,则需要通过更长的依赖性来预测学习,而 ERNIE 可以通过先验知识“黑龙江的省会”预测表示出遮掩词“哈尔滨”.

2　句级表示

仅依靠词级表示,无法获得对文本的准确理解,

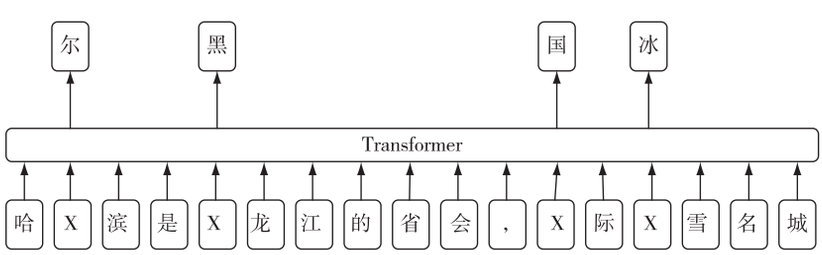

图 5　遮蔽语言模型[5]

Fig. 5　Mask language model[5]

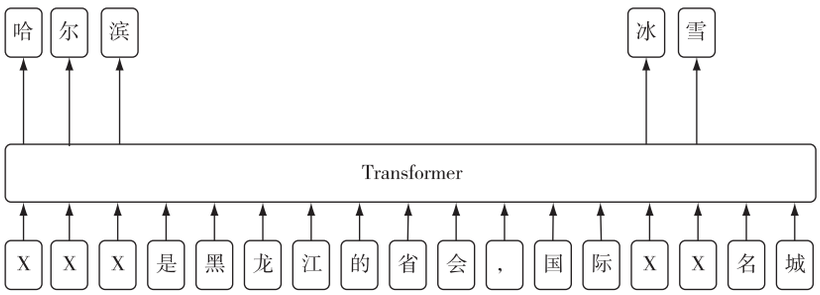

图 6　ERNIE 的学习

Fig. 6　Learnt by ERNIE

需要考虑词和词之间的关联对语义的影响,因此句子级表示的研究就应运而生了.句级表示方法最常用的是句子嵌入,具体来说是用向量来表示自然语言中的语句,使其携带句子中的语义信息.句子嵌入方法可分为基于词向量的线性组合和基于深度学习两类方法.

2.1　基于词向量的线性组合

把句子中所有词的词嵌入取平均值是一种非常成功和高效的获得句子嵌入的方法[13].具体来说,是将句子中每个词嵌入相加除以句中词数得到的向量值作为句嵌入.这一方法的缺陷在于忽略了句中词的权重和顺序. Kenter 等[14] 基于 word2vec 中的 CBOW 提出了 Siamese CBOW(图 7),与 CBOW 有着相同的原理,只不过该模型是将句中的词向量先做平均值处理表征句向量,然后通过周围的句子对目标句子进行预测来学习词嵌入以便达到优化的目的,最后对优化之后的词嵌入做平均值处理形成句向量.Arora 等[15]仅计算句子中词向量的加权平均,然后删除第一个向量上的平均投影,权重的计算来自于作者提出的 SIF,即一个词的权重:$w = \frac{a}{a + p(w)}$,其中,a 为参数,$p(w)$为预测词的词频.这样的加权方案具有十分不错的鲁棒性:使用从不同语料库得出的单词频率不会损害性能并且 a 的取值很广,可以让结果达到最佳.

2.2　基于深度学习的句级表示

近年来,随着深度学习在文本领域的广泛应用,越来越多的学者在句级表示上尝试引入深度学习模型,以建模词与词之间的复杂关系.目前基于深度学习的方法主要基于循环神经网络、卷积神经网络和 encoder-decoder.

在基于循环神经网络方面,Zhang 等[16] 提出 sentence-state LSTM,每次循环都对所有单词语义特征的隐藏状态进行建模,而不再是一次一个单词输入.将整个句子看成一个状态,这个状态是由各个词的子状态和一个整体的句子层状态组成.在每次循环时,单词的隐藏状态都能捕捉到越来越大的 n-gram 信息,并与句子状态进行信息交换.最终,循环得到一句话的表示.

