

知识图谱增强的科普文本分类模型

唐望径^{1,2}, 许 斌^{1*}, 全美涵¹, 韩美奂^{1,3}, 王黎明⁴, 钟 琦⁴

(1. 清华大学 计算机科学与技术系, 北京 100084; 2. 北京交通大学 计算机与信息技术学院, 北京 100044;

3. 清华大学 深圳国际研究生院, 广东 深圳 518055; 4. 中国科普研究所, 北京 100081)

(* 通信作者电子邮箱 xubin@tsinghua.edu.cn)

摘 要: 科普文本分类是将科普文章按照科普分类体系进行划分的任务。针对科普文章篇幅超过千字, 模型难以聚焦关键信息, 造成传统模型分类性能不佳的问题, 提出一种结合知识图谱进行两级筛选的科普长文本分类模型, 来减少主题无关信息的干扰, 提升模型的性能。首先, 采用四步法构建科普领域的知识图谱; 然后, 将该知识图谱作为距离监督器, 并通过训练句子过滤器来过滤掉无关信息; 最后, 使用注意力机制对过滤后的句子集做进一步的信息筛选, 并实现基于注意力的主题分类模型。在所构建的科普文本分类数据集(PSCD)上的实验结果表明, 基于领域知识图谱的知识增强的文本分类算法模型具有更高的 F1-Score, 相较于 TextCNN 模型和 BERT 模型, 在 F1-Score 上分别提升了 2.88 个百分点和 1.88 个百分点, 验证了知识图谱对于长文本信息筛选的有效性。

关键词: 科普文本分类; 知识图谱; 两级筛选; 长文本分类; 注意力

中图分类号: TP18 文献标志码: A

Popular science text classification model enhanced by knowledge graph

TANG Wangjing^{1,2}, XU Bin^{1*}, TONG Meihan¹, HAN Meihuan^{1,3}, WANG Liming⁴, ZHONG Qi⁴

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;

2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;

3. Tsinghua Shenzhen International Graduate School, Shenzhen Guangdong 518055, China;

4. China Research Institute for Science Popularization, Beijing 100081, China)

Abstract: Popular science text classification aims to classify the popular science articles according to the popular science classification system. Concerning the problem that the length of popular science articles often exceeds 1 000 words, which leads to the model hard to focus on key points and causes poor classification performance of the traditional models, a model for long text classification combining knowledge graph to perform two-level screening was proposed to reduce the interference of topic-irrelevant information and improve the performance of model classification. First, a four-step method was used to construct a knowledge graph for the domains of popular science. Then, this knowledge graph was used as a distance monitor to filter out irrelevant information through training sentence filters. Finally, the attention mechanism was used to further filter the information of the filtered sentence set, and the attention-based topic classification model was completed. Experimental results on the constructed Popular Science Classification Dataset (PSCD) show that the text classification algorithm model based on the domain knowledge graph information enhancement has higher F1-Score. Compared with the TextCNN model and the BERT (Bidirectional Encoder Representations from Transformers) model, the proposed model has the F1-Score increased by 2.88 percentage points and 1.88 percentage points respectively, verifying the effectiveness of knowledge graph to long text information screening.

Key words: popular science text classification; knowledge graph; two-level screening; long text classification; attention

0 引言

文本分类是信息处理和数据挖掘的重要研究方向,是自

然语言处理(Natural Language Processing, NLP)的核心任务之

一, 广泛应用于新闻分类、舆情分析、推荐系统等领域。随着

收稿日期: 2021-07-16; 修回日期: 2021-09-07; 录用日期: 2021-09-07。

作者简介: 唐望径(1996—), 男, 海南海口人, 硕士研究生, CCF 会员, 主要研究方向: 自然语言处理、知识图谱; 许斌(1973—), 男, 海南海口人, 副教授, 博士, CCF 会员, 主要研究方向: 知识图谱、数据挖掘; 全美涵(1995—), 女, 河南三门峡人, 博士研究生, CCF 会员, 主要研究方向: 知识工程、信息抽取; 韩美奂(1997—), 女, 内蒙古赤峰人, 硕士研究生, 主要研究方向: 知识工程、信息抽取; 王黎明(1982—), 男, 内蒙古巴彦淖尔人, 副研究员, 博士, 主要研究方向: 科学传播的社会网络分析方法; 钟琦(1970—), 女, 山西太原人, 研究员, 硕士, 主要研究方向: 大众传媒上的科学传播评估。

