浙江理工大学学报，2022，47(3)：424-432
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.03.018

基于 BERT 模型的无监督中文单文本关键词提取模型

顾　淳，俞成海，于　洋，关炜炜
（浙江理工大学信息学院，杭州 310018）

摘　要：针对现有方法存在的忽略语义信息及重复提取语义相近关键词等问题，提出了一种基于 Bidirectional encoder representation from transformers(BERT)模型的无监督中文单文本关键词提取模型。该模型首先对待提取文本进行预处理以选取候选词，接着使用 BERT 模型的隐藏层结合全文信息获取候选词的词向量，然后加入聚类层筛除语义重复的候选词，最后获取全文语义向量并计算候选词与全文的语义的相似度评分，经排序后提取关键词。实验结果表明：将模型用于混合主题中文论文摘要等较短文本，在提取关键词的数量分别为 5 和 8 时，该模型的准确率分别为 34.21%和 26.34%，优于 Text Rank、TF-IDF 等传统提取模型，表明该模型通过融合语义信息提升了中文单文本关键词提取的准确率，改善了关键词重复提取的问题，使提取的关键词更加准确，有效提升了中文单文本关键词提取质量。

关键词：关键词提取；无监督；BERT 模型；文本向量化；单文本
中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 05-0424-09

Unsupervised keyword extraction model for Chinese single text based on BERT model

GU Chun, *YU Chenghai*, *YU Yang*, *GUAN Weiwei*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Since the existing methods have inadequate consideration of semantic information and extracts keywords with similar semantics, an unsupervised Chinese single text keyword extraction model was proposed based on the bidirectional encoder representation from transformers (BERT) model. Firstly, this model preprocessed the extracted text and selected candidate words from it, then used the hidden layer of BERT model to obtain the vector of candidate words in combination with full-text information, and then added a clustering layer to eliminate candidate words with similar semantics. Finally, the full-text semantic vector was obtained, and keywords were extracted by calculating the similarity score between the candidate word and full text. It was found that in short texts, for example, the use of the model in abstracts of Chinese papers on mixed topics, when the number of extracted keywords was 5 and 8, the accuracy of this method was up to 34.21% and 26.34%, which was better than traditional extraction models such as Text Rank and TF-IDF. This implied that the proposed model effectively improved the keyword extraction accuracy of Chinese single texts by fusing semantic information, improved the problem of repeated keyword extraction, made the extracted keywords more accurate, and effectively enhanced the extraction

收稿日期：2021－12－11　　网络出版日期：2022－04－05
基金项目：浙江省重点研发计划项目(2020C03094)
作者简介：顾　淳(1997－　)，男，江苏无锡人，硕士研究生，主要从事自然语言处理方面的研究。
通信作者：俞成海，E-mail：ych@zstu.edu.cn

quality of Chinese single text keywords.

Key words: keyword extraction; unsupervised; BERT model; text vectorization; single text

0 引　言

关键词提取任务是自动地在一篇文本中提取表达文本关键信息的一组词作为关键词。目前,关键词提取方法主要分为有监督提取和无监督提取两种[1]。有监督提取使用大量经过标记的语料信息,利用神经网络类[2]将关键词提取任务转换为二分类任务,评判候选词是否为关键词。这类方法的难点在于训练有监督关键词提取模型时需要大量标注语料,但在有些任务中获取语料较困难,且训练出的模型仅适用于同类型文本的关键词提取[3]。无监督提取包含两类,即基于语料库的无监督关键词提取和单文本关键词提取[4]。

基于语料库的无监督提取方法是利用待提取文本与其所在语料库中的语义信息构建主题模型,构建主题模型的过程中不需要提供额外标注,通过迭代建模语料库中的语义信息形成主题模型。该类方法的典型算法有 LSA、LDA 等。基于语料库的无监督提取方法的优势在于,充分利用了待提取文本所处语料库中的语义信息,使得提取的关键词更有意义;但存在的问题是,主题模型的效果受限于训练的语料环境,在实际提取任务中,有大量单独存在的待提取文本,而不存在于特定的语料环境中[2]。

单文本关键词提取方法是指不依赖语料库,直接从独立的文本中提取关键词。如基于词图网络的 RAKE[5]、Topic Rank[6]、Position Rank[7] 等算法,将候选词作为节点、特征权重作为边,构建词图网络模型,计算权重并提取关键词[8];Matsuo 等[9] 利用候选关键词的共现关系、词频等特征提取关键词。此类方法虽能够提取单文本中的关键词,但仅利用了文本的特征,未将文本的语义信息作为筛选关键词的依据,而被忽略的语义信息往往对关键词提取的准确度有着较大影响[10]。Gagliardi 等[11] 使用 Word2Vec 模型将候选词向量化,Mahata 等[12] 使用 fastText 训练候选词向量,Bennani-Smires 等[13] 使用 Sent2Vec 及 Doc2Vec 模型获取候选词及全文的语义向量。此类方法通过计算候选词向量与主题的语义相似度,排序并提取关键词。此类方法存在的问题在于,虽然将词向量化可获取语义信息,但是使用的向量化模型都为静态向量化模型,在语义环境变化较大的情况下会有较大的误差;同时,计算相似度的方法可以看作是对候选词向量与主题向量的距离进行排序的过程,语义相近的候选词与主题的相似度也较大,因此在提取的过程中会出现重复提取语义相近的关键词的问题。