卷积神经网络方法在图像处理上已经取得了非常不错的效果,要求输入值是一个固定的图像分辨率.近年来,学者也在尝试将卷积神经网络应用在自然处理上,但是输入的文本或者句子长度不固定会

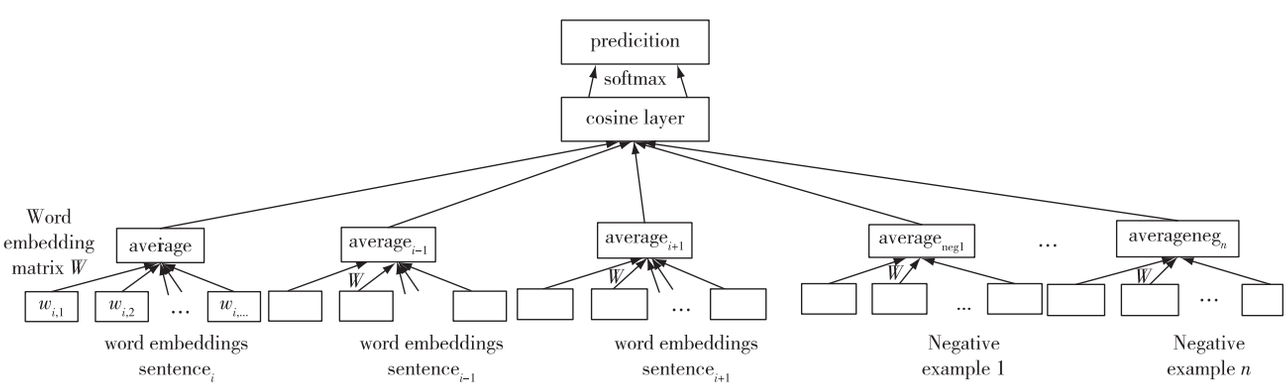

图 7　Siamese CBOW 网络结构[14]

Fig. 7　Siamese CBOW network structure[14]

造成多余的卷积结果丢失,从而对模型结果产生影响.Kim[17]先将词表示变成矩阵,然后通过一个仅有一层卷积层的简单 CNN,对其进行 Max-overtime pooling,最后经过全连接层得到句向量.Santos 等[18]让词嵌入和字符嵌入通过卷积神经网络联合表示形成句向量,其创新之处在于利用两层卷积层去提取词和句中的相关特征.第一层提取句子级的特征,第二层获取每个单词字符周围生成的局部特征用最大化的方式将其组合,最终生成一个固定大小的向量.

在 encoder-decoder 方面,句级表示主要是将词级表示中的 word2vec 模型推广到句子上.Kiros 等[19]提出了 Skip-Thought Vectors,通过大量连续的语料库训练出一个 encoder-decoder 模型,将多个词向量编码成句向量,并同时用一个句子来预测上下文另一个的句子.模型如图 8,模型中是用一个三元组 (s_{i-1},s_i,s_{i+1}) 表示连续的三句话,将来自连续语库 s_i 编码重建前一句 s_{i-1} 和后一句 s_{i+1}.图中未连接的箭头连接到编码器输出,颜色指示了共享参数的组件.受到 BOW 编码思想的启发,Hill 等[20]提出了对数线性语句模型——FastSent,将一个连续句子的三元组(s_{i-1},s_i,s_{i+1}),对于中间的句子 s_i 进行编码,编码方式是将 s_i 中的词向量求和即 $\sum_{w\in s_i} s_i$,这种方法没有考虑句中的词序,因此使得 FastSent 的训练速度大幅提升.根据实验用 Skip-Thought Vectors[19] 和 FastSent 两种模型训练得到参数的数据如表 1 所示,其中 * 表示在 GPU 上进行训练.

表 1　两种模型参数比较[20]

Table 1　Parameter comparison between two models[20]

	句向量维度/维	词向量维度/维	训练时间/h
Skip-Thought Vectors	4 800	620	336 *
FastSent	100	100	2

注:* 表示在 GPU 上进行训练.

3　语篇分析

事实上,句子之间也会存在着复杂的逻辑关系,因此需要引入语篇分析挖掘来进一步理解文本.语篇分析又称篇章分析,是通过对文本内部实体关系的挖掘和理解,对语篇整体进行分析,从而获得整个文档的高层语义.本章将分别介绍语篇分析中文本关系和隐式语篇表示嵌入两部分的研究.