互联网的发展,网络科普资源呈指数级增长,与此同时,海量的科普文本对人工进行文本分类提出严峻的挑战。一方面,传统人工识别文本类型的方法耗时耗力、效率低下^[1];另一方面,由于科普文章涉及领域广泛,科普工作者对科普文章类型识别往往受到自身专业领域知识的限制,难以掌握科普所有领域。因此,利用计算机智能辅助专家对文本进行自动分类,能够帮助科普工作者快速、准确地对文章归档。

早期浅层学习模型通常使用人工特征工程的方法获取适当的文本特征来表征样本,之后将特征输入到机器学习分类模型中,如朴素贝叶斯(Naive Bayes, NB)^[2]、支持向量机(Support Vector Machine, SVM)^[3]、和 K 近邻(K -Nearest Neighbor, KNN)^[4]等。浅层学习模型需要设计特征提取方法,往往代价昂贵。

随着深度学习理论及计算机硬件的发展,深度学习模型在文本分类领域得到广泛应用。深度学习模型将特征工程集成到模型拟合过程中,简化了分类流程,降低了模型的成本开销。深度学习常用模型主要有卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)以及近些年效果表现优异的预训练模型 BERT (Bidirectional Encoder Representation from Transformers)^[5]。Kim^[6]在2014年首次提出用于文本分类的 CNN 模型,其表现与复杂的基于特征学习的分类器模型相比毫不逊色,激发了深度学习文本分类模型的研究热潮。然而,当文本长度较长时, CNN 难以提取较好的全局特征。RNN 能够对复杂的单词序列学习特征,有利于在较长的文本中提取单词的全局时序特征,但是存在梯度消失和梯度爆炸等问题。预训练神经网络模型通过在大规模无监督的语料上预训练,再在文本分类任务上微调从而达到目前最优的效果;但是预训练神经网络模型输入的长度受限,无法捕获长文本语义,同时由于是端到端模型,缺乏知识的引导。

基于上述模型的问题,针对科普文章篇幅超过千字导致模型难以聚焦分类关键信息、性能不佳的问题,本文提出了知识图谱增强的科普文本分类模型。首先,采用四步法^[7]构建了科普领域知识图谱,之后将图谱中的实体信息作为远程监督器,训练一个二分类模型作为信息过滤器,对句子进行打分、排序并完成筛选,从而使得到的句子特征表示包含更多分类类别相关信息,以缓解层次模型无法突出关键词义信息的缺陷。然后,为了进一步降低主题无关信息干扰,提高分类模型性能,本文引入句级别注意力模型,对筛选后的句子信息进行二次过滤。最后,在构建的科普文本分类数据集(Popular Science Classification Dataset, PSCD)上进行对比实验,根据实验结果,本文所提模型取得良好的分类效果,验证了模型的有效性。相较于传统神经网络模型,本文模型在召回率和 F1 指标上有所提升。

本文的主要工作如下:1)提出了知识图谱增强的科普文本分类模型,通过科普知识图谱来进行主题句筛选,并引入句级别注意力模型,最终通过一个全连接神经网络实现科普文本分类;2)构造了一个科普文本分类数据集,包含了13 372篇科普文章,且本文模型在该数据集上取得了最优效果。

1 相关工作

1.1 文本分类

传统的 NB 和 SVM 等文本分类模型通常需要将分类流程划分为特征工程设计和分类器设计两部分,构建代价昂贵。随着深度学习的发展,深度学习模型将特征提取融合进模型拟合过程中,使得模型能够直接从输入中学习特征表示,无需过多的人工干预和先验知识。此外,由于深度学习模型具有更高的复杂度,相较于浅层模型具备更优异的效果。TextCNN^[6]作为一次成功的尝试,激发了研究人员在文本分类领域应用 CNN 的热情。由于 TextCNN 无法通过卷积获取文本的长距离依赖关系,为解决该问题,Johnson 等^[8]提出了深度金字塔 CNN (Deep Pyramid CNN, DPCNN),通过不计成本地增加网络深度,模型能够更精准地抽取长距离的文本依赖关系。RNN 能够通过递归计算捕获长距离依赖关系,但存在梯度消失和梯度爆炸问题。人们通常使用 RNN 的改进模型长短期记忆(Long Short-Term Memory, LSTM)^[9]网络。RNN 能够更好地捕捉上下文信息,但对于局部信息的提取能力欠佳。Lai 等^[10]提出了用于文本分类任务的循环 CNN (Recurrent CNN, RCNN)模型,利用双向 LSTM 结构替换 CNN 中的卷积层,对文本进行编码,较好地捕获了长文本上的序列信息;同时,通过 CNN 的最大池化层,较好地捕获了文本的局部特征。