针对现有方法存在的忽略语义信息及重复提取语义相近关键词的问题,本文提出一种基于 Bidirectional encoder representation from transformers (BERT)模型的无监督中文单文本关键词提取模型。该模型首先对原始文本预处理以选取候选词,接着使用 BERT 模型结合全文语义信息,将候选词输入隐藏层获取动态词向量,考虑 BERT 模型隐藏层对语义表示效果的影响,增强对文本语义信息的利用。然后,通过加入聚类层筛选重复候选词,以解决重复提取的问题。最后使用全文语义向量获取的方法,将全文语义向量化,并计算该向量与候选词向量的相似度评分,对评分排序,提取评分高的候选词向量所对应的候选关键词作为提取关键词的结果。由于本文模型利用 BERT 模型的动态词向量中包含的语义,并使用聚类层对重复提取进行改进,关键词提取的效果将得到提升。

1 模型设计

本文提出的关键词提取模型示意如图 1 所示,其中:w_i 表示使用 Jieba 工具包从原始文本中拆分出的单词;$i=1,2,3,\cdots$,表示每个词所在的位置;h_i 表示对 w_i 预处理后产生的候选词;v_i 表示候选词的词向量;v_d 表示全文语义向量;m 为预处理后得到的候选词总数。关键词提取步骤为:首先,将原始文本切分成长度为 n 的词序列$\{w_i\}$,经过文本预处理筛除部分词,得到 m 个候选词 h_i;然后,将候选词输入 BERT 模型将词向量化获得对应的词向量 v_i;接着,对得到的词向量集合进行候选关键词去重处理,得到候选关键词向量;同时,对全文使用 BERT 模型全文向量化的方法获得全文语义向量;最后,计算候选关键词向量与全文语义向量之间的相似度,并排序、提取目标关键词,完成全部提取过程。该模型的优势在于,通过预训练模型生成的候选词向量是动态的,并结合了全文语义信息来筛选关键词,同时解决了使用相似度在关键词选取任务中产生的词义重复的问题。