文本关系抽取需要深入理解语篇内所有实体之间的关系,由此学习到的文本关系嵌入可以用来扩充现有的关系提取模型,并能显著地提高它们的性能.Xu 等[21]通过卷积神经网络从实体间最短依赖路径学习更稳健的关系表示文本关系.但是这一方法需要依赖大量的标注句子作为训练集生成嵌入模型.Su 等[22]提出 GloRE,通过定义句子依赖图中两个实体的最短路径去改进关系提取,同时将文本关系和知识库关系的全局共现统计来学习文本关系的嵌入.可是由于手工标注的训练集太少,这一方面仅适用于小规模的训练数据的关系提取.2019 年,Chen 等[23]将 GloRE 方法与可以从缺少标签的数据中提取关系的远程监督方法[24]相结合进一步应用于大规模、领域无关的数据,目的是学习通用文本关系嵌入.

作为语篇分析另一重要分支,隐式语篇分析是在没有显式连接词的情况下提取关系,这很难从输入句子对的表面特征派生出来,所以需要从文本语义理解的角度中去寻找关系.近几年不少学者已经提出了基于神经网络的方法或高级表示的模型:CNN[25]、注意神经张量网络[26]、记忆网络(memory network)[27]和 RNN[28]等.还有一些方法考虑到了上下文段落和段落间相关性[29].但是对于机器来说,如何更好地理解文本成为了隐式语篇关系识别研究前进的障碍.因此,Bai 等[30]通过字词和 ELMo[2]的增强嵌入和深度剩余双注意力编码器,让表示更加丰富和深入模型结构(图 9).

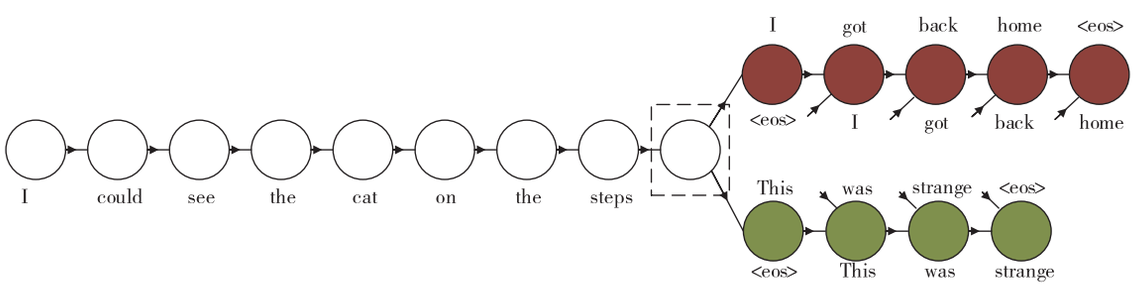

图 8　Skip-Thought Vectors 模型[19]

Fig. 8　Skip-Thought Vectors model[19]