1.2 注意力机制

神经网络能够很好地与注意力机制结合,在文本建模时将注意力聚焦于分类的关键信息上,从而提高文本分类效果。注意力机制在计算机视觉领域首先被提出,随后因被应用到机器翻译的端到端模型而引入 NLP 领域。注意力机制通过目标向量对输入序列进行打分,将注意力聚焦于输入序列中更为重要的部分,使输出结果更加精确。因此,注意力机制逐渐被应用于文本分类等 NLP 任务中。

在处理由许多句子组成的长文本时,直接将文本作为长序列进行处理往往容易忽略文本层级结构中蕴含的信息,因此 Yang 等^[11]提出了基于层级注意力网络(Hierarchical Attention Network, HAN)的文本分类模型,该模型以双向门控循环单元(Gate Recurrent Unit, GRU)^[12]作为编码器,通过两层注意力机制使模型更好地捕获文本中的重要信息,提高了长文本分类的模型性能表现。文献[13]通过一种无监督的段落向量生成模型 PV-DM (Distributed Memory model of Paragraph Vectors)^[14],基于结合注意力的 CNN 分类模型在长文本分类任务中取得了较好的效果。Choi 等^[15]以 ALBERT (A Lite BERT)^[16]作为编码器,通过类别表示层提取分类的类别信息,使用类别信息对句子信息进一步增强,之后通过句注意力机制对句子信息进行筛选,获取了文本重要信息,有效提高了长文本分类模型性能。

2 本文模型

图1为本文所提知识图谱增强的科普文本分类模型总体架构。

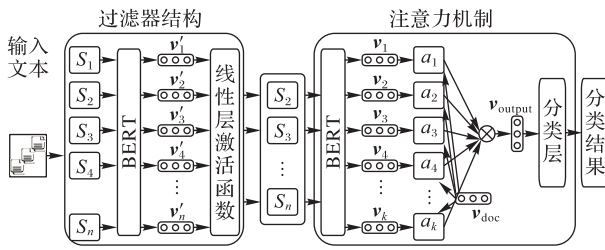


图1 知识图谱增强的科普文本分类模型架构

Fig. 1 Architecture of popular science text classification model enhanced by knowledge graph

通常,语料中的所有文本表示为 $T = \{t_1, t_2, \dots, t_i, \dots, t_{|T|}\}$, t_i 表示语料中第 i 篇文本, $|T|$ 表示语料中包含的文本数量;每篇文本可以被表示为 $t_i = \{S_1, S_2, \dots, S_j, \dots, S_n\}$, S_j 表示文本中的第 j 个句子, $n = |t_i|$ 表示文本中包含的句子数量;经过特征提取后,句子 S_j 被表示为句向量 v_j ,句向量维度为 768。

本文分类模型的建立流程如图2所示。

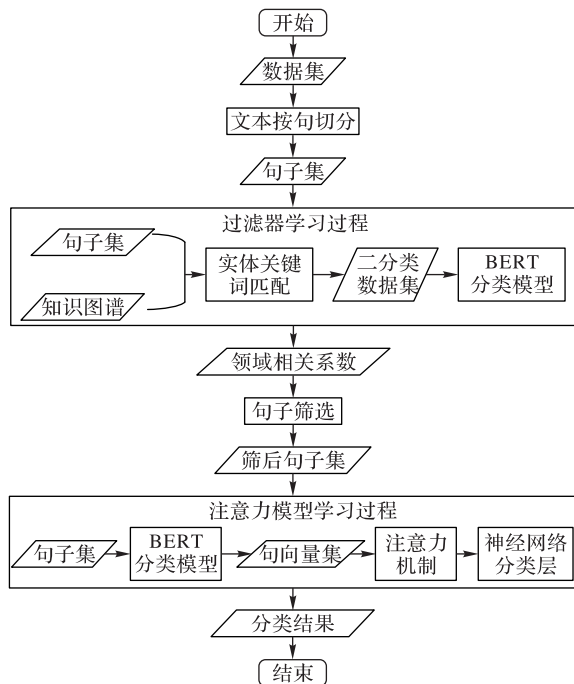


图2 模型流程

Fig. 2 Model flowchart

流程主要由以下四部分组成:

1) 科普领域知识图谱构建。借鉴“四步法”构造科普领域知识图谱。

2) 文本信息筛选。首先将文本切分为句子集,使用预训练模型 BERT 作为编码器,将句子编码为连续空间下的句向量,以实现文本句子特征提取。以二分类器作为过滤器,利用知识图谱中蕴含的实体信息对文本句向量进行筛选过滤,实现文本的信息筛选。