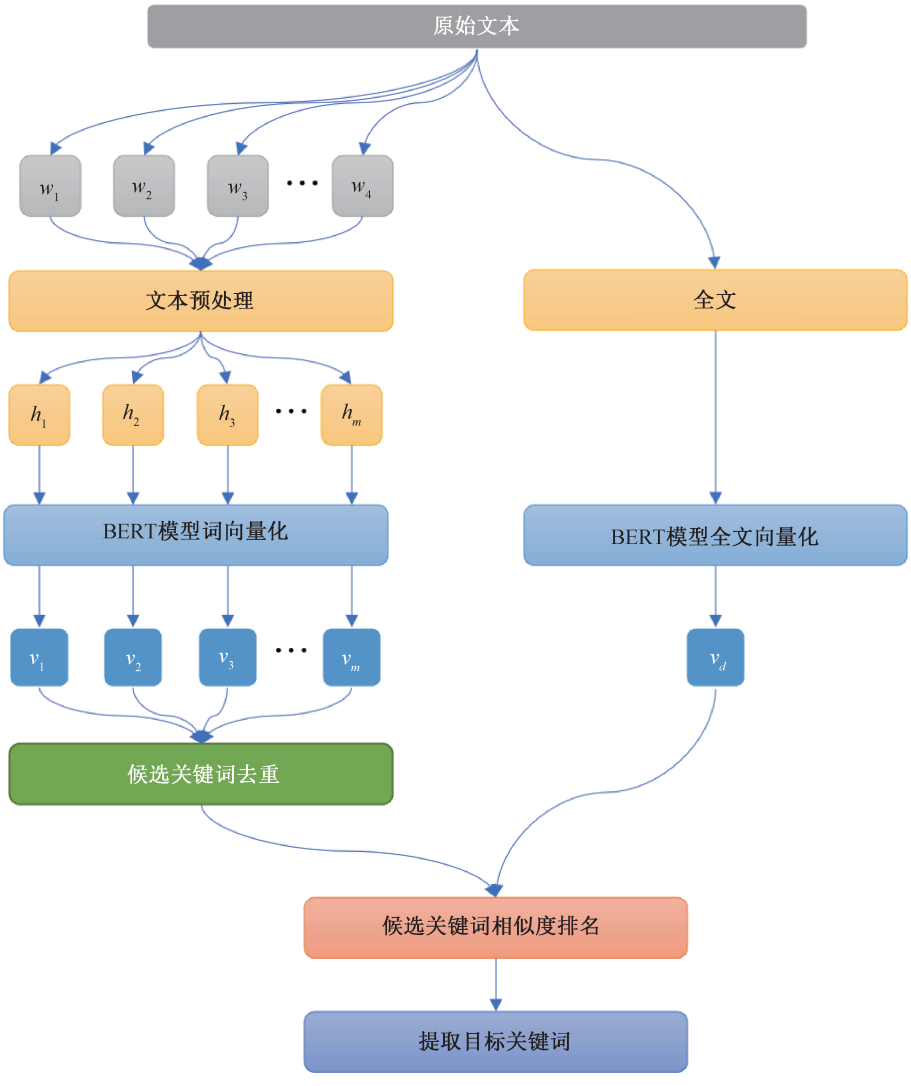

图 1 本文关键词提取模型示意

1.1 文本预处理与词向量获取

文本由各种词性的词、标点符号、注释等信息组成。在关键词提取任务中,并不需要利用所有的信息,若混入一些不必要的信息则会对提取结果产生负面影响。因此,对原始文本进行预处理并筛选出候选词很重要。

本文通过对文本进行预处理,将文本中的长句切分、去噪,并筛选出候选关键词。预处理使用 Jieba 工具包中分词模块的全模式进行分词处理,全模式是将文本中能够成词的词全部筛选出来,使用这种方式的优势在于为候选词提供更多的选择。主要过程为:首先将原始文本 D 切分为单词序列:$\{w_1,w_2,\cdots,w_n\}$,序列长度为 n;随后对词序列进行词性标注并筛选出名词、专有名词、名动词、副形词等关键词的常见词性,并将候选词与停用词表匹配,剔除停用词,得到 m 个候选词$\{h_1,h_2,\cdots,h_m\}$;接着,将原始文本 D 切分为单字序列:$\{c_1,c_2,\cdots,c_T\}$,T 表示全文单字的总量。设组成候选词 h 的单字在$[a,b]$区间内,则候选词对应的字序列为:$\{c_a,c_{a+1},\cdots,c_b\}$。

预处理后的文本已经被分割为候选词$\{h_1,h_2,\cdots,h_m\}$,随后需要将其转换为语义向量。本文所选用的 BERT 模型能够根据上下文动态地生成词向量,相较于静态词向量的优势在于能够更好地联合上下文语义,并将语义信息融入候选词的向量表示中,使获得的向量能够更准确的表示其在当前语境下的含义。词向量获取的具体过程如图 2 所示。

首先将原始文本 D 输入 BERT 分词器中进行切分处理,其对中文的切分方式是以单字为单位,将文本的切分成单字序列$\{c_1,c_2,\cdots,c_T\}$。将单字序列输入 BERT 模型的隐藏层中。通过 BERT 隐藏层获取向量序列,其过程可用式(1)表示:

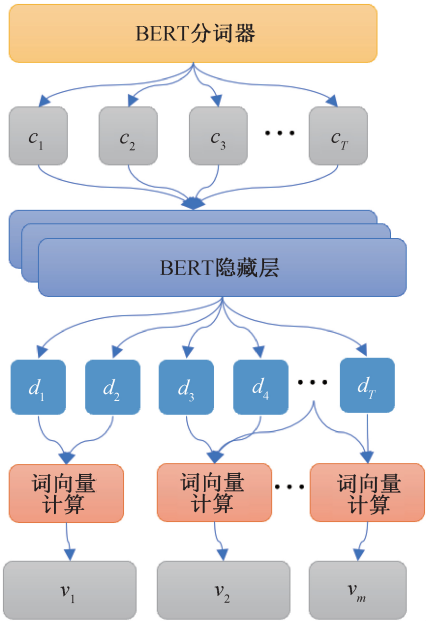

图 2　词向量获取的具体过程

$$\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_T\}=Bert\_encoder(\{c_1,c_2,\cdots,c_T\}) \tag{1}$$

其中:d_p,$p=1,2,\cdots,T$,为 BERT 模型的隐藏层 $Bert\_encoder$ 输出的单字向量,与输入单字 c_p 对应,且包含上下文语义信息。若候选词 h 所对应的单字序列为$\{c_a,c_{a+1},\cdots,c_b\}$,则通过式(1)中取得单字序列的对应位置字向量,得到候选词的对应单字向量序列$\{d_a,d_{a+1},\cdots,d_b\}$,并计算出词向量 v,v 的计算公式为:

$$\boldsymbol{v}=\mathrm{mean}(\boldsymbol{d}_a,\boldsymbol{d}_{a+1},\cdots,\boldsymbol{d}_b) \tag{2}$$

式(2)是对括号中的向量序列求平均向量来表示组成词的词向量,以保证不同长度的词向量维度统一,同时此向量包含了该词在上下文中的语义信息。

1.2　初选关键词词义去重

一篇文本中需要提取的主题词往往有多个,提取出的关键词需要表示该文本所属主题的不同层面。经过上述步骤筛选的候选关键词向量集合$\{v_1,v_2,\cdots,v_m\}$包含大量主题词,其中可能包含涵义相近的词。在一次关键词提取中,若提取出意思相似的词,则可以认为是无效的提取。因此本文使用语义聚类的方法对前述工作筛选出的候选词的词向量去重。