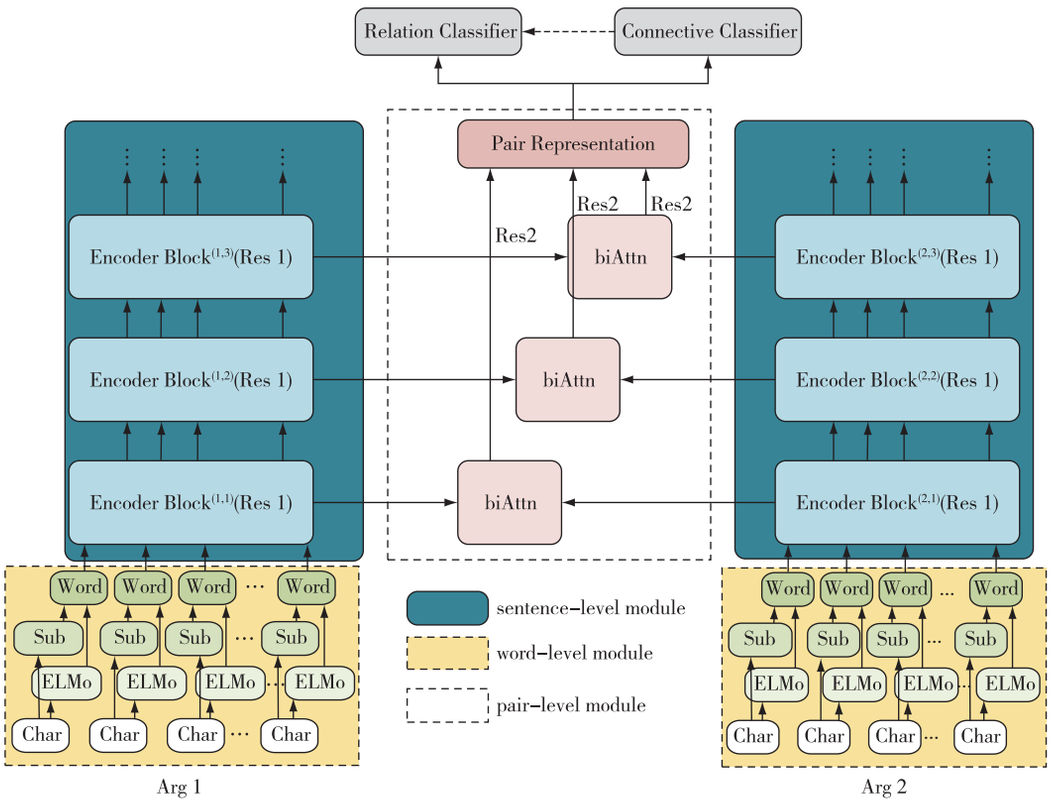

图 9　增强嵌入和深度剩余双注意力编码器[30]

Fig. 9　Enhanced embedding and deep remaining dual attention encoder[30]

4　文本特征提取结合实际的应用

问答系统是文本特征提取的一个典型应用,任务是能够准确地理解用户用自然语言提出的问题,并通过检索语料库、知识图谱或问答知识库返回简洁、准确的答案.相较于搜索引擎,问答系统能够根据已有语料库学习问答知识,进而更有效地满足用户的信息需求.将文本特征提取的技术应用在问答系统中可以很好地帮助计算机理解人类语言的重点,同时在提高训练速度、检索答案质量等方面都会有很好的表现.

在问答系统领域方面,有效的提取问句的意图识别和填槽可以为快速准确匹配出答案和使其更加人性化奠定基础.表 2 显示了一个用户查询的意图分类和填槽的实例.

表 2　用户查询的意图分类和填槽的实例

Table 2　An example of intention classification and slot filling of user query

	今天南京天气怎么样?
意图	查询天气
信息槽	具体查询哪里的天气? 哪一天的天气?

Chen 等[31]将之前 BERT[5] 扩展到一个联合意图分类和槽填充模型.基于第一个特殊 token 的隐藏状态 h_1 的意图被表示为 $y^i = \mathrm{softmax}(W^i h_i + b^i)$,而对于槽填充模型,将会提供除去第一个 token 的最终隐藏状态: $h_2, \cdots, h_T$ 进入 Softmax 层,对槽填充标签进行分类.当两个任务联合训练时,使目标函数 $p(y^i, y^s \mid x) = p(y^i \mid x) \prod_{n=1}^{N} p(y_n^s \mid x)$ 最大化.经过在 Snips 和 ATIS 数据集上测试的结果如表 3,可以看出基于 BERT 的意图分类和槽填充在准确率方面相较于其他方法都取得了最好的结果.

表 3　不同测试集上的实验结果[31]

Table 3　Experimental results on different test sets[31]　%

	Snips		ATIS	
	意图分类	槽填充	意图分类	槽填充
RNN-LSTM	96.9	87.3	92.6	94.3
Atten-Bi-LSTM	96.7	87.8	91.1	94.2
Slot-Gated	97.0	88.8	94.1	95.2
Joint Bert	98.6	97.0	97.5	96.1

此外,对于问题的理解对于问答系统来说也是

十分重要的.Dong 等[32]介绍了多列卷积神经网络,模型不依赖于手工特征和规则,通过共享相同的词向量来表示问题单词,使用不同列的网络来提取答案类型、关系和上下文信息.同时,也会在知识库中共同学习实体和关系的低维嵌入.使用问题-答案组合对模型进行训练,以对候选答案进行排序.如图 10 不同网络列获取得问题表示.

2017 年,Seo 等[33]提出 BIDAF(Bidirectional Attention Flow for Machine Comprehension)双向注意力矩阵来计算上下文的每个字与查询问题之间的相关性,从而建立查询感知的上下文表示.然而这种模型却不能像人类一样对候选答案进行反思,因此 Gong 等[34]将 BIDAF 扩展成 Ruminating Reader 使其能够进行第二遍阅读和推理,通过门控机制让第一遍和第二遍阅读的内容进行融合(模型框架如图 11),在选择答案的时候能够有效地使用上下文并更好地权衡答案.