3) 注意力机制。使用注意力机制对信息筛选后结果做进一步的增强。

4) 通过全连接层和 Softmax 函数实现科普文本分类。

2.1 构建科普领域知识图谱

本文借鉴了领域知识图谱构建方法“四步法”。首先,对果壳网、环球科学、Science 科学等共 68 家科普微信公众号进行数据采集,获取近 20 万篇科普文章。构建领域图谱首先需要进行本体构建,通过使用统计方法结合人工归纳、借鉴高质量通用图谱和专家指导构建了图谱本体。首先通过词频—逆文本频率(Term Frequency-Inverse Document Frequency, TF-IDF)算法^[17]和 TextRank 算法^[18]提取采集文章的关键词,同时通过 K-Means 聚类算法^[19]对文章进行聚类,通过人工观察关键词和聚类结果,得到了初步的领域概念。之后结合百度百科、Wikipedia 等通用图谱对领域概念进行修改并定义领域关系及其约束。其中:一级概念共 10 个,分别为军事、农业、科学文化、医学、健康、生态环境、信息科技、空间科学、基础学科、生活百科;二级概念共 45 个。最后,根据中国科普研究所专家指导,对构建本体进行检查和评估,修改完善后得到最终的科普领域本体。根据定义好的科普领域本体结构,从互联网上公开的知识图谱(如:百度百科、Wikipedia、XLORE 等)以及其他结构化较好的网站中获取大量结构较好、质量较高的实体数据,并结合半监督和远程监督方法,通过人工筛选获取实体间关系。

2.2 文本信息筛选

文本信息筛选是本文模型对文本信息进行提取过滤的部分,本节将按顺序分别介绍以下两步:第一步,以预训练模型 BERT 作为编码器对文本中所有句子进行特征提取;第二步,以科普知识图谱为监督源构建二分类过滤器,使用构建的过滤器计算句子领域相关系数,并按照该系数对每篇文本句子进行定量筛选。

2.2.1 特征提取

Google 的 Devlin 等^[5]在 2018 年提出了大规模预训练模型 BERT,在特征提取任务上取得了极佳效果。BERT 模型由多层双向的 Transformer^[20]解码器构成,Transformer 模型的核心是注意力机制。BERT 采用多头注意力(Multi-Head Attention)机制,能够更好地获取目标字在多种语义场景下与其上下文构造的语义信息。 H 的计算公式为:

$$H = \text{Concat}(h_1, h_2, \dots, h_i) * W^0 \quad (1)$$

其中: W^0 表示权重矩阵,为输出向量重新分配权重; h_i 表示第 i 个头的输出向量。 h_i 的计算公式为:

$$h_i = \text{Attention}(Q_i, K_i, V_i) \quad (2)$$

$$\text{Attention}(Q_i, K_i, V_i) = \left(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}} \right) * V_i \quad (3)$$

其中: Q_i, K_i, V_i 表示输入向量经过线性映射后得到的矩阵; d_k 表示 K 向量的维度。

2.2.2 信息过滤

本文通过四步法构建了科普领域知识图谱,该图谱的实体类别与科普文本分类类别一致。首先将科普文本分类数据集测试集按照 7:3 比例划分出用于训练信息过滤二分类器的训练集和测试集,对所得训练集和测试集进行句子切分,得到以句子为单位的训练句子集和测试句子集。之

后,将科普图谱中实体与上述所得句子集中句子进行链接,为句子赋予标签 $L_j = [0, 1]$,实体链接成功句子标签赋1,链接失败句子标签赋0。通过上述预处理,将语料测试集转化为如下所示数据二元组:

〈“天王星和海王星都有一个完全倾斜的磁场”,“1”〉

其中,“天王星”和“海王星”均为科普图谱中实体,故该句子实体链接成功,其对应标签为1。之后,使用构建好的数据二元组〈句子 S_j ,二分类标签 L_j 〉训练一个二分类模型作为句子过滤器。其模型结构如图3所示。

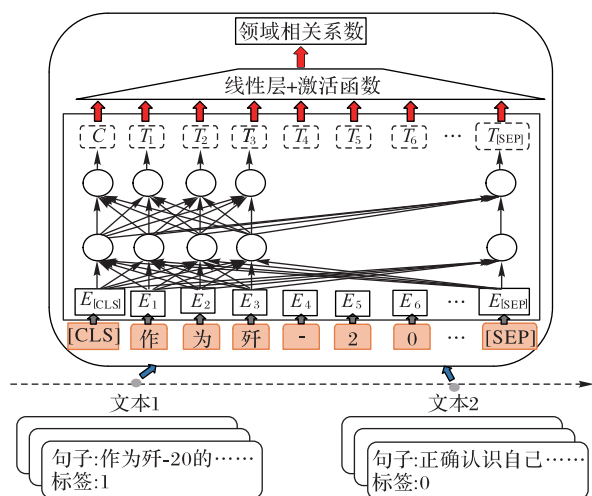


图3 句子过滤器结构

Fig. 3 Structure of sentence filter

将输入文本切分为句子集合,集合中句子 S_j 输入训练好的二分类模型后,模型通过预训练模型BERT对句子进行特征提取,提取的特征信息将会通过输出层得到输出。输出层包括线性层、激活函数,计算公式如下所示:

$$p = \tanh(WH + b) \quad (4)$$

线性层维度为 $[768, 768]$,为随机初始化的全连接层。为了避免出现线性模型表达能力不足的问题,将线性层输出结果输入激活函数,以此为模型添加非线性因素。将经过激活函数后结果输入Softmax函数进行归一化处理,取标签 $L_j = 1$ 的概率,将其视为领域相关系数 $q_j \in [0, 1]$,计算公式如下:

$$q_j = \frac{\exp[p_j(L_j = 1)]}{p_j(L_j = 1|S_j) + p_j(L_j = 0|S_j)} \quad (5)$$

按照领域相关系数 q_j 对相应的句向量 v_i 进行排序,选出分数值排名靠前的 n 个句向量 $V_{input} = [v_1; v_2; \dots; v_n]$ 作为下一部分注意力模型输入。同时,保持了各句向量在原文本中的先后顺序,以避免造成位置信息丢失,影响模型分类性能。

2.3 注意力机制

文章中不同的句子对于文章内容理解的贡献程度不同。为计算句向量对于区分文章类别的贡献程度,本文引入了注意力机制,其具体结构如图4所示。

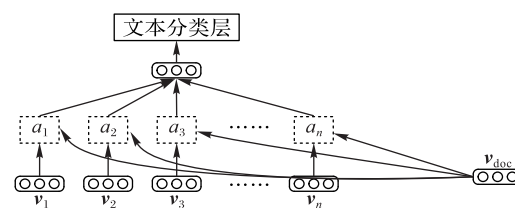


图4 注意力机制结构

Fig. 4 Structure of attention mechanism

注意力模型的计算公式如下:

$$v_{doc} = \frac{1}{n} \sum_{i=1}^n v_i \quad (6)$$

$$v_{output} = \sum_{i=1}^n a_i v_i \quad (7)$$

$$a_i = \frac{\exp\left(\frac{(Wv_i, v_{doc})}{\sqrt{d}}\right)}{\sum_{j=1}^n \exp\left(\frac{(Wv_j, v_{doc})}{\sqrt{d}}\right)} \quad (8)$$

其中:文本向量 v_{doc} 由向量集中所有向量相加求平均后得到; v_{output} 为经过注意力模型后得到的输出向量; $V_{input} = [v_1; v_2; \dots; v_n]$ 为经过句子过滤器筛选过滤后得到句子集通过句向量编码器BERT后得到的句向量集合, v_i 为此向量集中第 i 个句向量; W 为随机初始化的全连接层; d 为词向量维度,模型中 $d = 768$ 。由于使用乘法计算注意力,添加了 $\sqrt{d}$ 作为注意力模型计算的缩放因子,避免因乘法所得结果较大,导致Softmax函数结果进入“饱和区”,使得梯度过小,从而对注意力计算造成影响。

2.4 文本分类

通过注意力机制,得到了句向量对应分类的注意力。通过注意力对句向量进行聚合,得到 v_{output} ,该向量包含了经过过滤器和注意力机制双重筛选后的文本信息。将 v_{output} 输入由全连接层和Softmax函数构成的多分类器中,得到类别概率:

$$p = \text{Softmax}(Wv_{doc} + b) \quad (9)$$

通过最小化分类的交叉熵误差,以有监督的方式训练模型。损失函数如下:

$$\text{loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} p_{i,c} \quad (10)$$

其中: N 为样本数; C 为多分类类别总数;若预测类别和样本所属类别相同则 $y_{i,c}$ 为1,否则为0; $p_{i,c}$ 为文本 t_i 属于类别 c 的概率。

3 实验与结果分析

3.1 实验数据

为了验证模型效果,本文构建了科普文本分类数据集PSCD。首先,对照科普领域知识图谱实体分类,从科普中国等国内著名科普网站中爬取相应栏目的科普文章。剔除内容过短和过长的科普文本,对筛选后文本进行去重和数据清洗,最终得到科普文本分类数据集。科普长文本分类数据集中单个文本长度为102~26 722。数据集中包含10个分类,