本文对候选关键词向量集合使用 K-means 聚类算法。首先根据需要提取的关键词数量选取聚类中心数,聚类的目的在于将语义相近甚至同义的词从候选关键词列表中排除,聚类中心数的选取首先需要保证比需要提取的关键词数量多,同时保证足够的语义丰富度。接着,分别计算候选词向量集合中的每一个向量至聚类中心位置距离,并根据距离长短,分别将集合中的词归入对应的类中。完成一次计算后,重新计算每个类的中心,再次计算词距离,迭代至收敛。聚类完成后取出距离聚类中心最近的词作为候选关键词输出。

1.3　全文语义向量获取

全文语义向量的生成步骤如图 3 所示。

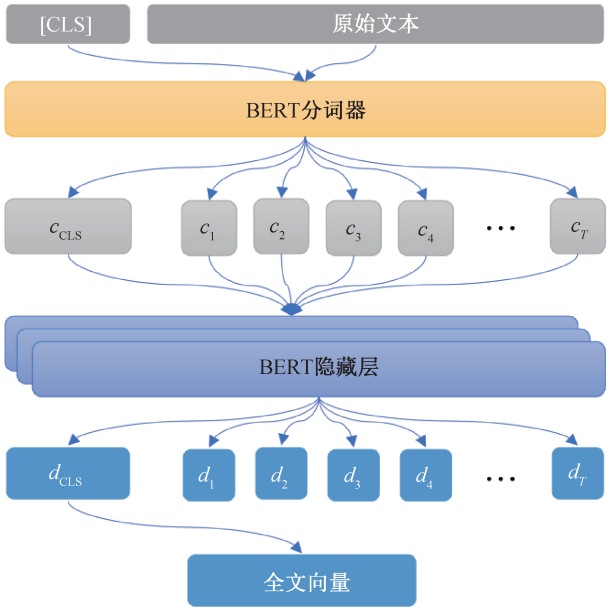

图 3　全文语义向量的生成步骤

首先在原始文本的起始位置添加“[CLS]”标识,加入此标识的目的是利用此标识获取全文语义向量。然后将原始文本 D 输入 BERT 分词器,即:

$$\{c_{\mathrm{CLS}},c_1,c_2,\cdots,c_T\}=Tokenizer(D) \tag{3}$$

其中:$Tokenizer$ 表示 BERT 模型分词器。由于 BERT 模型的隐藏层输入序列长度限制在 512 维以内,因此经过 BERT 模型分词器拆分的序列长度若超过 512 维,则将超过 512 维的部分截去。然后将得到的单字序列输入 BERT 模型的隐藏层中,使用的隐藏层与获取词向量任务的隐藏层一致,即:

$$\{\boldsymbol{d}_{\mathrm{CLS}},\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_T\}=Bert\_encoder(\{c_{\mathrm{CLS}},c_1,c_2,\cdots,c_T\}) \tag{4}$$

其中:$\boldsymbol{d}_{\mathrm{CLS}}$表示起始位置“[CLS]”标识所对应的输出向量,该向量包含了下文语义信息,因此考虑直接将此向量作为全文语义信息的表示向量。

1.4　候选词相似度计算排序

经过以上步骤处理完成的向量已处于同一维度下,对其进行相似度评判是有意义的。将上述步骤中获取的 $\boldsymbol{d}_{\mathrm{CLS}}$作为全文语义表示向量,该向量可以表示为:

$$\boldsymbol{d}_{\mathrm{CLS}}=(t_1,t_2,\cdots,t_{768}) \tag{5}$$

其中:t 表示全文向量中的每一个参数,共有 768 维。同样,经过处理的词向量 $\boldsymbol{v}_k$ 可以表示为:

$$\boldsymbol{v}_k=(u_1,u_2,\cdots,u_{768}) \tag{6}$$

与式(5)类似,u 表示词向量中的每一个参数,维数与全文向量相同。

由于都为经过同一层隐藏层所输出的向量,因此这组向量位于同一个二维空间中,可以使用余弦相似度、欧氏距离等距离算法来计算语义距离从而判断相互之间的相似度。本文选用余弦相似度来计算候选关键词与全文语义向量的相似度,根据式(5)和式(6)的向量,相似度的计算过程可用式(7)表示:

$$S_k(\boldsymbol{d}_{\mathrm{CLS}},\boldsymbol{v}_k)=\frac{\sum_1^{768}(t_i\times u_i)}{\sqrt{\sum_1^{768}t_i^2}\times\sqrt{\sum_1^{768}u_i^2}} \tag{7}$$