5 展望

本文根据语义层面的由低到高依次总结了词、句和篇章三个层次上文本特征提取方法的研究进展.近年来,学者们注意到图作为一种特殊的数据结构,能够面对一组对象和对象之间的联系进行建模.由于这一强大的优点,把基于图神经网络的方法用于机器学习的方向越来越受人追捧.同时,现在数据平台的多样性使得数据结构变得极为复杂,给文本特征提取带来了不小的挑战,而图神经网络作为一种可以在图结构上运行的神经网络,能够保存图形嵌入中的全局结构信息,因此在处理具有丰富关系结构的任务时可以得到很好的效果.所以,利用图神经网络来应对结构复杂的文本信息也成为了一个新的研究方向.在问答系统方面,生成的回答也更加人性化,因此,在未来的文本特征提取中,应该建立新的文本特征表示模型,并结合领域知识快速定位用户的兴趣反馈,以达到更加流畅的使用感受.

列1(回答路径)	列2(回答类型)	列3(上下文信息)
what to do in hoollywood can this weekend	where be george washington originally from	where do charle draw go to college
what to do in midland tx this weekend	what to do in midland tx this weekend	where do kevin love go to college
what to do in cancun with family	where be george bush from	where do pauley perrette go to college
what to do at fairfield can	where be the thame river source	where do kevin jame go to college
what to see in downtown asheville nc	where be the main headquarters of google	where do charle draw go to high school
what to see in toronto top 10	in what town do ned kelly and he family grow up	where do draw bree go to college wikianswer
where do draw bree go to college wikianswer	who be the leader of north korea today	who be judy garland father
who found the roanoke settlement	who be the leader of syrium now	who be clint eastwood date
who own skywest	who be the leader of cuba 2012	who be emma stone father
who start mary kay	who be the leader of france 2012	who be robin robert father
who be the owner of kfc	who be the current leader of cuba today	who miley cyrus engage to
who own wikimedium foundation	who be the minority leader of the house of representative now	who be chri cooley marry to
what type of money do japanese use	what be the two official language of paraguay	what be the timezone in vancouver
what kind of money do japanese use	what be the local language of israel	what be my timezone in californium
what type of money do jamaica use	what be the four official language of nigerium	what be los angeles california time zone
what type of currency do brazil use	what be the official language of jamaica	what be my timezone in oklahoma
what type of money do you use in cuba	what be the dominant language of jamaica	what be my timezone in louisiana
what money do japanese use	what be the official language of brazil now	what be the time zone in france

图 10 使用不同列网络获得的问题表示来查询最近的上下文[32]

Fig. 10 Using question representations obtained by different column networks to query the nearest neighbors[32]