共 13 372 篇文本,其中短文本 730 篇,长文本 11 195 篇,超长文本 1 447 篇。实验随机选取 80% 作为训练数据,20% 作为测试数据。数据集具体细节统计如表 1 和图 5 所示。

表 1 数据集详细情况

Tab. 1 Dataset details

类别	样本数	平均长度	最长长度	最短长度
基础学科	3 156	2 711	26 722	120
健康科普	2 935	2 459	19 970	102
军事科普	1 074	2 661	22 750	130
科学文化	1 230	1 976	15 408	182
空间科学	900	2 641	19 646	134
农业科普	425	1 720	16 899	207
生活百科	232	2 906	19 962	148
生态环境	857	3 037	16 816	148
信息科技	1 943	3 003	24 504	148
医学科普	620	1 063	18 860	102

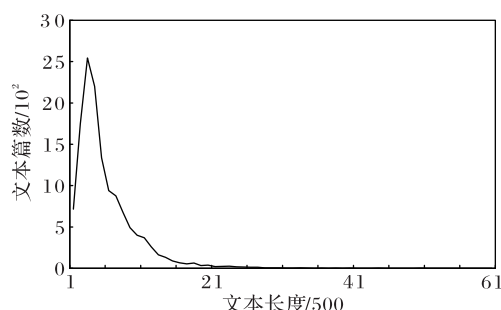


图 5 输入文本长度分布统计情况

Fig. 5 Statistics on length distribution of input text

3.2 基线方法及实验设置

3.2.1 基线方法

为评价模型在长文本分类任务上的性能,引入了多种典型模型进行实验对比,以下详细描述了所有进行比较的模型:

1) FastText^[21]: 该模型是一个使用浅层神经网络实现的快速文本分类算法,不需要使用预训练词向量,模型会在训练过程中训练词向量。该模型分类效果较好,需要的训练时间极少,是一种高效的工业级分类模型。

2) TextCNN: 该模型是 CNN 在 NLP 中的一次成功的应用,能够利用 CNN 对局部特征捕捉的特性,获取不同抽象层次的语义信息。

3) TextRNN^[22]: 该模型是一种 RNN,将经过预训练词向量模型后的词向量表示按顺序输入双向 LSTM 中,将 LSTM 最后一个时间步长中的隐藏状态输入 Softmax 层后输入全连接层进行分类,最终输出文本分类。

4) TextRCNN^[10]: RNN 能够更好地捕捉上下文信息,而 CNN 能够有效地捕捉局部特征。该模型结合了 RNN 和 CNN 的特点,将 CNN 中的卷积层换成了双向的循环结构,使其能够有效减少噪声,最大限度地捕捉上下文信息。同时该模型保留了最大池化层,保证了模型对局部特征的提取能力。

5) BERT^[5]: 该模型为 2018 年谷歌提出的大规模预训练模型,模型框架基于 Transformer,实现了多层双向 Transformer 编码器。该模型一经发布,就在多项 NLP 任务中

取得了 SOTA 结果。虽然该模型主要任务并不是文本分类,但由于其优秀的文本表征能力,本文使用中文 BERT 预训练模型作为编码器,将结果输入 Softmax 函数和全连接层,实现文本分类任务。

3.2.2 实验设置

所有基线模型均为基于字级别的分类模型。由于 BERT 输入长度限制为 510 字,且原数据集中文本长度较长,故对 BERT 模型输入文本进行预处理,对长度超过 510 字文本,从文本开头截取长度为 510 字文本作为模型输入,对长度未超过限制的文本则不做处理。

对比实验中,TextCNN、TextRNN、TextRCNN 均使用预训练词向量模型 Word2Vec 生成词向量,设置单词嵌入维度为 300,文本长度均未作处理。

实验运行服务器配置为 Intel Core i9-10900K CPU,显卡为 RTX 3090。

3.3 评测标准

本文采用准确率 Acc (Accuracy)、召回率 R (Recall) 以及 F1 值 (F1-score) 作为对本文分类模型效果的评价指标。计算公式如下:

$$Acc = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

其中: TP (True Positives) 表示预测为正的样本, FP (False Positives) 表示预测为正的负样本, FN (False Negatives) 表示预测为负的正样本, TN (True Negatives) 表示预测为负的负样本。