其中:S_k 为相似度;$\boldsymbol{d}_{\mathrm{CLS}}$为全文语义向量;$\boldsymbol{v}_k$ 为第 k 个词向量。经过计算后得到的余弦相似度结果范围为[0,1],因此将 S_k 直接作相似度评分分数,并依此进行排序,选出分数最高的 N 个关键词作为提取的最终结果输出。

2 实验与结果分析

2.1 实验环境与数据

为了验证分析本文所提出模型的性能,本文通过网络,随机筛选出 900 篇构建评估数据集,选用 4 种不同种类的无监督关键词提取模型与本文提出的模型在该数据集上进行对比实验。

2.1.1 实验环境

实验中计算机配置及环境如下:CPU 为 Intel (R) Core(TM) i5-7300HQ,主频为 2.5 GiHz,内存为 16 GiB,GPU 为 NVIDIA GeForce GTX 1050 Ti,系统为 Windows 10 64 位,python 版本号为 3.8,pytorch 版本号为 1.7.0。

2.1.2 数据准备

实验通过中国知网获取的混合领域论文摘要以及关键词信息作为实验评估语料来源,并参考 Bennani-Smires 等[13]选取实验数据的量级,随机选择的领域范围包括自然、教育、计算机、电子、交通、医疗、社会等方向的论文摘要,筛选出关键词标记完整且表述通顺的测试语料共 900 段,提取作者标记的关键词作为语料集的标注关键词。数据预处理的过程主要包括:去除标点符号、去除特殊字符、替换英文字符。在准备进行词向量获取之前需要对文本进行词性标注和分词处理。分词使用 Jieba 库的全模式分词,将能够成词的词语都提取出来,这样做的目的是使预处理产生的词尽可能考虑多种组合可能性,同时对分词结果产生的词语集进行词性标注,筛选出名词、专有名词、名动词、形容词、副形词等更常用来表示语义信息的词放入候选词表中。

2.1.3 预训练模型

预训练模型采用 Google 公开的 base 版 BERT 模型 Bert-base-chinese,包含 12 层 Transformer,并取用第 3 层隐藏层的输出向量作为词向量表示。最大序列长度为 512,学习率为 1×10^{-5},Batch_size 为 4,在全文语义表示的过程中,虽然大部分的摘要语料的长度都小于 512 的最大序列长度,但是仍有少部分摘要长度超过此值,由于摘要的核心内容大多位于开头和中间的部分,因此对于长度超过最大序列长度的数据,本文使用“去尾”的方法,去尾后对文献语义造成的影响相对更小。词向量维度为 768 维,文本词向量维度也为 768 维。

2.2 对比模型与评价指标

为了验证并分析本文提出算法的有效性,分别选取不同类型的无监督关键词提取模型,设置 2 组对比实验在数据集上进行测试,并与本文模型的提取结果进行对比。实验分别以提取 5 个和 8 个关键词作为标准,同时比较提取数量对提取准确度的影响。对比模型主要为以下 4 种:基于统计的关键词提取算法 TF-IDF,基于词图模型排序的 Text Rank 算法,基于预训练词向量的 Key-BERT[14]、Word2Vec[15] 关键词提取算法。

用于对比的 TF-IDF 算法步骤为:首先对原始文本进行分词操作,接着对其预处理,且与本文模型的操作保持一致,得到候选词;随后计算每个候选词的 TF 值,IDF 值则使用 Jieba 库中预设的 IDF 值表,然后对每个词分别计算其 TF-IDF 值作为评分,最终对评分排序并提取前 N 个词作为提取关键词。

用于对比的 Text Rank 使用 textrank4zh 库中提供的 Text Rank 算法,步骤为:首先对原始文本进行分词,接着进行预处理,且与本文模型的操作保持一致,得到候选词;随后构建词图模型,以每个候选词为节点,以 2 为窗口大小构建共现词的边,迭代直至收敛并得到每个节点的权重评分;最后对权重评分进行排序并提取前 N 个词作为提取关键词。

用于对比的 Key-BERT 使用 Key-BERT 库中封装的算法进行提取,步骤为:首先对原始文本进行分词和预处理的方法与前述方法一致,得到候选词;

随后将分词完成后的数据输入该库中的模型中;然后设置提取词汇长度为 1,且提取关键词的数量为 5 和 8;最后其输出的结果为提取的关键词。

用于对比的 Word2Vec 关键词提取算法所使用的预训练模型为基于简体中文维基百科训练的通用预训练模型,提取步骤为:首先对原始文本预处理,预处理的方法与步骤和本文模型一致,得到候选词;随后将候选词输入预训练模型中获得候选词向量;接着使用 $K-$means 聚类算法对候选词向量聚类,聚类中心数与需要提取的关键词数量相等,分别取聚类中心数为 5 和 8;最后取出每个聚类簇中心的词作为提取关键词。

实验使用准确率(P)、召回率(R)以及 F 值(F)来评价关键词提取的效果,3 种计算指标的计算方法如下:

计算提取的准确率 P 可用式(8)计算:

$$P=\frac{T}{N} \tag{8}$$

其中:P 为正确提取关键词的数量 T 与提取关键词的总数量 N 之比。

召回率 R 可用式(9)计算:

$$R=\frac{T}{K} \tag{9}$$

其中:提取出正确的关键词的总数 T 与作者标记的关键词数量 K 之比。

最后根据以上两个计算结果和式(10)计算 F 值:

$$F=\frac{2PR}{P+R} \tag{10}$$

关键词提取的结果见表 1。表中展示了关键词提取数量为 5 个及 8 个时,算法及模型关键词提取的准确率、召回率和 F 值。可以看出:在没有融合语义信息的算法中:基于统计的关键词提取算法 TF-IDF 提取 5 个关键词时的准确率为 31.21%,提取 8 个关键词时的准确率为 20.27%。基于词图模型排序的 Text Rank 算法提取 5 个关键词的准确率为 31.62%,提取 8 个关键词的准确率为 25.61%。这两种算法在提取 5 个关键词时的准确率、召回率和 F 值相近。提取 8 个关键词时,TF-IDF 算法的准确率和 F 值下降较多,召回率也略有下降。Text Rank 算法的准确率和 F 值下降,但召回率提高。本文模型相较于以上两种算法的准确率、召回率和 F 值更高,说明本文模型通过融合语义信息提升了关键词提取的质量。在使用预训练模型融合语义的关键词提取模型中:基于 Word2Vec 模型的关键词提取模型提取 5 个关键词时准确率为 27.13%,提取 8 个关键词时准确率为 22.56%。Key-BERT 模型提取 5 个关键词时准确率为 32.84%,提取 8 个关键词时准确率为 24.59%。可以看出,使用动态词向量的 Key-BERT 模型相比使用静态词向量的 Word2Vec 模型的准确率、召回率和 F 值更高。提取 8 个关键词时,准确率和 F 值均下降,但召回率提高。本文模型相较于以上两种模型,准确率、召回率和 F 值更高。说明本文模型通过融合 BERT 预训练模型提供的动态词向量中的语义信息,并添加聚类层对语义去重提升了关键词提取的质量。

表 1　关键词提取结果　%

模型	提取 5 个关键词结果			提取 8 个关键词结果		
	准确率	召回率	F 值	准确率	召回率	F 值
Text Rank	31.62	33.11	32.35	25.61	36.18	29.99
TF-IDF	31.21	32.43	31.81	20.27	31.24	24.59
Word2Vec	27.13	29.37	28.21	22.56	29.63	25.62
Key-BERT	32.84	35.28	34.02	24.59	41.76	30.95
本文	34.21	36.65	35.39	26.34	43.59	32.84

2.3　对比实验

为了验证在不同层数的 BERT 模型隐藏层中获取的向量在本数据集上对提取效果的影响,本文分别从上往下选取 1、3、5、7 和 11 层隐藏层获取的向量进行测试。在实验中,首先都通过 Jieba 库的全模式对初始数据进行分词,并进行词性标注来选择初步候选关键词,保证每次实验中选取的候选关键词一致,随后将初始文章输入同一个 BERT 模型中,控制取出向量所需的隐藏层层数,然后分别获得对应隐藏层数的候选词词向量,并通过相同的聚类算法得到候选关键词,接着从对应的隐藏层中取出全文语义向量与对应层中的词向量进行相似度计算后得到相似度评分并排序前 5 名作为最终关键词,最后使用相同的方法对提取词的 F 值进行计算评估,结果见表 2。

在同一隐藏层的实验中,获取词向量与获取全文语义向量所使用隐藏层的层数相同。在不同隐藏层对比实验中,提取流程一致,预处理使用相同的停

表 2 不同的 BERT 模型隐藏层提取 5 个关键词的 F 值结果

隐藏层数/层	F 值/%
1	34.23
3	35.39
5	32.68
7	28.56
11	37.41

用词表以及 Jieba 分词模式,保证获取的初步候选关键词相同,聚类去重步骤中的中心点选取数都为 10。其中第 1 层表示 BERT 模型隐藏层的最顶层,12 层为最底层。根据表 2 中的结果数据可以看出:底层向量作为词句向量进行关键词提取任务的效果要差于高层向量,但是高层向量中,第 3 层的提取效果更佳,直到最顶层后,提取效果降低。其原因在于:底层包含更多词语级别的表层信息,全文语义信息的捕获较弱;中层句法级别的信息逐渐增加,捕获到了更多句子级别的语义信息;高层句子、全文级别信息更加丰富,但词语级别的信息逐渐稀释遗忘。本文模型使用词向量与全文语义向量相结合的方式,且全文语义捕获对正确率的影响更大,因此在高层隐藏层拥有更高的正确率,但是由于词语级别信息在高层中逐渐稀释,效果在到达峰值后下降。