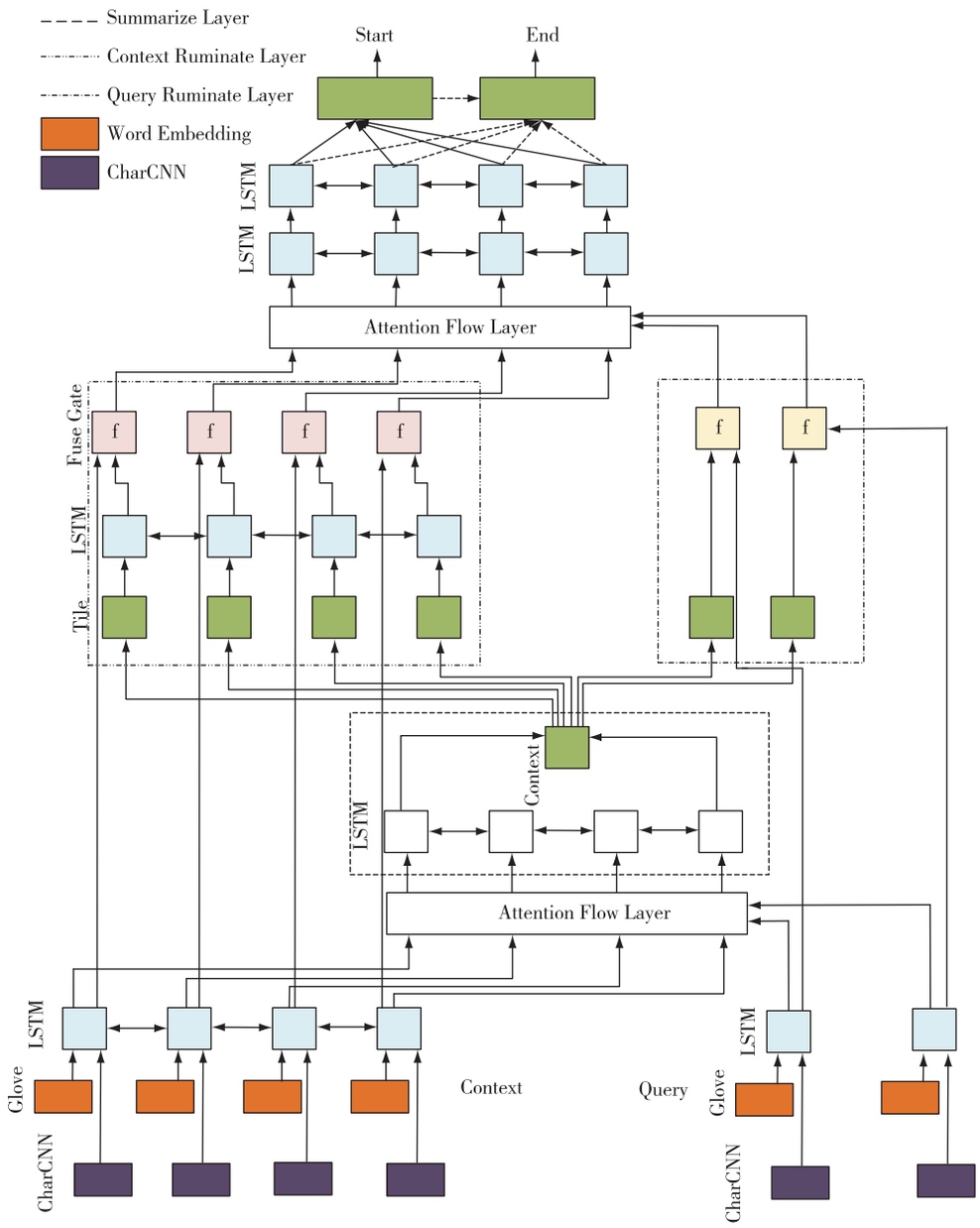

图 11　反思器模型结构[34]

Fig. 11　Model structure of Ruminating Reader[34]

参考文献

References

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

[2] Melamud O,Goldberger J,Dagan I.Context2vec:learning generic context embedding with bidirectional LSTM[C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning,2016:51-61

[3] Peters M E,Neumann M,Iyyer M,et al.Deep contextualized word representations[J]. arXiv Preprint, 2018, arXiv:1802.05365

[4] Radford A,Narasimhan K,Salimans T,et al.Improving language understanding by generative pre-training[EB/OL].[2019-10-12]. https://s3-us-west-2. amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf,2018

[5] Devlin J,Chang M W,Lee K,et al.Bert:pre-training of deep bidirectional transformers for language understanding [J]. arXiv Preprint, 2018, arXiv:1810.04805

[6] Dai Z,Yang Z,Yang Y,et al.Transformer-xl:attentive language models beyond a fixed-length context[J].arXiv Preprint,2019,arXiv:1901.02860

[7] Cao Y X,Huang L F,Ji H,et al.Bridge text and knowledge by learning multi-prototype entity mention em-