3.4 实验结果分析

表 2 为上述各基线模型与本文模型在科普文本分类数据集 PSCD 上分类表现。由表 2 可知,本文模型在平均召回率以及平均 F1 值上最高,相较于 TextCNN 和 BERT,分别提高了 2.88 和 1.88 个百分点;BERT 在所有基线模型中表现最佳,TextCNN、TextRCNN 均在数据集上取得了较好的分类效果。虽然 TextRNN 使用 LSTM 替代普通 RNN,解决了处理较长文本时会出现的梯度消失问题,但是由于科普文本中长文本数量多、字数长,该模型在 PSCD 上的分类表现仍较差。FastText 表现仅优于 TextRNN,但其模型的训练耗时明显少于其他模型,同时不需要预训练词向量,故其仍是一个非常优秀的分类模型。

表 2 PSCD 上各模型的表现

单位: %

Tab. 2 Performance of each model on PSCD unit: %

模型	平均准确率	平均召回率	平均 F1 值
FastText	82.22	76.54	78.88
TextCNN	82.84	79.51	80.71
TextRNN	61.42	55.46	56.69
TextRCNN	85.24	76.48	79.62
BERT	85.53	79.47	81.71
本文模型	85.47	82.65	83.59

为了更进一步探索模型的效果,故对测试集中分类错误案例进行了分析。当文本句子个数上限设置为30时,共在测试集中得到类别错误划分案例453例,其中,文本句子数量少于30的案例共272例,文本句子数多于或等于30的案例共181例。即文本中句子数量对于模型的分类效果造成了一定的影响。之后,对案例中各分类数量占比进行了统计,发现错误常常出现在涉及学科领域交叉应用的文章的分类中,如图6所示。

据外媒NewAtlas报道,以色列国防军(LDF)利用主战坦克的最新变种将人工智能带入装甲战斗中。Merkava Mk4 Barak被认为是第一款拥有先进人工智能和传感器系统的坦克,它可以完成许多日常任务,让士兵可以集中精力完成手头的任务。其先进的AI系统,它接管日常的乘员任务。根据以色列国防军的说法,这不仅可以使任务时间减少30%,而且还可以更快、更准确地实现瞄准攻击目标。此外,计算机可以与乘员和附近的其他坦克共享数据。另一项创新是IronView VR头盔,为士兵提供360度的外部环境视图。这不仅为指挥官提供了更好的态势感知,而且还为学员提供了更加逼真的训练体验。以色列国防军表示,这使得从训练到真正坦克任务的过渡变得更加容易。“装甲军团在不同地区结束了一段重要的作战成功,这证明了军团在保护以色列国及平民安全方面的重要性。”一名高级官员表示“新坦克使装甲军团在作战领域向前迈出了一步,并构成了一种创新武器,将改变地面作战的面貌。”事实上在军事领域应用先进的人工智能技术,一直以来是业界一个极为敏感的话题。人们担心,一旦主要军事大国开始推进人工智能武器的研发,一场全球性的自主军备竞赛将不可避免。这种技术轨道的重点是大规模生产的廉价杀人机器。7月19日以马歇尔和DeepMind的两位联合创始人为主的科技界最近签署了一份承诺书,内容是他们所属的公司在未来将不会研发拥有AI的智能武器。这是一个由科研人员和科技领域的高管组成的非正式全球联盟。此举主要是反对AI武器的研发。承诺中警告:在人类使用人工智能武器“选择和接触”目标时,如果决定涉及目标前没有人干预,那么最终结果将会导致道德沦丧,更严重的结果将会对现实生活产生危害。签署此次承诺书的人表示:“从道义上讲,不应该把结束人类生命的决定委托给机器。而实际上,如果此类武器在未来得到扩散,那么最终的结果将危及每个国家甚至牵涉到每个人的生命安全。”谷歌曾助力五角大楼开发致命的人工智能无人机,之后遭到了员工们的集体抗议。几周后,公司发布了公告,承诺以后不会开发AI和武器相关的系统。因为谷歌对AI武器的态度,韩国开普特大学的学生曾经对AI武器方面进行了研发,由于受到外部的强烈抵制,最终开普特的校长承诺不会开发“违背人类医院,以及没有认为控制的自动化武器”。

图6 错误案例

Fig. 6 Error case

该文章开头提及了以色列国防军配备了一种装载人工智能系统的坦克,并在后续对该坦克进行了详细介绍,之后花费了大量篇幅对军事领域内容进行了描述。文章中大量出现“国防军”“坦克”“装甲战斗”等军事领域词汇和科普图谱中属于军事科普分类下的“Merkava Mk4 Barak 坦克”“IronView VR 头盔”等军事实体。虽然该文本主要内容为人工智能技术在军事领域应用及相关的伦理问题,应属于信息科技分类,但大量的军事领域词汇和军事领域实体形成的噪声误导了分类器,以至于将本属于信息科技分类下的文本划分到军事科普分类中。类似情况在错误案例中频繁发生。