2.4 案例分析

本文展示了所使用数据集中的 3 段摘要提取案例,结果如图 4—图 9 所示。图 4、图 6 和图 8 为从数据集中随机选取的摘要原文案例。图 5、图 7 和图 9 的第一行为作者标记的准确关键词,第二行为使用本文模型提取出的 Top5、Top8 关键词以及移除聚类算法后的 Top5、Top8 关键词。作者标记关键词作为准确关键词,使用下划线标识,并在原文中标出。模型提取的关键词与作者标记的关键词完全相同的词用蓝色标出,同时标识在作者标记关键词与原文中。模型提取出的关键词中,包含部分语义与作者标记关键词相似的词用红色字体标出;同一组关键词中,语义相近的关键词用斜体字标出;提取出的关键词与作者标记的关键词相似,且在提取结果中拥有语义相近提取的关键词,用红色斜体标出。

根据中国服装协会发布数据显示,受国际市场需求减弱因素影响,我国服装产业出口已连续13个月出现负增长,2012年1至9月,14328家规模以上企业销售收入的增幅从年初的13%降至10%左右。企业销售利润降低,亏损面扩大,亏损额增多。2012年以来,服装出口形势和企业盈利状况依然不容乐观。面对纷繁复杂的世界经济形势,如何转变发展方式,实现品牌升级,提升在国际市场的竞争力,已成为中国服装业扭转出口颓势的当务之急。

图 4 摘要案例内容 1

作者标记关键词:亏损 服装出口 服装产业 品牌升级

模型提取关键词:
Top5:服装出口 不容乐观 亏损 品牌 产业
Top8:亏损额 销售 竞争力
移除聚类算法Top5:不容乐观 服装出口 *亏损额* 产业 *亏损面*
移除聚类算法Top8:销售 竞争力 规模

图 5 摘要案例内容 1 提取结果

中国城市医疗卫生体制的演变逻辑,通过建立一个综合性理论框架,阐释中国城市医疗卫生体制的演变逻辑发现,新中国成立后形成的城市医疗卫生计划体制与当时的外部环境相适应,且内部各组成要素之间协调互补。改革开放以来,整体社会经济制度走向市场体制,人口特征与疾病谱显著改变,居民收入水平显著提高且差距拉大,医疗卫生人员素质普遍提高,外部环境这些不可逆转的变化,内在地要求城市医疗卫生事业走出传统计划体制。这期间发生的诸多问题,源于体制转变期的冲突,是渐进式改革难以避免的。因此,不能以"改革"之名退回传统计划体制,应该逐步完善新体制,使其既与外部环境相适应,又能实现内部要素间的协调互补。

图 6 摘要案例内容 2

在以上三段案例中:所有使用下划线标出的作者标记关键词都在原文中出现。由图 4—图 5 的第一段案例本文模型提取的 Top5 关键词中:精确提取的有 2 个,分别是"服装出口"与"亏损",且"服装出口"的得分最高,表示该词在于全文语义的相似度计算中得到了最高的分值,可以认为该词与全文语义最接近。同时在前 5 名中,"品牌"与"产业"与准确关键词中的"品牌升级"与"服装产业"语义相近,

作者标记关键词:医疗卫生体制 计划体制 市场体制 渐进式改革

模型提取关键词:
Top5:人员素质 医疗卫生 渐进式 外部环境 计划体制
Top8:难以避免 经济 居民收入
移除聚类算法Top5:人员素质 *疾病谱* 冲突 *医疗卫生* *渐进式*
移除聚类算法Top8:外部环境 *逐步完善* 计划体制

图 7　摘要案例内容 2 提取结果

“微课导学”教学模式构建与实践——以中小学机器人教学为例,目前智能机器人已逐渐进入到我们日常生活中的各个应用领域,但我国中小学机器人教育起步较晚,近几年只有个别学校将机器人引入中小学课堂,亟待寻求一种适合中小学机器人教学的教学模式。在“微时代”、移动学习、在线教育等日趋盛行的背境下,全国掀起了“微课”研究热潮,但如何应用微课、采用何种教学模式组织教学的研究较少。文章通过对“翻转课堂”“研学后教”等教学模式进行研究和分析,结合机器人教学实操性强的特点,构建了以“微课”和“研学案”为教学载体的“微课导学”教学模式,以期为中小学机器人教学模式的研究和应用提供参考。

图 8　摘要案例内容 3

作者标记关键词:微课 机器人教育 微课导学 教学模式 构建

模型提取关键词:
Top5:教学模式 微课 学案 机器人 教育
Top8:起步 全国 学校
移除聚类算法Top5:教学模式 微课 学案 机器人 *教育*
移除聚类算法Top8:起步 *在线教育* 全国

图 9　摘要案例内容 3 提取结果

为相似提取,能够帮助概括原文语义,因此也可判定为有效提取。由图 6—图 7 的第二段案例可以看出:精确提取的关键词有 1 个,为“计划体制”;同时,在前 5 名中,“医疗卫生”与“渐进式”与准确关键词中的“医疗卫生体制”与“渐进式改革”语义相近。图 6—图 8 中的第三段案例与前两段类似,精确提取的关键词为“教学模式”与“微课”;同时,在前 5 名中,“机器人”与“教育”和准确关键词中的“机器人教育”语义相近。