bedding[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers),2017:1623-1633
[8] Sarma P K,Liang Y,Sethares W A.Domain adapted word embeddings for improved sentiment classification[J]. arXiv Preprint,2018,arXiv:1805.04576
[9] Hotelling H.Relations between two sets of variates[J]. Biometrika,1936,28(3/4):321.
[10] Hardoon D R,Szedmak S,Shawe-Taylor J.Canonical correlation analysis:an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664
[11] Xu H,Liu B,Shu L,et al.Double embeddings and CNN-based sequence labeling for aspect extraction[J].arXiv Preprint,2018,arXiv:1805.04601
[12] Zhang Z,Han X,Liu Z,et al.ERNIE:enhanced language representation with informative entities[J]. arXiv Preprint,2019,arXiv:1905.07129
[13] Faruqui M,Dodge J,Jauhar S K,et al.Retrofitting word vectors to semantic lexicons[J].arXiv Preprint,2014, arXiv:1411.4166
[14] Kenter T,Borisov A,De Rijke M.Siamese CBOW:optimizing word embeddings for sentence representations[J]. arXiv Preprint,2016,arXiv:1606.04640
[15] Arora S,Liang Y,Ma T.A simple but tough-to-beat baseline for sentence embeddings[C]//International Conference on Learning Representations,2017
[16] Zhang Y,Liu Q,Song L.Sentence-state LSTM for text representation[J]. arXiv Preprint, 2018, arXiv: 1805.02474
[17] Kim Y.Convolutional neural networks for sentence classification[J].arXiv Preprint,2014,arXiv:1408.5882
[18] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proceedings of COLING 2014,the 25th International Conference on Computational Linguistics: Technical Papers, 2014:69-78
[19] Kiros R,Zhu Y,Salakhutdinov R R,et al.Skip-thought vectors[C]//Advances in Neural Information Processing Systems,2015:3294-3302.
[20] Hill F,Cho K,Korhonen A.Learning distributed representations of sentences from unlabelled data[J].arXiv Preprint,2016,arXiv:1602.03483
[21] Xu K,Feng Y S,Huang S F,et al.Semantic relation classification via convolutional neural networks with simple negative sampling[J]. arXiv Preprint, 2015, arXiv: 1506.07650
[22] Su Y,Liu H L,Yavuz S,et al.Global relation embedding for relation extraction[J].arXiv Preprint, 2017, arXiv: 1704.05958
[23] Chen Z Y, Zha H W, Liu H L, et al. Global textual relation embedding for relational understanding[J].arXiv Preprint,2019,arXiv:1906.00550
[24] Mintz M,Bills S,Snow R,et al.Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-ACL-IJCNLP,2009:1003-1011
[25] Qin L H,Zhang Z S,Zhao H.A stacking gated neural architecture for implicit discourse relation classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2263-2270
[26] Guo F Y,He R F,Jin D,et al.Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning[C]//Proceedings of the 27th International Conference on Computational Linguistics,2018:547-558
[27] Jia Y Y,Ye Y,Feng Y S,et al.Modeling discourse cohesion for discourse parsing via memory network[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2:Short Papers), 2018:438-443
[28] Ji Y F,Eisenstein J.One vector is not enough:entity-augmented distributed semantics for discourse relations[J]. Transactions of the Association for Computational Linguistics,2015,3:329-344
[29] Dai Z Y, Huang R H. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph[J].arXiv Preprint,2018, arXiv:1804.05918
[30] Bai H X, Zhao H. Deep enhanced representation for implicit discourse relation recognition[J]. arXiv Preprint,2018,arXiv:1807.05154
[31] Chen Q,Zhuo Z,Wang W.BERT for joint intent classification and slot filling[J].arXiv Preprint,2019,arXiv: 1902.10909
[32] Dong L,Wei F R,Zhou M,et al.Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1:Long Papers),2015:260-269
[33] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[J].arXiv Preprint,2016,arXiv:1611.01603
[34] Gong Y C, Bowman S R. Ruminating reader: reasoning with gated multi-hop attention[J].arXiv Preprint,2017, arXiv:1704.07415

南京信息工程大学学报(自然科学版),2019,11(6):706-715
Journal of Nanjing University of Information Science and Technology(Natural Science Edition),2019,11(6):706-715

Research progress on text feature extraction

ZENG Mingrui[1] YUAN Mengqi[1] SHAO Xi[1] BAO Bingkun[1] XU Changsheng[1,2]

1 School of communication and information engineering, Nanjing University of Posts and Telecommunications, Nanjing 210044

2 Institute of Automation, Chinese Academy of Sciences Institute of Automation, Chinese Academy of Sciences, Beijing 100190

Abstract Text understanding is an important research branch in artificial intelligence, which avails the effective interaction between human and computer with natural language. Text feature extraction is one of the basic and key steps for computers to understand and perceive the textual data. In this paper, we introduce the development history of text feature extraction and the mainstream feature extraction methods in recent years, and prospects the future research directions of text feature extraction. The three semantic hierarchies, namely word representation, sentence representation and discourse relationship mining are elaborated, then a case is given to show the typical application of text feature extraction on question answering system.

Key words natural language processing; text feature extraction; question answering system