上述结果表明,利用知识图谱实体信息构建的过滤器,对长文本进行句子筛选,能够有效地提取包含领域相关实体、领域词汇句子,对长文本分类模型性能有所提升。

4 结语

本文研究基于领域知识图谱的文本分类模型,提出了知识图谱增强的科普文本分类模型。首先,将输入文本切分为句子集,使用BERT作为句子编码器,将所有句子转化为句向量集;然后,构建了一个以科普领域图谱实体信息作为监督数据源,科普文本句子集作为输入的二分类过滤器,基于过滤器,对输入文本进行筛选,输出固定数量且排名靠前的句子向量;最后,通过注意力模型获得对文本信息高度总结的输出向量,将其输入分类层获得文本对应类别概率,以实现文本分类。通过在科普分类数据集PSCD进行对比实验表明,所提出模型在分类性能优于基线模型中召回率最高的TextCNN模型以及F1-Score最高的BERT模型,与BERT模型相比,召回率和F1-Score分别提升了3.18个百分点和1.88个百分点。对错误案例的分析可知,仅利用实体及类别信息不足以解决交叉领域中次要领域信息对分类造成影响的问题。在后续工作中,将尝试通过引入更多图谱信息解决上述问题。

致谢:此项工作得到了中国科普研究所2020年委托合作项目“自

然语言处理方法在科普领域的应用研究”(200110EMR028)支持。

参考文献 (References)

- [1] 庞观松,蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践, 2012, 35(2): 123-128. (PANG G S, JIANG S Y. A summary of research on automatic text classification technologies [J]. Information Studies: Theory and Application, 2012, 35(2): 123-128.)
- [2] LEWIS D D. Naive (Bayes) at forty: the independence assumption in information retrieval [C]// Proceeding of the 1998 European Conference on Machine Learning, LNCS 1398. Berlin: Springer, 1998: 4-15.
- [3] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.
- [4] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.
- [5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.
- [6] KIM Y. Convolutional neural networks for sentence classification [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.
- [7] 杨玉基,许斌,胡家威,等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报, 2018, 29(10): 2931-2947. (YANG Y J, XU B, HU J W, et al. Accurate and efficient method for constructing domain knowledge graph [J]. Journal of Software, 2018, 29(10): 2931-2947.)
- [8] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2017: 562-570.
- [9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [10] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification [C]// Proceeding of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2015: 2267-2273.
- [11] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 1480-1489.
- [12] CHO K, MERRIENBOER B van, GULÇEHRE Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,

- PA: Association for Computational Linguistics, 2014: 1724-1734.
- [13] 卢玲, 杨武, 王远伦, 等. 结合注意力机制的长文本分类方法[J]. 计算机应用, 2018, 38(5): 1272-1277. (LU L, YANG W, WANG Y L, et al. Long text classification combined with attention mechanism[J]. Journal of Computer Applications, 2018, 38(5): 1272-1277.)
- [14] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]// Proceedings of the 31st International Conference on International Conference on Machine Learning. New York: JMLR.org, 2014: 1188-1196.
- [15] CHOI G, OH S, KIM H. Improving document-level sentiment classification using importance of sentences[J]. Entropy, 2020, 22(12): No. 1336.
- [16] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. (2020-02-09) [2021-05-28]. <https://arxiv.org/pdf/1909.11942.pdf>.
- [17] SALTON G, YU C T. On the construction of effective vocabularies for information retrieval[J]. ACM SIGPLAN Notices, 1975, 10(1): 48-60.
- [18] MIHALCEA R, TARAU P. TextRank: bringing order into text[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2004: 404-411.
- [19] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles, CA: University of California Press, 1967: 281-297.
- [20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceeding of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 6000-6010.
- [21] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Stroudsburg, PA: Association for Computational Linguistics, 2017: 427-431.
- [22] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. California: IJCAI Organization, 2016: 2873-2879.
- TANG Wangjing**, born in 1996, M. S. candidate. His research interests include natural language processing, knowledge graph.
- XU Bin**, born in 1973, Ph. D., associate professor. His research interests include knowledge graph, data mining.
- TONG Meihan**, born in 1995, Ph. D. candidate. Her research interests include knowledge engineering, information extraction.
- HAN Meihuan**, born in 1997, M. S. candidate. Her research interests include knowledge engineering, information extraction.
- WANG Liming**, born in 1982, Ph. D., associate research fellow. His research interests include social network analysis method to scientific communication.
- ZHONG Qi**, born in 1970, M.S., research fellow. Her research interests include evaluation of science communication on mass media.