产生相似提取的原因是本模型提取关键词的质量一定程度上受到预处理质量的限制,在预处理阶段完成后,筛选出的候选关键词的质量会影响实际提取的效果。

最后展示了本文模型在移除聚类算法后,仅计算候选词与原文语义相似度排序时,第一段案例的前 8 名关键词中:“亏损额”与“亏损面”有着相似的词义,都对应准确关键词中的“亏损”。第二段案例的前 8 名关键词结果中:“渐进式”与“逐步完善”有着相似的含义。第三段案例的前 8 名关键词提取结果中:“教育”与“在线教育”有着相似的含义,属于重复提取。在计算相似度的过程时,在二维空间中可以视为候选关键词次与全文语义距离的排序,聚类模块不仅考虑了候选词与全文语义的相似度距离,同时也考虑了候选词与候选词之间的语义距离,对语义相似的候选关键词筛选去重,因此能够使部分虽距离语义中心较远,但也能够作为关键词表示全文信息的词提取出来,从而提升关键词提取的准确性。

3　结　语

本文基于 BERT 模型,提出了一种无监督中文单文本关键词提取模型。该模型首先使用预处理技术将原始文本中选取多个候选关键词,然后使用预训练模型结合全文信息,将候选关键词向量化;接着对候选关键词的向量使用聚类算法对获取的候选关键词的语义去重处理,挑选出语义不重复的候选关键词;最后通过添加段落标记的方式,使用预训练模型获取全文语义向量,并计算全文语义与候选关键词语义的相似度,挑选出相似度较高的多个词作为文本的关键词。

与 4 种典型的关键词提取模型(基于数理统计的 TF-IDF、基于词图模型的 Text Rank 与基于预训练模型的 Word2Vec、Key-BERT)相比较,本文模型在提取论文摘要的关键词时,准确率有所提升;本文模型与去除聚类层的本文模型进行了对比,结果发现:本文添加的聚类层对候选关键词语义去重是有效果的,并能够提取出更高质量的关键词。因此,

本文模型通过融合预训练模型中丰富的语义信息并对候选关键词语义去重,提升了关键词提取的效果和质量。

参考文献:

[1] 李俊, 吕学强. 融合 BERT 语义加权与网络图的关键词抽取方法[J]. 计算机工程, 2020, 46(9): 89-94.

[2] Alzaidy R, Caragea C, Giles C L. Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents[C]//The World Wide Web Conference on - WWW '19. New York: ACM Press, 2019: 2551-2557.

[3] Campos R, Mangaravite V, Pasquali A, et al. YAKE! Keyword extraction from single documents using multiple local features[J]. Information Sciences, 2020, 509: 257-289.

[4] Liang X, Wu S, Li M, et al. Unsupervised keyphrase extraction by jointly modeling local and global context [EB/OL].(2021-9-15)[2021-10-18] https://arxiv.org/abs/2109.07293.

[5] Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents[M]// Berry M W, Kogan J. Text Mining: Applications and Theory. Chichester, UK: John Wiley & Sons, 2010: 1-20.

[6] Bougouin A, Boudin F, Daille B. TopicRank: Graph-based topic ranking for keyphrase extraction[C]// International Joint Conference on Natural Language Processing. Nagoya, Japan: IJCNLP, 2013: 543-551.

[7] Florescu C, Caragea C. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017, 1: 1105-1115.

[8] Wan X, Xiao J. Single document keyphrase extraction using neighborhood knowledge[C]// Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. Beijing: AAAI, 2008: 855-860.

[9] Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information[J]. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.

[10] Wang R, Liu W, McDonald C. Corpus-independent generic keyphrase extraction using word embedding vectors[C]//Software Engineering Research Conference. Australia: AAAI, 2014: 1-8.

[11] Gagliardi I, Artese M T. Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods[J]. Multimodal Technologies and Interaction, 2020, 4(2): 30.

[12] Mahata D, Kuriakose J, Shah R, et al. Key2Vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Louisiana: Association for Computational Linguistics, 2018, 2: 634-639.

[13] Bennani-Smires K, Musat C, Hossmann A, et al. Simple unsupervised keyphrase extraction using sentence embeddings[EB/OL]. (2018-01-13)[2021-10-21]. https://arxiv.org/abs/1801.04470.

[14] Sharma P, Li Y. Self-supervised contextual keyword and keyphrase retrieval with self-labelling[EB/OL]. (2019-09-06)[2021-10-21]. https://www.preprints.org/manuscript/201908.0073/v1.

[15] 李跃鹏, 金翠, 及俊川. 基于 word2vec 的关键词提取算法[J]. 科研信息化技术与应用, 2015, 6(4): 54-59.

(责任编辑:康 锋)